UNIVERSIDADE
BEIRA INTERIOR

How Can Deep Learning Aid Human Behavior Analysis?

Tiago Filipe Dias Santos Roxo

Tese para obtenção do Grau de Doutor em
Engenharia Informática
(3º ciclo de estudos)

Orientador: Professor Doutor Hugo Pedro Martins Carriço Proença
Coorientador: Professor Doutor Pedro Ricardo Morais Inácio

Covilhã, junho de 2025

Provas públicas realizadas em 23 de maio de 2025 com início às 10:30 horas.

Composição do júri

Presidente

Doutor Luís Filipe Barbosa de Almeida Alexandre
Professor Catedrático da Faculdade de Engenharia da Universidade da Beira Interior

Vogais

Doutor João Manuel Ribeiro da Silva Tavares
Professor Catedrático da Faculdade de Engenharia da Universidade do Porto

Doutor Hugo Pedro Martins Carriço Proença
Professor Catedrático da Faculdade de Engenharia da Universidade da Beira Interior

Doutor Paulo Luís Serras Lobato Correia
Professor Catedrático do Instituito Superior Técnico da Universidade de Lisboa

Doutor João Carlos Raposo Neves
Professor Associado da Faculdade de Engenharia da Universidade da Beira Interior

Doutor Nuno Miguel Mendonça da Silva Gonçalves
Professor Auxiliar da Faculdade de Ciência e Tecnologia da Universidade de Coimbra

Declaração de Integridade

Eu, Tiago Filipe Dias Santos Roxo, estudante com o número de inscrição 2807 do 3º ciclo de estudos em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridade da Universidade da Beira Interior**.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã, 03/06/2025

Assinado por: **Tiago Filipe Dias dos Santos Roxo**
Num. de Identificação: 13660462
Data: 2025.06.03 14:22:29+01'00'



This thesis was prepared at the University of Beria Interior, IT - Instituto de Telecomunicações, Soft Computing and Image Analysis Laboratory (SOCIA Lab), Covilhã Delegation, and was submitted to the University of Beira Interior for defense in a public examination session.

This work is funded by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the project UIDB/50008/2020 and research grant 2020.09847.BD. Also, it was supported by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT.





Acknowledgments

The research work presented in this thesis would not have been possible without the help and support of the individuals and organizations mentioned here.
First and foremost, I would like to express my utmost gratitude to my supervisor and coadvisor, Professor Dr. Hugo Proença and Professor Dr. Pedro Inácio, for their constant guidance, trust, patience, and support. Without their motivation, encouragement, expertise, research insight, and invaluable help it would have not been possible to complete this thesis.
The support of the University of Beira Interior, and Instituto de Telecomunicações is also acknowledged, as well as the financial support of the Fundação para a Ciência e a Tecnologia (FCT) through National Funds and co-funded by EU funds under Project UIDB/50008/2020; in part by the FCT Doctoral Grant 2020.09847.BD.
Also, I would also like to express my gratitude to all my colleagues of the Soft Computing and Image Analysis Lab (SOCIA-LAB) research group, for the enjoyable working environment that they contributed to.
Last, but not least, I would like to thank all the people close to myself in the last years, for their strong support, encouragement, friendship, and love.

List of Publications

Articles included in the thesis resulting from this doctoral research program

1. **Roxo, Tiago**, and Proença, Hugo. *"Is Gender "In-the-Wild" Inference Really a Solved Problem?."* IEEE Transactions on Biometrics, Behavior, and Identity Science (2021): 573-582.

2. **Roxo, Tiago**, and Proença, Hugo. *"YinYang-Net: Complementing Face and Body Information for Wild Gender Recognition."* IEEE Access 10 (2022): 28122-28132.

3. **Roxo, Tiago**, Costa, Joana Cabral, Inácio, P. R. and Proença, Hugo. *"How to Squeeze An Explanation Out of Your Model"*, 4th Workshop On Explainable & Interpretable Artificial Intelligence for Biometrics (xAI4Biometrics), Europen Conference on Computer Vision (ECCV) (2024), pp. 1-16;

4. **Roxo, Tiago**, Costa, Joana Cabral, Inácio, P. R., and Proença, Hugo. *"WASD: A Wilder Active Speaker Detection Dataset."* IEEE Transactions on Biometrics, Behavior, and Identity Science (2025), vol. 7, no. 1, pp. 61-70.

5. **Roxo, Tiago**, Costa, Joana Cabral, Inácio, P. R., and Proença, Hugo. *"BIAS: A Body-based Interpretable Active Speaker Approach."* IEEE Transactions on Biometrics, Behavior, and Identity Science (2025): 1-1.

6. **Roxo, Tiago**, Costa, Joana Cabral, Inácio, P. R., and Proença. *"ASDnB: Merging Face with Body Cues For Robust Active Speaker Detection"*, under review.

Other publications resulting from this doctoral research program not included in the thesis

1. Costa, Joana Cabral, **Roxo, Tiago**, Sequeiros, J. B, Proença, Hugo, and Inácio, P. R. *"Predicting CVSS metric via description interpretation."* IEEE Access (2022), vol. 10, pp. 59125-59134;

2. Costa, Joana Cabral, **Roxo, Tiago**, Lopes, C., Sequeiros, J. B., Simões, T. M., and Inácio, P. R. *"Theoretical and practical assessments over SSH."* Research and Practice in Technology Enhanced Learning 18 (2023): 030-030;

3. **Roxo, Tiago**, Costa, Joana Cabral, Inácio, P. R. and Proença, Hugo. *"On Exploring Audio Anomaly in Speech"*. In IEEE International Workshop on Information Forensics and Security (WIFS) (2023), pp. 1-6;

4. Costa, Joana Cabral, **Roxo, Tiago**, Proença, Hugo, and Inácio, P. R. *"How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses"*. in IEEE Access (2024), vol. 12, pp. 61113-61136;

Resumo

Recentemente, têm surgido vários tópicos de investigação motivados pelo aumento de dados de videovigilância e pela robustez de modelos de aprendizagem profunda, nomeadamente a nível de biometria humana, rastreio e reidentificação. Uma outra área que conjuga videovigilância com modelos de aprendizagem profunda à base de imagem, não explorada com a mesma extensão que as referidas, é a avaliação de comportamento humano. O menor volume de trabalho neste tópico no contexto de Ciência de Computadores e Informática não é surpreendente dadas as dificuldades inerentes na categorização de comportamento humano nas condições referidas, em particular com a ausência de cooperação humana. Tendo como referência a área de Psicologia, a análise de comportamento humano requer ambientes de experiência controlados, com cooperação dos participantes, onde a avaliação de características comportamentais é feita à base de um questionário em grelha. Nesse sentido, não há uma forma evidente de como os modelos de aprendizagem profunda podem auxiliar especialistas no comportamento humano, sendo por isso este tema um dos contributos desta tese para a comunidade científica.

Nesta tese exploramos a literatura em Psicologia para definir um conjunto de características com influência comprovada no comportamento humano e que sejam passíveis de serem avaliadas a partir de uma câmara de videovigilância. Deste modo, definimos o comportamento humano via perfilagem de pessoas usando sete características comportamentais: interação, posição relativa (à câmara), vestuário, biometria, proximidade a outras pessoas, postura e uso de objectos portáteis. Note que este tipo de análise não categoriza o comportamento humano para estados específicos (como agressivo ou depressivo, *e.g.*), mas permite criar um conjunto de características que podem ser usadas para perfilar pessoas, sendo assim aplicáveis para auxíliar/complementar o trabalho de especialistas em comportamento e/ou comparar traços comportamentais entre pessoas num mesmo contexto. Adicionalmente, para motivar o desenvolvimento de trabalhos nas áreas das características comportamentais referidas, revemos as abordagens estado-da-arte e os conjuntos de dados usados para realçar as limitações existentes e discutir os tópicos com maior potencial para desenvolvimento de trabalhos futuros.

Após a definição das características comportamentais, começamos por explorar a limitação de modelos de biometria atuais em condições de videovigilância, em particular a resiliência de métodos para classificação de género. Demonstrámos o desempenho reduzido destas abordagens em dados de videovigilância usando os conjuntos de dados *Pedestrian Attribute Recognition (PAR)* e realçámos as limitações de treinar modelos em contextos cooperativos para posteriormente os aplicar em cenários mais desafiantes. Suportado pelos resultados destas experiências, propomos um modelo mais robusto para classificação de género em contextos mais desafiantes, assente na premissa de complementar cara com informação corporal. Assim, estendemos o conhecimento da comunidade científica de um tema amplamente estudado (biometria, em particular classificação de género) ao analisar a aplicabilidade de modelos atuais em cenários onde estes, geralmente, não atuam (videovigilância).

Nesta tese também explorámos outro tema relacionado com as caraterísticas comportamentais: interação humana, nomeadamente a deteção da pessoa que fala (*Active Speaker Detection (ASD)*), em particular num contexto similar a videovigilância. Contrariamente ao tema de biometria/género, este é menos explorado na comunidade e os trabalhos desenvolvidos assentam principalmente na avaliação de pessoas a falar em cenários cooperativos e com boa qualidade de som e imagem (filmes). Deste modo, para demonstrar as limitações das abordagens atuais, começámos por criar um conjunto de dados de ASD (*Wilder Active Speaker Detection (WASD)*), composto por diferentes categorias com desafios incrementais para deteção de pessoas a falar, nomeadamente com degradação de qualidade de som e imagem e em contextos de não cooperação de participantes. Este conjunto de dados demonstrou a incapacidade dos modelos ASD atuais para atuar em cenários livres (similares a videovigilância), e evidenciou a importância de informação corporal em condições onde a qualidade de audio é reduzida e o acesso a dados faciais não é garantido. Seguindo esta premissa, desenhámos o primeiro modelo que complementa informação de áudio, cara e corpo para atingir resultados estado-da-arte em cenários mais desafiantes, em particular em condições de videovigilância, contendo adicionalmente uma nova forma de utilizar blocos *Squeeze-and-Excitation (SE)*, com o intuito de apresentar as razões para a decisão do modelo via representação visual. Esta permissa de utilização de blocos de SE foi também extendida para outros modelos e áreas relacionadas a ASD com o intuito de comprovar a sua viabilidade como aplicação geral de interpretabilidade independente do modelo. Embora o primeiro modelo proposto de ASD fosse superior ao estado-da-arte em cenários desafiantes, a sua performance não era igualmente robusta em contextos mais cooperativos. Nesse sentido, desenvolvemos um modelo melhorado que combina cara e corpo em simultâneo aquando da extração de informação visual que, em conjunto com pretreino em dados desafiantes, tem desempenhos estado-da-arte em cenários cooperativos e desafiantes (como videovigilância). Dada a sua inovação, estes trabalhos servem como base para a avaliação de interação humana em dados mais desafiantes e interpretabilidade de modelos.

Palavras-chave

Análise de Comportamento Humano, Sistema de Vigilância Automático, Biometria Não Cooperativa, Videovigilância, Classificação de Género em Condições Adversas, Conjunto de Dados para Deteção de Orador, Deteção de Orador com Informação Corporal, Interpretabilidade Visual.

Resumo Alargado

Recentemente, têm surgido vários tópicos de investigação motivados pelo aumento de dados de videovigilância e pela robustez de modelos de aprendizagem profunda, nomeadamente a nível de biometria humana, rastreio e reidentificação. Uma outra área que conjuga videovigilância com modelos de aprendizagem profunda à base de imagem, não explorada com a mesma extensão que as referidas, é a avaliação de comportamento humano. O menor volume de trabalho neste tópico no contexto de Ciência de Computadores e Informática não é surpreendente dadas as dificuldades inerentes na categorização de comportamento humano nas condições referidas, em particular com a ausência de cooperação humana. Tendo como referência a área de Psicologia, a análise de comportamento humano requer ambientes de experiência controlados, com cooperação dos participantes, onde a avaliação de características comportamentais é feita à base de um questionário em grelha. Nesse sentido, não há uma forma evidente de como os modelos de aprendizagem profunda podem auxiliar especialistas no comportamento humano, sendo por isso este tema um dos contributos desta tese para a comunidade científica.

Trabalhos relacionados com análise de comportamento humano no contexto de Ciência da Computação não são abundantes e podem, geralmente, associar-se a dois grandes grupos: 1) à base de ações e 2) à base de emoções. Estes tipos de abordagens acarretam limitações pois o primeiro grupo limita a avaliação de comportamento humano a uma classificação de ações e o segundo grupo restringe a quantidade de informação passível de ser analisada e depende, fortemente, de cooperação de participantes. Nesse sentido, a ausência de mais trabalhos associados a comportamento parece padecer de uma falta de motivação clara entre o que é passível de ser analisado visualmente por um modelo de aprendizagem profunda e a sua relação com comportamento humano. Por estas razões, nesta tese exploramos a literatura em Psicologia para definir um conjunto de características com influência comprovada no comportamento humano e que sejam passíveis de serem avaliadas a partir de uma câmara de videovigilância. A motivação para tal prende-se à noção de querer extrapolar a análise para algo mais que simples ações (distinção do grupo 1) e que esta possa ser feita sem restrição do espaço de experiência e sem dependência de cooperação de sujeitos (distinção do grupo 2). Assim, definimos o comportamento humano via perfilagem de pessoas usando sete características comportamentais: interação, posição relativa (à câmara), vestuário, biometria, proximidade a outras pessoas, postura e uso de objectos portáteis. Note que este tipo de análise não categoriza o comportamento humano para estados específicos (como agressivo ou depressivo, *e.g.*), mas permite criar um conjunto das características que podem ser usadas para perfilar pessoas, sendo assim aplicáveis para auxiliar/complementar o trabalho de especialistas em comportamento e/ou comparar traços comportamentais entre pessoas num mesmo contexto. Este tipo de aplicação tem particular interese em cenário de avaliação comportamental de clientes (com intuito de melhorar a experiência de compra), alunos (como medida preventiva de abandono escolar), ou controlo de segurança de ambientes (detetar a presença de comportamentos anómalos de várias pessoas num local pode indicar

problemas). Para motivar o desenvolvimento de trabalhos nas áreas das características comportamentais referidas, revemos as abordagens estado-da-arte e os conjuntos de dados usados para realçar as limitações existentes e discutir os tópicos com maior potencial para desenvolvimento de trabalhos futuros.

Após a definição das características comportamentais, começamos por explorar a limitação de modelos de biometria atuais em condições de videovigilância, em particular a resiliência de métodos para classificação de género. O foco inicial para esta primeira característica foi motivado pela biometria ser um tema amplamente estudado e trabalhado e é essencialmente prototipada para cenários de cooperação de sujeitos. Assim, primeiramente definimos um conjunto de dados que nos permitisse demonstrar as dificuldades de classificação de biometria fraca, como o género, tendo usado para tal conjuntos de dados PAR. Embora estes não sejam usados especificamente para este efeito (classificação de género), são usados para classificação de vários atributos pedestres, um dos quais é o género. Assim, um dos primeiros trabalhos demonstra o desempenho reduzido destas abordagens em dados de videovigilância usando os conjuntos de dados PAR e realça as limitações de treinar modelos em contextos cooperativos para posteriormente os aplicar em cenários mais desafiantes. Suportado pelos resultados destas experiências, propomos um modelo mais robusto para classificação de género em contextos mais desafiantes, assente na premissa de complementar cara com informação corporal. Naturalmente, este conceito depende do acesso a informação facial para a poder usar no contexto de videovigilância (PAR). Para tal, fizemos uso de informação de pose (*pose estimation*) para detetar se a pessoa estava de frente ou de lado, permitindo obter a região facial via uma nova abordagem à base de infomação de pose, capaz de competir com abordagens específicas para deteção de região da cabeça. A conjugação de informação facial com corporal permite obter melhores resultados em ambientes desafiantes, facilitando o foco de modelo para zonas de interesse adequadas. Com estes trabalhos, estendemos o conhecimento da comunidade científica de um tema amplamente estudado (biometria, em particular classificação de género) ao analisar a aplicabilidade de modelos atuais em cenários onde estes, geralmente, não atuam.

Nesta tese também explorámos outro tema relacionado com as caraterísticas comportamentais: interação humana, nomeadamente a deteção da pessoa que fala (ASD), em particular num contexto similar a videovigilância. Contrariamente ao tema de biometria/género, este é menos explorado na comunidade e os trabalhos desenvolvidos assentam principalmente na avaliação de pessoas a falar em cenários cooperativos e com boa qualidade de som e imagem (filmes). O contexto de videovigilância representa, por isso, desafios para os modelos tendo em conta a abordagem usada pelo estado-da-arte: conjugação entre imagem facial e som. Uma outra evidência para a necessidade de conjuntos de dados mais desafiantes é a saturação de resultados dos modelos atuais no conjunto de dados mais usado (AVA-ActiveSpeaker). Assim, para demonstrar as limitações das abordagens atuais, começámos por criar um conjunto de dados de ASD (WASD), composto por diferentes categorias com desafios incrementais para deteção de pessoas a falar, nomeadamente com degradação de qualidade de som e imagem e em contextos de não cooperação

de participantes. Este conjunto de dados demonstrou a incapacidade dos modelos ASD atuais para atuar em cenários livres (similares a videovigilância), e evidenciou a importância de informação corporal em condições onde a qualidade de audio é reduzida e o acesso a dados faciais não é garantido.

Seguindo a premissa da importância corporal para ASD, desenhámos um modelo, **BIAS**, que, pela primeira vez, complementa informação de áudio, cara e corpo para atingir resultados estado-da-arte em cenários desafiantes, em particular em condições de videovigilância. A proposta deste modelo foi também conjugada com uma nova forma de utilizar blocos SE, capaz de apresentar as razões para a decisão do modelo via representação visual. A lógica para este novo uso prende-se à estrutura do bloco de SE, que é composto por um vetor de importância de *features*, normalmente usado intra *backbones* para melhorar a extração de *features*. Neste trabalho incluímo-lo em pós extração para obter as *features* mais importantes a partir deste vetor, permitindo a representação da atenção do modelo, de forma similar a *Class Activation Maps (CAM)*, o que confere interpretabilidade visual à decisão do modelo. Finalmente, dada a falta de descrição textual para cenários mais subtis de ações relacionadas com ASD (falar, tapar boca, cruzar braços, *e.g.*), este trabalho foi complementado por um conjunto de dados para este fim, o qual usámos para treinar um modelo ViT-GPT2 para complementar a interpretabilidade visual de BIAS com a descrição textual da cena, produzindo um *setup* para interpretabilidade completa.

A aplicação de blocos SE para interpretabilide de modelos foi também extendida para outros modelos e áreas relacionadas com o intuito de mostrar a viabilidade desta abordagem para interpretabilidade independente de modelos. Em particular, a aplicação destes blocos foi explorada em contexto de imagens (usando conjuntos de dados de objetos e caras) e vídeo/multi-modal (contexto ASD), avaliando características variadas e diferentes modelos. As experiências mostraram a robustez desta abordagem comparativamente a variantes de GradCAM, aplicável em modelos tradicionais e customizáveis, sem comprometer o desempenho do modelo na tarefa original.

Embora o modelo inicial de ASD (BIAS) tenha dado um passo importante no desenvolvimento de modelos de ASD mais robustos, em particular para cenários desafiantes como videovigilância, a sua performance em cenários mais cooperativos (onde o acesso a informação facial é garantido) não era superior aos melhores modelos existentes. Nesse sentido, propusemos um novo modelo **ASDnB** que conjuga informação corporal e facial ao nível de extração visual, com complemento gradual a diferentes níveis de extração. Esta nova abordagem de conjugação de informação visual, em conjunto com pretreino em dados mais desafiantes, permitiu a inclusão de informação corporal relevante para ASD, produzindo resultados superiores ao estado-da-arte em cenários cooperativos e desafiantes, e corrigindo assim as limitações do primeiro modelo proposto. Dada a inovação destas contribuições, estas servem como base para a avaliação de interação humana em dados mais desafiantes e interpretabilidade de modelos.

Abstract

With the increase of available surveillance data and robustness of state-of-the-art deep learning models, various recent research topics focus on human biometric assessment, tracking and person re-identification. However, one other area of work not extensively explored that can combine surveillance and visual-based models is assessing human behavior. The lack of work in this topic is not surprising given the inherent difficulties on categorizing human behavior in such conditions, in particular without subject cooperation. Based on the psychology literature, human behavior analysis typically requires controlled experimental environments, with subject cooperation and assessing features via grid-based survey. As such, it is not clear on how deep learning models can aid psychology experts in human behavior analysis, which is where this thesis intents to contribute to the body of knowledge.

We extensively review psychology literature to define a set of features that have been proven as influential towards human behavior and that can be assessed via camera in surveillance-like conditions. This way, we define human behavior via subject profiling using seven behavioral features: interaction, relative position, clothing, soft biometrics, subject proximity, pose, and use of handheld devices. Note that this analysis does not categorize human behavior into specific states (*e.g.* aggressive, depressive) but rather creates a set of features that can be used to profile subjects, usable to aid/complement behavioral experts and to compare behavioral traits between subjects in a scene. Furthermore, to motivate the development of works in these areas, we review state-of-the-art approaches and datasets to highlight the limitation of certain areas and discuss the topics worth exploring for future works.

After defining a set of behavioral features, we start by exploring the limitation of current biometric models in surveillance conditions, in particular the resilience of gender inference approaches. We demonstrate that these models underperform in surveillance-like data, using PAR datasets, highlighting the limitations of training in cooperative settings to perform in wilder conditions. Supported by the findings of our initial experiments, complementing face and body information arouse as a viable strategy to increase model robustness in these conditions, which lead us to design and propose a new model for wild gender inference based on this premise. This way, we extend the knowledge of an extensive discussed literature topic (gender classification) by exploring its application in settings where current models do not typically perform (surveillance).

We also explore the topic of human interaction, namely Active Speaker Detection, in particular towards more uncooperative scenarios such as surveillance conditions. Contrary to the gender/biometrics topic, this is a lesser explored area where works are mainly based on assessing active speakers via face and audio information in cooperative conditions and with good audio and image quality (movie settings). As such, to clearly demonstrate the limitations of state-of-the-art ASD models we start by creating a wilder ASD dataset (WASD), composed of different categories with increasing challenges towards ASD, namely with audio and image quality degradation, and containing uncooperative

subjects. This dataset highlighted the limitations of current models to deal with unconstrained scenarios (*e.g.* surveillance conditions), while also displaying the importance of body information in conditions where audio quality is subpar and face access is not guaranteed. Following this premise, we design the first model that complements audio, face, and body information to achieve state-of-the-art performance in challenging conditions, in particular surveillance settings. Furthermore, this model also proposed a novel way to combine data via SE blocks, which allowed to provide reasoning behind model's decision by visual interpretability. The use of SE blocks was also extended to other models and ASD-related areas to highlight the viability of this approach for model-agnostic interpretability. Although this initial model was superior to the state-of-the-art in challenging data, its performance in cooperative settings was not as robust. As such, we develop a new model that simultaneously combines face and body information in visual data extraction which, in conjunction with pretraining in challenging data, leads to state-of-the-art performance in both cooperation and challenging conditions (such as surveillance settings). These works pave a new way to assess human interaction in more challenging data and with model interpretability, serving as baselines for future works.

Keywords

Human Behavior Analysis, Automated Surveillance Systems, Non-Cooperative Biometric, Surveillance Settings, Wild Gender Classification, Active Speaker Detection Dataset, Body-based Active Speaker Detection, Visual Interpretability.

Contents

1 Introduction **1**

1.1 Problem Definition . . . 1
1.2 Work Relevance and Motivation . . . 2
1.3 Research Objectives . . . 3
1.4 Thesis Work Plan . . . 3
1.5 Main Contributions . . . 5
1.6 Thesis Organization . . . 7

2 Human Behavior Analysis: A Survey **9**

2.1 Introduction . . . 9
2.2 Human Behavior Analysis Approaches . . . 10
2.2.1 Behavior Analysis as Action Recognition . . . 10
2.2.2 Behavior Analysis as Emotion Recognition . . . 11
2.2.3 Considered Behavior Analysis . . . 11
2.3 Features Considered . . . 12
2.3.1 Verbal Social Interaction . . . 13
2.3.2 Physical Social Interaction . . . 17
2.3.3 Relative Position . . . 21
2.3.4 Handheld Devices . . . 25
2.3.5 Proximity . . . 28
2.3.6 Clothing . . . 32
2.3.7 Pose Estimation . . . 35
2.3.8 Soft Biometrics . . . 40
2.4 Current Limitations and Future Directions . . . 44
2.5 Conclusion . . . 46

3 Is Gender "In-the-Wild" Inference Really a Solved Problem? **47**

3.1 Introduction . . . 47
3.2 Gender Inference State-of-the-Art . . . 49
3.2.1 Face-Based Methods . . . 49
3.2.2 Body-Based Methods . . . 49
3.3 Methodologies . . . 50
3.3.1 Methods and Datasets . . . 50
3.3.2 In-the-Wild Major Variability Factors . . . 51
3.3.3 Shapley Value . . . 53
3.3.4 Head Detection . . . 53
3.3.5 Face Importance Analysis . . . 54
3.4 Experiments . . . 54
3.4.1 Dataset and Evaluation Metrics . . . 54
3.4.2 Implementation Details . . . 55

3.4.3 Cross-Domain Performance . . . 56
3.4.4 Dataset Analysis . . . 56
3.4.5 Methods Evaluation . . . 57
3.4.6 Interpreting Gender Inference . . . 58
3.4.7 Face Importance in Gender Inference . . . 61
3.5 Conclusions . . . 63

4 YinYang-Net: Complementing Face and Body Information for Wild Gender Recognition 65
4.1 Introduction . . . 65
4.2 Related Work . . . 67
4.2.1 Face-Based Gender Recognition . . . 67
4.2.2 Body-Based Gender Recognition . . . 67
4.2.3 Attention-Based Gender Recognition . . . 68
4.3 YinYang-Net . . . 69
4.3.1 Network Architecture . . . 69
4.3.2 Face Attention Module . . . 70
4.3.3 Implementation Details . . . 70
4.4 Frontal PAR and Face Datasets . . . 71
4.4.1 Head Detection . . . 71
4.4.2 Data Analysis . . . 73
4.5 Experiments . . . 74
4.5.1 Datasets, Methods, and Evaluation Metrics . . . 74
4.5.2 Gender Recognition . . . 75
4.5.3 Within and Cross-Domain Performance . . . 76
4.5.4 Face Attention Influence . . . 77
4.5.5 Ablation Studies . . . 79
4.6 Conclusions . . . 80

5 WASD: A Wilder Active Speaker Detection Dataset 81
5.1 Introduction . . . 81
5.2 Related Work . . . 83
5.3 Dataset . . . 84
5.3.1 Video and Category Selection . . . 85
5.3.2 Main Characteristics . . . 86
5.3.3 WASD Annotations . . . 87
5.4 Experiments . . . 87
5.4.1 Datasets, Models, and Evaluation Metric . . . 87
5.4.2 Limitations of AVA-ActiveSpeaker Training . . . 88
5.4.3 Models Robustness in WASD . . . 90
5.4.4 Qualitative Analysis . . . 92
5.5 Proposed Baseline . . . 92
5.5.1 TalkNet Architecture Modifications . . . 93

5.5.2 Body Importance Assessment . . . 93
5.5.3 Training Influence in Model Performance . . . 94
5.5.4 Body Importance in Columbia . . . 95
5.6 Conclusion . . . 95

6 BIAS: A Body-based Interpretable Active Speaker Approach 97
6.1 Introduction . . . 97
6.2 Related Work . . . 99
6.3 BIAS Approach . . . 100
6.3.1 Visual and Audio Encoders . . . 101
6.3.2 Squeeze-and-Excitation for Interpretability . . . 101
6.3.3 Self-attention and Loss Function . . . 102
6.3.4 ASD-Text Dataset . . . 103
6.3.5 Implementation Details . . . 103
6.4 Experiments . . . 104
6.4.1 Datasets, Models, and Evaluation Metrics . . . 104
6.4.2 BIAS Performance in WASD . . . 104
6.4.3 BIAS Performance in Other Datasets . . . 105
6.4.4 Ablation Studies . . . 106
6.5 BIAS Interpretability . . . 108
6.5.1 Squeeze-and-Excitation Visual Interpretability . . . 108
6.5.2 Feature Importance Assessment . . . 110
6.5.3 ASD-Text Dataset . . . 112
6.6 Conclusion . . . 113

7 How to Squeeze An Explanation Out of Your Model 115
7.1 Introduction . . . 115
7.2 Related Work . . . 116
7.3 Squeeze-and-Excitation for Interpretability . . . 117
7.3.1 SE Definition and Concepts . . . 117
7.3.2 SE Visual Interpretability . . . 118
7.3.3 Implementation Details . . . 119
7.4 Experiments . . . 120
7.4.1 Datasets, Methods, Models, and Evaluation Metrics . . . 120
7.4.2 SE Inclusion on Model Performance . . . 120
7.4.3 Qualitative Performance of SE Interpretability . . . 123
7.4.4 Multi-Modal Feature Importance via SE . . . 123
7.4.5 Quantitative Performance of SE Interpretability . . . 125
7.5 Conclusion . . . 125

8 ASDnB: Merging Face with Body Cues For Robust Active Speaker Detection 127
8.1 Introduction . . . 127

8.2 Related Work . . . 129
8.3 ASDnB . . . 130
8.3.1 Visual Encoder . . . 130
8.3.2 Audio Encoder . . . 131
8.3.3 Temporal Modeling in Classifier . . . 132
8.3.4 Loss Function . . . 132
8.3.5 Obtaining Body Data . . . 133
8.3.6 Implementation Details . . . 133
8.4 Experiments . . . 134
8.4.1 Datasets and Evaluation Metrics . . . 134
8.4.2 ASDnB Performance in AVA-ActiveSpeaker . . . 134
8.4.3 ASDnB Performance in Other Datasets . . . 135
8.4.4 Ablation Studies . . . 136
8.4.5 ASDnB Performance Analysis . . . 138
8.5 Conclusion . . . 140

9 Conclusion and Future Directions **141**
9.1 Conclusion . . . 141
9.2 Future Directions . . . 142

Bibliografia **191**

List of Figures

1.1 Timeline of tasks, with the respective milestones. Each milestone corresponds to a scientific contribution, whose descriptions are in Fig. 1.3. . . . 4
1.2 Description of the proposed tasks. Task numbers refer to the numbers presented in Fig. 1.1. . . . 4
1.3 Description of the achieved milestones. Milestone numbers refer to the numbers presented in Fig. 1.1. . . . 5
1.4 Main features for human behavioral analysis via automated vision-based approaches. . . . 6

2.1 Overview of the use of behavioral features for subject profiling. Each feature is analysed from images/video of a surveillance camera in a non-cooperative manner. . . . 12
2.2 Overview of ASD models: combining face images with audio features to predict active speakers. . . . 15
2.3 Representative images from AVA ActiveSpeaker (top row) and WASD datasets (bottom row). . . . 16
2.4 Variations of action recognition tasks. . . . 19
2.5 Representative images from University of Texas (UT)-Interaction (top row) and AVA datasets (bottom row). . . . 20
2.6 Overview of the two key steps for MOT. . . . 22
2.7 Representative images from MOT20 (top row) and KITTI Multi-Object Tracking (MOT) datasets (bottom row). . . . 24
2.8 Overview of a two-stage approach for Human-Object Interaction (HOI). . . . 26
2.9 Representative images from Humans Interacting with Common Objects Detection (HICO-Det) (top row) and Verbs in Common Objects in Contex (VCOCO) datasets (bottom row). . . . 28
2.10 Examples of depth estimation in indoor and outdoor scenarios. . . . 30
2.11 Representative images from KITTI (top row) and NYU Depth (bottom row) dataset. . . . 31
2.12 Clothing related tasks. . . . 33
2.13 Representative images from DeepFashion (top row) and PETA (bottom row) datasets. . . . 35
2.14 Overview of top-down and bottom-up approaches for multi-person 2D pose estimation. . . . 37
2.15 Representative images from Common Objects in Context (COCO) (top row) and Max Planck Institut Informatik (MPII) datasets (bottom row). . . . 38
2.16 Representative images from PA-100K (top row, first four images), Labeled Faces in the Wild (LFW) (top row, last 3 images), and UTKFace datasets (bottom row). . . . 43

3.1 Overview of the used approaches for image and subject-based features correlation with gender inference. We divide the work into three major phases: 1) image quality perception; 2) gender inference interpretation; and 3) face importance analysis. We divide the PAR datasets considering image-based features and correlate image and subject-based features to gender inference accuracy of a meta-model. This model is a combination of five PAR models. The face importance is analyzed in two ways: by evaluating its reliability in wild conditions and assessing pose influence in face-based gender inference. 51

3.2 Dataset examples of the evaluated features, divided into image (*luminosity* and *blurriness*) and subject-based (*head*, *upper body*, and *lower body* availability) features. Each row corresponds to one of the three PAR datasets. Each column contains opposite cases of each feature, with the left one representing *bad quality* (with feature value closer to its minimum) and the right one representing *good quality*. For the blurriness feature, high feature value represents *bad quality*. All images are resized to the same resolution for visualization purposes. 52

3.3 Examples of images considered as *frontal*, *sideways* and *backside*. Each row is related to one of the three PAR datasets and each column regards examples of the respective pose. 53

3.4 Phases of the head detection process, displaying pose estimation and head bounding box size. Faces are obtained exclusively from the frontal images of each dataset. 54

3.5 PAR dataset quality division, taking into account image-based feature values. Each dataset illustrates examples of the corresponding quality levels. 58

3.6 Bar plot of the mean absolute SHapley Additive exPlanations (SHAP) values for image and subject-based features. Blue colored bars correlate with gender correct inference, while the red one correlates with gender misinference. Features are ordered by its influence in gender inference accuracy. 59

3.7 Dependence scatter plots of Face (a), Lower Body (b), Upper Body (c), Resolution (d), Luminosity (e), and Blurriness (f) SHAP values. Face and Lower Body plots present a correlation with the named feature and Lower Body and Luminosity, respectively. These plots contain a right y axis with color variation, representing correlated feature value variance. Color variation from blue to red in the other plots represent increasing values of the feature. Image-based features are normalized. 59

3.8 Bar plots of the evaluated features, presenting the SHAP values for low, medium, and high quality image datasets. Datasets are represented by different colors. High SHAP values represent high feature relevance for gender correct inference. The bar plot is broken in the vertical axis for better visualization. 61

4.1 Illustration of the effect of the learnable fusion matrix, a major component of YinYang-Net (YY-Net), as an attention mechanism with respect to the (non)availability of facial information. 66

4.2 Cohesive perspective of the YY-Net, divided into three parts: *1) Model Architecture*; *2) Backbone*; and *3) Classifier*. The *Model Architecture* displays our processing flow, where the input data feeds two different backbones: 1) for the body; and 2) for facial information. The facial regions are obtained according our head detection approach, using pose information. The backbones derive from ResNet50, with classifier-related layers ignored, as shown in the *Backbone* portion. The combination of the face and body inputs is done through a classifier, which uses the proposed Face Attention Module to fuse both types of information effectively. 69

4.3 Examples of images considered as *frontal*, *sideways* and *backside*. Each row is related to one of the three PAR datasets and each column regards examples of the respective pose. 71

4.4 Comparison of the head Region of Interest (ROI) detection results between Multi-Task Cascaded Convolutional Neural Networks (MTCNN), EXTD, RetinaFace, and our approach, in images with varying quality, head pose and blurriness. Green boxes denote cases where the method accurately locates the face, and red boxes denote the opposite. Dotted red boxes refer to cases where the model over-predicted the number of existent faces. All images are resized to the same resolution for visualization purposes. 72

4.5 Examples of image quality variability in *wild* face datasets. The datasets contain subjects with varying facial expressions, levels of cooperativeness, and partial occlusions. Each row refers to one variability factor, and each group of four columns regards examples of a PAR dataset. 73

4.6 Comparison of StrongBase and YY-Net attention focus from images where YY-Net correctly classifies gender and StrongBase does not. Each row is related to one of the categories of diverging focus. 77

4.7 Crowd effect on attention focus from StrongBase and YY-Net. Each column displays an example where YY-Net accurately classifies gender and StrongBase misclassifies it. 77

4.8 Scatter plot of the difference between YY-Net and StrongBase predictions, and Ground Truth (GT) labels, in RAP and PA-100K images. Points in blue represent cases where YY-Net outperforms StrongBase, and points in red denote the opposite. Four different regions are highlighted in the plot, displaying the cases where a model outperforms the other the most, for male (M) and female (F) recognition. Representative images, from the four regions, are presented sideways of the plot, accompanied by descriptive labels. 78

5.1 AVA-ActiveSpeaker (AVA) state-of-the-art models achieve over 94% mean Average Precision (mAP) in active speaker detection, solely based on **face** and **audio data**. However, this approach may not be suited for uncooperative poses, non-guaranteed face access, or unreliable image/audio quality. How well do these models perform in such scenarios? And can body information aid in this task? To answer these questions (and more), we propose WASD, a Wilder Active Speaker Detection dataset. 82

5.2 Considered categories of WASD, with relative audio and face quality represented. Categories range from low (Optimal Conditions) to high (Surveillance Settings) ASD difficulty by varying audio and face quality. Easier categories contain similar characteristics to AVA-ActiveSpeaker (AVA-like), while harder ones are the novelty of WASD. 84

5.3 Gender, language, race, speaking activity, and number of speakers distribution of WASD. Afro refers to African and Afro-American people. On the right, distribution of head-body and body-image proportions of WASD categories. WASD is a balanced demographics dataset, with *talking* being the predominant speaking activity, mainly composed of few people conversations, where audio impaired categories (Speech Impairment and Human Voice Noise) have speakers closer to the camera, and Surveillance Settings has speakers further from it. 86

5.4 Average performance (mAP) variation of the four models on WASD categories, when trained on AVA-ActiveSpeaker and WASD. AVA-ActiveSpeaker is represented as AVA. 89

5.5 Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for models trained in AVA-ActiveSpeaker and WASD, and evaluated in Easy (left) and Hard (right) groups of WASD. AVA-ActiveSpeaker is represented as AVA. 90

5.6 Incorrect model inference in different scenarios. Source of misconception: a) awe expression, with sudden and subtle mouth movement, while having human voice in the background; b) partial facial occlusion from scene object; and c) slight mouth occlusion from hand movement. 91

5.7 Incorrect model inference by mixing active speakers. Hand and arm movement (from right speaker, better viewed with zoom in) suggest a change in conversation between the two speakers, whose analysis would aid understanding speaker swap mid conversation. 92

5.8 Adaptation of TalkNet [1] architecture to include body data and serve as our baseline: inclusion of a visual temporal encoder for body data and cross-attention for audio and body. 93

5.9 WASD group training performance (mAP) influence on: a) Proposed Baseline; and b) TalkNet with face as input. AVA refers to TalkNet trained in AVA-ActiveSpeaker, and WASD refers to training using all training data. . 94

6.1 Illustration of the **BIAS insight**: in surveillance settings, where facial and audio-based features might not be always available, body data should be crucial to accurately detect the active speakers. In such challenging conditions, providing reliable explanations for the reasoning behind the provided responses is also an important feature. This paper describes BIAS, which singularly uses facial, audio, and body-based features, also providing visual interpretability and feature importance assessment for its responses. 98

6.2 Overview of the BIAS architecture and pipeline, with GPT model integration: body and face-based data is fed into the respective visual encoders, while audio is processed into MFCC before encoding. SE blocks are used in visual encoders and feature combination for attention heatmaps and feature relative importance, respectively. SA refers to self-attention blocks. Heatmaps are created by combining channel features of the respective top 10% SE vector values. BIAS prediction is based on feature combination, accompanied with visual interpretability and feature importance assessment, complemented by text descriptions from a GPT model finetuned in ASD-related actions data (ASD-Text). 100

6.3 SE vector values from feature (audio, body, and face) combination, for BIAS trained in AVA-ActiveSpeaker and WASD. Both datasets follow a normal distribution. AVA refers to AVA-ActiveSpeaker. 101

6.4 Comparison of face and body area, relative to image dimension, in percentage. AVA-ActiveSpeaker contains data with subjects closer to the camera, expressed by higher face and body percentage, relative to WASD and any of its categories. Surveillance Settings (SS) is the category with further distance of subjects from camera. 106

6.5 Performance of BIAS, TalkNet, and BIAS$_F$ relative to Head-Body Proportion (HBP) in WASD, across 5 equidistant intervals based on minimum (0.1) and maximum (0.7) HBP. BIAS$_F$ refers to BIAS with only face as visual input. 108

6.6 Context, body and face crops, and respective BIAS attention heatmaps of various scenarios. We group the examples into 4 overlapping sets: 1) face available; 2) occluded or non available face; 3) focus on ASD-body movements, and 4) ignoring body information. 109

6.7 Variation of body movement attention in different time lapses of the same context. Although sudden hand movements related to speech are considered by BIAS, subtle ones are not perceived as important, within the same challenging set. 109

6.8 Relevance of audio, body, and face-based features for ASD on the five different categories of WASD. Hard-BIAS gives higher audio importance in audio impaired categories, perceives body importance similar to WASD-BIAS for the most body-reliant scenario (Surveillance settings) and gives, overall, less face importance to face-dependent contexts, relative to WASD and Easy-BIAS. 110

6.9 Influence of the different features on BIAS performance (mAP): *left*) Effect of audio, face and body absence in WASD categories; *middle*) Feature variation influence relative to different training and testing sets (Easy, Hard, and WASD); and *right*) Feature effect on AVA-ActiveSpeaker (AVA) and WASD categories. 110

6.10 Distribution of ASD-Text dataset actions (left) and pair of actions more commonly associated (right). Action labels refer to their numbers in Table 6.1. 112

6.11 ASD-Text dataset finetuned ViT-GPT2 model predictions in representative images of the 14 considered ASD-related actions of Table 6.1. Each figure number relates to the respective ASD-action number. ViT-GPT2 predictions contain the main caption and a related one (other). 112

6.12 Combination of SE visual interpretability with ViT-GPT2 model subject description in various scenarios with varying ASD challenges. Bold descriptions are the main prediction of ViT-GPT2. 113

7.1 Overview of the inclusion of SE blocks for model-agnostic interpretability. Two different settings are depicted (multi-modal, on the left, and image, on the right) where SE block is similarly included in feature extractors to output attention heatmaps of models predictions. Heatmaps are created using channel features of the respective top 10% SE vector values, via channel interpolation and combination with the original image. Audio encoding of the multi-modal setting (ASD) is not displayed for simplicity. 117

7.2 Overview of a Squeeze-and-Excitation block, where input U is squeezed, for each channel (u_c, in (1)), via average pooling through spatial dimensions $H \times W$ to output a vector of weight importance (s, in (2)), used to highlight the most important feature via channel-wise multiplication of weight importance vector and respective channel of the inputted feature map. Retrieved from the original paper [2]. 118

7.3 SE value distribution of all considered models for the datasets CIFAR-10, CIFAR-100, CelebA, AVA, and WASD (from left to right). SE values are normalized with mean to 0. 119

7.4 Inclusion of SE block in ResNet50 (left), and overview of how standard ASD models perform (right). 119

7.5 SE Interpretability (last column per original image) relative to other standard visual interpretability approaches: GradCAM, GradCAM++, EigenGrad, and FullGrad. For each original image the top row relates to interpretability of ResNet50 and the bottom to InceptionV3. 121
7.6 SE Interpretability of ResNet18 for different CelebA attributes. Each column after the original image refers to attributes more commonly located to specific parts of the face: top, middle, and bottom, respectively. For each original image the top row relates to interpretability of our approach (SE-based) and the bottom to FullGradCam. 122
7.7 Attention of Light-ASD for facial cues, via SE blocks, in AVA-ActiveSpeaker data. 122
7.8 Body attention via SE blocks of WASD Baseline in challenging sets of WASD. 124
7.9 Accuracy with top % of SE channels used in CIFAR-100, relative to using all available feature channels. 125

8.1 State-of-the-art ASD models solely rely on facial cues as visual input to perform. This approach is only reliable with cooperative (subjects) conditions, increasing uncertainty in model prediction in more challenging settings. Our approach aims to complement face with body cues to create more robust models that are able to perform in cooperative and unconstrained scenarios. 128
8.2 Overview of the ASDnB architecture. Body and face data are fed into the visual encoder, allowing intra encoder feature fusion and complement, while audio is processed through its respective encoder. Audio and visual features are combined to predict subject speaking label, using an adaptive weighted loss for combined and visual features ($\mathcal{L}_{av}$ and $\mathcal{L}_{v}$, respectively). 130
8.3 Overview of the flow of face and body combination in ASDnB visual encoder. The first convolution for both inputs downsamples via stride. 131
8.4 Bidirectional GRUs of visual and audio combination in ASDnB classifier, before inputting to FC layers for speaker classification. 132
8.5 Comparison of ASDnB performance relative to ASD state-of-the-art models for (a) number of faces per frame and (b) various face sizes in AVA-ActiveSpeaker. 138
8.6 Relative body importance of ASDnB and Light-ASD with decremental head-body area proportion in AVA-ActiveSpeaker. 139
8.7 ASDnB and Light-ASD (Light) qualitative performance assessment in challenging scenarios of AVA-ActiveSpeaker and WASD. Red bounding boxes denote model prediction of subject not talking, green to speaking, and predictions with red cross denote missclassification relative to the ground-truth. In both examples with subjects far from the camera, Light-ASD misclassified the switch of speakers while ASDnB was more resilient by analysing the hand movement that preceded subject speaking. 140

List of Tables

2.1 Comparison between the current strategies and the considered approach for Human Behavior analysis. 11

2.2 State-of-the-art results in AVA-ActiveSpeaker (AVA) and categories of WASD. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. Adapted from [3] and [4]. Best performance in bold and second best underlined. 17

2.3 State-of-the-art results in Kinetics 400 (K400) and AVA Actions (AVA) using Top-1 accuracy and the mAP metric, respectively. SSv2 refers to Something Something-v2 [5] dataset. 21

2.4 State-of-the-art results in MOT17 dataset. 25

2.5 State-of-the-art results in HICO-Det. TS, OS, and T indicate Two-Stage, One-Stage, and Transformer, respectively. F, R, and NR refer to Full, Rare, and NonRare, respectively. 29

2.6 State-of-the-art results in KITTI at a depth of 80m. Results extracted from [6]. 32

2.7 Clothes Detection (mAP_{box}) and Landmark Detection (mAP_{pt}) in DeepFashion2. 35

2.8 COCO.TD and BU refer to Top-down and Bottom-up approaches, respectively. 39

2.9 MPII.The table is divided by the type of approach used, the top refers to Heatmap and bottom to Regression. 39

2.10 State-of-the-art results in multiple PAR datasets. 44

3.1 Gender mA of different models trained and evaluated on various PAR datasets. Models named with *Gender* suffix refer to models trained solely for the gender attribute. The outperforming methods for each dataset are shown in bold. 57

3.2 Dataset image-based feature analysis. Values are normalized for the combination of the three dataset values, for each feature. 57

3.3 Image and subject-based feature values for PETA, PA-100K, and RAP datasets. *Correct* represents the correctly inferred images (when all models inferred the correct gender value) and *All* refers to all test data. Subject pose values represent the portion of test images with the given pose. 58

3.4 Comparison of our face-based gender model, in the Adience dataset. 61

3.5 Gender mA of face-based gender model, trained and evaluated in within and cross-domain settings. Only frontal face images of PETA, PA-100K, and RAP datasets were considered. 62

3.6 Gender mA of face-based gender model, trained for *Any* and *Frontal* pose images, and evaluated in frontal image faces (*Face*) and full-body images of corresponding pose (*Body*). *Any Pose* refers to all dataset images and *Frontal Pose* images refer to frontal images. `FI` stands for Face Importance. 62

4.1 Dataset image-based feature analysis. Values are normalized for the combination of all dataset values, for each feature. Datasets named with *Frontal* suffix refer to datasets composed solely of frontal images. 73

4.2 Number of images, test/all image number ratio for each dataset, and frontal/PAR image number ratio, for train and test set. Frontal datasets are represented with the subscript term *Frontal* and test/all ratios are represented by Dataset Ratio. 74

4.3 Gender mean Accuracy (mA) of different models on frontal PAR datasets. The outperforming method for each dataset is shown in bold. 75

4.4 Gender mA of different models on PAR datasets. The outperforming method for each dataset is shown in bold. 75

4.5 Gender mA of YY-Net in frontal datasets, when trained in PAR datasets and in its frontal version. Frontal datasets are represented with the subscript term *Frontal*. The highest evaluation accuracy for each dataset is shown in bold. 76

4.6 Performance comparison, on $\text{PETA}_{\text{Frontal}}$ dataset, when gradually adding each component to the baseline model. Variants of the same key parts lie in the same group. Components in bold are the adopted approach in YY-Net. 79

5.1 Feature comparison of ASD datasets. AVA-ActiveSpeaker is represented as AVA. If datasets contain information regarding a feature, its absence is presented with $\times$, while its presence with ✓. WASD has a high number of hours, with increased number of faces and reduced face tracks (culminating in higher average video duration), Frames Per Second (FPS) variability (Var), and increased talking percentage. The most discriminative factors are demographic representation (DR), surveillance conditions (SC), and body data annotations. 82

5.2 Category feature matrix. Feature description: FA, Face Availability; SO, Speech Overlap; DS, Delayed Speech; FO, Facial Occlusion; HVB, Human Voice as Background Noise; SS, Surveillance Settings. The absence of a certain feature is presented with $\times$, while its presence with ✓. Features containing ? refer to non-guarantee of its presence or absence. Green cells refer to features favorable for ASD, while red ones are unfavorable. 85

5.3 Comparison of AVA-ActiveSpeaker trained state-of-the-art models on AVA-ActiveSpeaker and categories of WASD, using the mAP metric. We train and evaluate each model following the authors' implementation. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. AVA refers to AVA-ActiveSpeaker, Light refers to Light-ASD, and TS-Talk to TS-TalkNet. 88

5.4 Comparison of state-of-the-art models on the different categories of WASD, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. Light refers to Light-ASD, and TS-Talk to TS-TalkNet. 88

5.5 Comparison of TalkNet, using face and body data, and proposed Baseline on WASD categories, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. 93

5.6 Comparison of F1-Score (%) on the Columbia dataset. 95

6.1 ASD-Text dataset action labels, grouped by body parts. 102

6.2 Category feature matrix. Feature description: FA, Face Availability; SO, Speech Overlap; DS, Delayed Speech; FO, Facial Occlusion; HV, Human Voice as Background Noise; SS, Surveillance Settings. The absence of a certain feature is presented with $\times$, while its presence with ✓. Features containing ? refer to non-guarantee of its presence or absence. Green cells refer to features favorable for ASD, while red ones are unfavorable. Retrieved from [3]. 103

6.3 BIAS and state-of-the-art models performance on the different categories of WASD, grouped by Easy and Hard, using the mAP metric. 105

6.4 Comparison of BIAS and state-of-the-art models on the AVA-ActiveSpeaker, grouped by the visual encoder used. Models with $*$ customized the reported backbones. 105

6.5 Comparison of F1-Score (%) on the Columbia dataset. 106

6.6 Ablation studies on the effect of SE for feature combination and the contribution of different features towards BIAS performance (mAP) in AVA-ActiveSpeaker (AVA) and WASD. 107

6.7 BIAS performance using different backbones in AVA-ActiveSpeaker (AVA), while maintaining the remaining of the architecture. 107

6.8 BIAS performance (mAP) in AVA-ActiveSpeaker (AVA) and WASD, with varying body inputs and as the only model input for ASD. 107

6.9 Effect of group training, in mAP, on the categories of WASD. Easy group training contains data of Optimal Conditions and Speech Impairment, while Hard group contains data of the remaining WASD categories. 111

6.10 ViT-GPT2 performance increase with finetune on ASD-Text dataset. Metric description: RL, ROUGE-L; M, METEOR; B1-4, BLEU1-4. 112

7.1 SE inclusion performance (Acc) influence on CIFAR-10, CIFAR-100, and CelebA. Results on CelebA refer to gender attribute prediction. Par(M) refers to the number of parameters in millions. 121
7.2 SE inclusion performance (mAP) influence on AVA-ActiveSpeaker and WASD. AVA refers to AVA-ActiveSpeaker and Par(M) refers to the number of parameters in millions. 121
7.3 ResNet18 performance (Acc) on different attributes of CelebA. 122
7.4 Comparative evaluation in terms of Deletion (lower is better) and Insertion (higher is better) scores of a ResNet50 on CIFAR-10 and CIFAR-100 datasets.124

8.1 Comparison of ASDnB with state-of-the-art models in AVA-ActiveSpeaker. 135
8.2 Comparison of ASDnB with state-of-the-art models on the different categories of WASD, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. 135
8.3 Comparison of F1-Score (%) on the Columbia dataset. 136
8.4 Ablation studies on the effect of WASD pretraining, face and body influence towards ASDnB performance (mAP) in AVA-ActiveSpeaker. 136
8.5 Variation of audio and visual encoders regarding the number of parameters and model performance in AVA-ActiveSpeaker. 137
8.6 Loss effect on ASDnB performance (mAP) in AVA-ActiveSpeaker. I_v refers to the importance of visual features, while I_{av} refers to the importance of combined audio and visual inputs towards ASD prediction. All approaches have Cross-Entropy has the underlying training loss. τ refers to the temperature coefficient. 137
8.7 Performance of temporal modeling methods in ASDnB classifier. 138

Acronyms and Abbreviations

ALM	Attribute Localization Module
AP	Average Precision
ARM	Attribute Re-weighting Module
ASD	Active Speaker Detection
A.Rel	Absolute Relative Error
BCELogits	Binary Cross Entropy with Logits
BGRU	Bidirectional Gated Recurrent Units
BMI	Body Mass Index
BN	Batch Normalization
CAM	Class Activation Maps
CNN	Convolutional Neural Network
CVAT	Computer Vision Annotation Tool
COCO	Common Objects in Context
ConvLSTM	Convolutional LSTM
CSV	Comma-Separated Values
DMTL	Deep Multi-Task Learning
DPM	Deformable Part-based Model
DS-Conv1D	Depth-wise Separable Convolutional Layers
EEG	Electroencephalogram
FAM	Face Attention Module
FC	Fully Connected
FI	Face Importance
FIW-BMI	Face In the Wild for BMI analysis
FPS	Frames Per Second
GCN	Graph Convolutional Networks
GT	Ground Truth

GRU	Gated Recurrent Units
GUI	Graphical User Interface
HBP	Head-Body Proportion
HICO	Humans Interacting with Common Objects
HICO-Det	Humans Interacting with Common Objects Detection
HOI	Human-Object Interaction
HOTA	(Higher Order Tracking Accuracy
IMDB	Image Movie DataBase
IoU	Intersection over Union
KP	Keypoints
LFW	Labeled Faces in the Wild
LSTM	Long Short-Term Memory
mA	mean Accuracy
mAP	mean Average Precision
MFCC	Mel-Frequency Cepstral Coefficients
MOT	Multi-Object Tracking
MOTA	Multiple Object Tracking Accuracy
MPII	Max Planck Institut Informatik
MTCNN	Multi-Task Cascaded Convolutional Neural Networks
OKS	Object Keypoint Similarity
PAR	Pedestrian Attribute Recognition
PCK	Percentage of Correct Keypoints
PR	Precision-Recall
Re-ID	Re-identification
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROI	Region of Interest
ROC	Receiver Operating Characteristic

RoR	Residual Networks of Residual Networks
SE	Squeeze-and-Excitation
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
STN	Spatial Transformer Network
SVM	Support Vector Machine
S.Rel	Squared Relative Error
UT	University of Texas
VCOCO	Verbs in Common Objects in Contex
V-TCN	Video-Temporal Convolutional
WASD	Wilder Active Speaker Detection
YOLO	You Only Look Once
YY-Net	YinYang-Net

Chapter 1

Introduction

This thesis regards the underlying challenges of assessing human behavior, how deep learning approaches can be used towards this goal, and the current limitations of state-of-the-art to perform in conditions where human behavior analysis is most needed: wild (surveillance-like) conditions. We make our key contributions in highlighting the limitations of state-of-the-art and propose improvements of two behavioral features topics: soft biometrics and verbal interaction (Active Speaker Detection). In this chapter, we complement our contributions' discussion with problem definition, work relevance and motivation, research objectives, thesis work plan and organization.

1.1 Problem Definition

Human behavior can be defined by a set of action(s) and reaction(s) one has in response to their environment, which differs between individuals based on numerous variables such as cultural background, ethnicity, past experiences, traumas, among others. With such variability, analyzing human behavior is a very complex task, which is corroborated by the methodologies used in the area more closely linked to behavioral assessment: Psychology. In this area, studies and experiments are done in very controlled/restrictive settings and conclusions drawn are specific to the experiment context (*i.e.*, hard to extrapolate to other settings), which is mainly due to the amount of variables needed to control to compare behaviors of different people. Furthermore, studies' results are tied to observations from experts and survey completion (*i.e.*, answer a set of questions) either from observers and/or participants, and subjects analyzed know that they are in experiments which may condition/affect their natural reaction. Given the restrictive setting where human behavior is typically assessed, we believe it is important to **explore other ways to study this topic given the importance of human behavioral analysis** in today's society.
One of the major limitations of Psychology-based methodologies is the controlled conditions in which subjects' behavior is analyzed. To fully explore the potential of human behavior assessment, it is important to extend this topic to more unconstrained/wild conditions, contrasting with the standard Psychology methodology of controlled scenarios and subject cooperation. The ideal setting for this improvement is surveillance-like data, where subject actions/behavior is unconstrained, uncooperative, and not conditioned to specific experiments. Although this type of data is ideal for human behavior assessment in uncooperative conditions, current Psychology-based methodologies are not applicable in such settings. As such, one of the current state-of-the-art limitations is how to **combine deep learning approaches, surveillance-like data and Psychology methodologies** to assess human behavior in wild conditions.

One potential approach for deep learning in human behavior analysis is to mimic (to an extent) how Psychology studies assess human behavior: via grid of behavioral features. To achieve this, it is necessary to define a set of features important for human behavior that can be assessed via surveillance cameras and without requiring subject cooperation which, although not applicable to obtain intrinsic motives from people behavior, can be used to profile subjects based on their behavioral features. **The biggest challenge in this topic is to define which set of behavioral features** are suited (and which are not) for deep learning models to assess in surveillance-like settings.

1.2 Work Relevance and Motivation

One of the main motivations of our work relates to the restrictive context where human behavior is typically analyzed. Analyzing human behavior is extremely complex, which justifies the grid-based approaches and controlled experimental settings where Psychology generally operates, but it is very time and resource consuming. It requires expert analysis, environment control and, most importantly, subject cooperation for analyzing human behavior in such conditions. As such, our work proposes an alternative strategy to overcome some of the existing limitations by analyzing behavioral features without requiring subject cooperation to obtain behavioral subject profiling. In particular, using behavioral subject profiling in surveillance-like conditions it would be possible to identify abnormal behaviors related to malicious intent or detect subjects in need of aid, assess emergency conditions, or assess customer service/preferences in shop/mall settings. Note that with this proposal we do not intend to replace behavioral specialists, but rather rather to serve as complement to them in cases where intrinsic motive analysis is required, such as assessing potential scholar dropouts, depressive-like behavior, among others.

The ideal setting for our proposal is surveillance-like data, where subject actions/behavior is unconstrained, uncooperative, and not conditioned to specific experiments. Although this type of data is ideal for human behavior assessment in uncooperative conditions, current Psychology-based methodologies are not applicable in such settings. One major advantage of this type of data is that it is widely available for different tasks, namely PAR, Re-identification (Re-ID), action recognition, among others, where deep learning models performance have improved gradually over the years. This highlights a potential point of convergence between deep learning models and human behavioral analysis which is motivated by our proposal of assessing behavioral features for subject profiling.

Defining a set of features available for human behavior analysis that can be applied in surveillance-like conditions is an important aspect of our proposal, which was only achieved via extensive Psychology literature review and discussion with area experts of characteristics to confirm their influence in human behavior. Among the various features considered, we assessed that some of the existing approaches can already be applied to profile subject which shows that our proposal can serve as motivation for future works/companies to apply current deep learning approaches towards this goal. Additionally, since not all the features are currently explored/applicable in unconstrained settings, this also highlights

the room for improvement in various Artificial Intelligence and Deep Learning areas. In particular, most of our work focused on improving the area of soft biometrics and ASD since the experimental settings of the state-of-the-art were mainly in cooperative conditions. Similar to our work, the behavioral analysis proposal can be used to motivate future researchers to extend the application of existing models/datasets to more uncooperative settings.

1.3 Research Objectives

Our proposal started by defining a set of behavioral features analyzable using deep learning approaches in surveillance-like data to remove the requirement of subject cooperation and environment restriction of Psychology-based methodologies. It is important to highlight that a full subject profiling of all the proposed behavioral features is an extensive task, beyond the scope of this PhD thesis. Rather than working on improving all the features we selected two key ones where their use was mainly linked to cooperative settings, with limited/unproven application in uncooperative conditions: soft biometrics and ASD, with greater incidence on the latter topic based on the limitation of existing datasets. Given our goals, we define the following research objectives:

- Define a set of behavioral features, supported by Psychology studies, that can be applied to profile subjects in wild conditions, without requiring subject cooperation;
- Explore the limitations of state-of-the-art deep learning models to assess soft biometrics in surveillance conditions, using PAR data;
- Propose a state-of-the-art model to assess gender in PAR data, which can be extended to other soft biometrics;
- Propose a new dataset for ASD (WASD) that contrasts with existing datasets by having challenging and surveillance-like data, highlighting the limitations of current ASD models to perform in non-cooperative conditions;
- Propose a new ASD model (BIAS) that, for the first time, combines body, face, and audio for ASD, provides visual interpretability for its decision, and is able to perform in cooperative and challenging conditions;
- Propose an improved model for ASD (ASDnB), based on the premise of body, face, and audio combination, that has improved results by intra visual encoder combination of dual inputs (face and body) and training with adaptive weight feature importance to effectively combine relevant features for robust ASD.

1.4 Thesis Work Plan

Based on the research objectives defined, we followed the work plan in Fig. 1.1, with the tasks' description in Fig. 1.2.

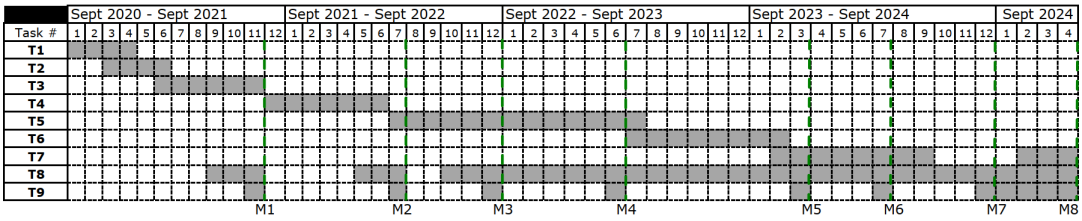


Figure 1.1: Timeline of tasks, with the respective milestones. Each milestone corresponds to a scientific contribution, whose descriptions are in Fig. 1.3.

Task #	Description
T1	Definition of behavioral features retrievable from surveillance scenarios, based on psychology literature
T2	Review of the existing deep learning approaches, datasets and metrics for behavioral features
T3	Comparison analysis of state-of-the-art methods for soft biometrics, in wild conditions
T4	Implementation of a method, based on the analysis, to improve soft biometrics evaluation
T5	Data gathering and annotation for ASD dataset containing uncooperative data
T6	Implementation of a method to improve ASD, in particular, for uncooperative conditions
T7	Improvement of ASD model to reliably performance in varying conditions (cooperative and surveillance)
T8	Writing for publication of gathered knowledge in conference papers/journals
T9	PhD thesis writing and collaboration in projects

Figure 1.2: Description of the proposed tasks. Task numbers refer to the numbers presented in Fig. 1.1.

The first-year tasks were important to define the path of this PhD work. The first task (T1) proposes a pioneering approach to evaluate human behavior. Instead of relying on action or emotion recognition, we focus our attention on defining a set of features that can provide an in-depth analysis. This was achieved through multiple interactions with Psychology investigators, culminating in our attempt to mimic the approach used by humans to evaluate human behavior: *grid-based approaches*. In these cases, the environment is extremely controlled, and a set of features, specifically targeted to evaluate a given behavior, are assessed. The features' choice to evaluate behavior in a wilder context required the Psychology investigators' cooperation and the support of Psychology literature. We complement our analysis by compiling relevant deep learning methods, datasets and metrics usable for each behavioral feature (T2).

After this step, we decided to target gender, which was one of the behavioral features derived from our literature research. We moved to our third task (T3), where we performed an extensive analysis of the factors influencing gender misinference in wild conditions. This proved to be an important work, given that most (specific) gender inference is typically made in face datasets. Our analysis focused on wilder scenarios, targeting PAR datasets, whose focus is to infer attributes from surveillance images. We adapted several PAR methods to only infer gender and displayed the conditions/factors more influential for wild gender inference.

The following years' task was based on the findings of our third task, where we propose our own method (T4) to improve wild gender inference, targeting the combination of face and body information to achieve our goal, extendable to evaluate other soft biometrics. This work highlighted one of the limitations of current gender approaches that mainly focus on cooperative images instead of surveillance-like conditions. After successfully contributing to the body of knowledge for one of the proposed behavioral features we explored the limitations of another feature: ASD. Given the state-of-the-art ASD datasets, that focused

on more cooperative settings, and existing models that base their performance on face and audio information, we decided to propose a new dataset to mitigate the limitations of the area. To achieve this goal we gathered data containing uncooperative subjects, face occlusion, audio impairment, subjects variability (language, ethnicity, etc.) from available YouTube data: WASD (T5).
After successfully gathering a wilder dataset that showed the limitations of current models, we focused on proposing a new model that is able to deal with such conditions, on the later part of the third year (T6). The main key improvement is the inclusion of body data information which aids in scenarios where face is not reliably accessed. Although this showed improvements in WASD the results in the state-of-the-art ASD datasets (more cooperative settings) could be improved. This motivated us to propose a model that performs in both cooperative and unconstrained scenarios mainly due to face and body combination intra encoding for adequate feature assessment, resulting in state-of-the-art performance in varying conditions (T7).
Throughout the years, we wrote this PhD thesis continuously at each successful milestone reach, with a more significant portion of time dedicated to its writing in the latter stages of our work, translated by task 9 (T9). Additionally, since throughout the years we collaborated with other projects and works that resulted in publications, we dedicated significant portions of the work plan to writing and/or reviewing papers for publication (T8). The proposed work plan also displays a set of milestones, whose description is exhibited in Fig. 1.3.

#	Milestone
M1	Publication on IEEE Transactions on Biometrics, Behavior, and Identity Science
M2	Publication on IEEE Access
M3	Publication on IEEE Access (co author)
M4	Publication on Research and Practice in Technology Enhanced Learning (co author)
M5	Publication on IEEE International Workshop on Information Forensics and Security
M6	Publication on IEEE Access (co author) Publication on IEEE Transactions on Biometrics, Behavior, and Identity Science
M7	Publication on Explainable & Interpretable Artificial Intelligence for Biometrics
M8	Publication on IEEE Transactions on Biometrics, Behavior, and Identity Science

Figure 1.3: Description of the achieved milestones. Milestone numbers refer to the numbers presented in Fig. 1.1.

1.5 Main Contributions

One of the main contributions of this thesis is the definition of behavioral features that can be analyzed via camera for behavior profiling, and that have been proven as influential towards human behavior. We achieve this via discussion with Psychology researchers and extensive review of Psychology literature to clearly define seven major behavioral features (Figure 1.4), which we discuss in Chapter II. There, we also explore the areas of deep learning more closely related to each feature and existing applications/models usable to assess each of them. The reasoning for filtering all potential behavioral features into only

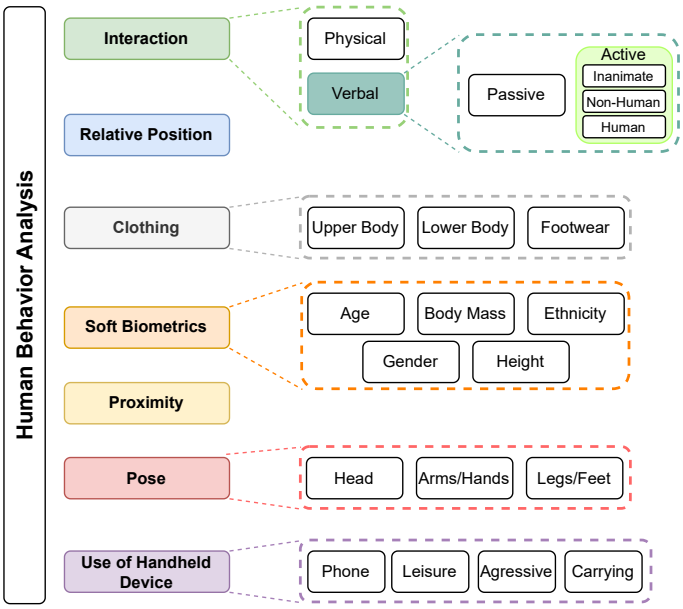


Figure 1.4: Main features for human behavioral analysis via automated vision-based approaches.

seven is that these are the most mainstream topics explored in the Psychology literature review, thus our focus on these. Note that each feature also has ramification to other (sub)features, and our proposal of assessing human behavior via the profiling of these seven features does not limit the inclusion of others, as long as they are dully supported by Psychology literature. As such, future works can further contribute towards this behavioral feature grid by extending it with other features. Furthermore, this profiling can be used to give more meaning to other/future works by complementing their motivations with behavioral analysis (*i.e.*, someone working on PAR can extend their motivation to use in behavioral analysis). With our work in Chapter II, we also highlight the areas currently less explored, helping future researches identifying potential areas of work as well as motivating current works to explore their applicability in more wilder (surveillance-like) conditions.

Our first publicly available contribution relates to the limitations of current models to assess gender (soft-biometrics) in wilder conditions. Soft biometrics is extensively explored but most of the works are done in cooperative settings and not in wilder conditions, in particular surveillance settings-like (*e.g.* PAR data). Although the analyzed PAR models/datasets are not exclusively used for gender, they are the ones more closely related to wild gender inference, which also highlights the limitations of state-of-the-art in wilder/uncooperative biometric assessment. As such, the novelty of our works is exploring the main limitations of current approaches for gender inference in these wilder settings (Chapter III), and proposal of a new model based on these findings (Chapter IV), using PAR data as representative of surveillance conditions.

The following contributions relate to the ASD topic. We found out that ASD also contains similar limitations to the ones found in our first two works: assessment of ASD is done in cooperative settings (Hollywood-like movies), with AVA-ActiveSpeaker. Given our goal to analyze human behavioral features in uncooperative conditions (surveillance-

like scenarios), we need to assess how existing approaches are able to perform in wilder scenarios. However, contrary to the first two works, there were no available data that represent such conditions, which made us propose a wilder ASD dataset (WASD, in Chapter V) to highlight the limitations of existing ASD approaches. We divided WASD data into five categories, with varying image and audio quality, which highlighted the inability of current models to perform in impaired quality of audio and with limited/absent face access. We also found out that in such scenarios, complementing body with traditional ASD data (audio and face) leads to performance improvement, given the access to information besides facial data such as non-verbal communication cues, particularly important in uncooperative conditions without reliable face access (*i.e.*, surveillance settings).
Based on the existing limitation of the state-of-the-art, our next contribution is the proposal of a new ASD model (BIAS, in Chapter VI) that, for the first time, combines face, audio, and body information to reliably perform in cooperative (AVA-ActiveSpeaker and Columbia) and challenging (WASD) conditions. Furthermore, given the lack of explanation of deep learning models behind their decision, we focus on providing visual interpretability by a novel use of SE blocks that identify the visual regions more influential towards a model decision based on its (SE) feature importance vector. Finally, we complement our interpretability model with a textual description dataset of ASD-related actions, given the lack of available data for these tasks in the literature. We also complement the results of this work by extending the applicability of SE-based interpretability to various model and datasets, in different contexts (image, video/multi-modal), achieving competitive results with GradCAM variants without their drawback/limitations (Chapter VII).
Our final contribution is the proposal of an improved ASD model (ASDnB, in Chapter VIII), based on the same premise of face, audio, and body information. Our new approach is to conjugate face and body inputs intra visual encoder, combining them at different steps of the extraction, which contributes to a complement of facial cues with relevant body information, translating into state-of-the-art performances in both cooperative (AVA-ActiveSpeaker and Columbia) and challenging (WASD) conditions.

1.6 Thesis Organization

This thesis consists of a collection of papers, meaning that the content of each chapter (except Chapters 1, 2, and 9) relates to a paper published in a journal, conference proceedings, or currently under review. The remainder of this thesis is organized as follows: Chapter 2 presents the features we propose for human behavior analysis in wilder conditions and without requiring subject cooperation, with a review of existing approaches applicable to assess each feature and discussing the areas worth exploring given state-of-the-art limitations. Chapter 3 extensively discusses the limitations of existing methods for biometric evaluation, in particular gender, in uncooperative scenarios, highlighting the unreliability of most existing gender approaches for surveillance-like conditions. Chapter 4 presents a new approach for wild gender inference that complements body with face information for resilient gender classification in wilder conditions, where face may not always be reliably

accessed. Chapter 5 presents a new ASD dataset (WASD) that contrasts with existing ones by containing data with impaired audio and image quality, highlighting the limitations of current approaches to perform in wilder conditions (surveillance scenarios). This is mainly due to the existing data available, where the state-of-the-art (AVA-ActiveSpeaker) allows ASD via audio and face assessment given the cooperative settings and good quality data available. Chapter 6 proposes BIAS, a model that explores the importance of completing body information with audio and face for ASD, in particular for surveillance conditions where face access is not guaranteed and audio quality is subpar. This model is able to perform in cooperative and challenging settings and also provides interpretability in decisions with visual representation of the key areas of interest via a novel use of SE blocks. Chapter 7 further explores model interpretability using SE blocks by assessing its use in different models and context (image and video/multi-modal). The extensive experiments corroborate the relevance of SE-based interpretability given its application to traditional and custom model with competitive results relative to traditional interpretability approaches such as GradCAM variants. Chapter 8 further extends the discussion of body inclusion importance for robust ASD with the proposal of an improved model, ASDnB. By combining face and body inputs intra visual encoder, complemented with pretraining in challenging data, ASDnB is able to achieve state-of-the-art performances in both cooperative and challenging conditions, serving as basis for robust ASD models for future works. Finally, chapter 9 discusses the main conclusions of the work presented in this thesis, summarizing the key achievements and discussing future directions.

Chapter 2

Human Behavior Analysis: A Survey

It is an incredibly complex task to understand human behavior at its fullest: action recognition, intrinsic motives, and context evaluation are amongst the variables needed for complete behavior analysis. Current deep learning-based approaches focused on human behavior analysis are either too high-level, failing to capture intrinsic motives, or require human cooperation and a constrictive testing environment. In this survey, we introduce a novel behavioral analysis perspective that focuses on analysing features associated with human behavior, such as interaction, relative position, clothing, proximity, pose, handheld devices, and soft biometrics. The selected features are assessed via deep learning approaches that can be applied in wild conditions, without requiring human cooperation. We support the choice of behavioral features using psychology studies, and discuss the most relevant deep learning areas, highlighting the state-of-the-art methods and datasets available, limitations and future directions.

2.1 Introduction

Human behavior analysis difficulty lies in how diverse people are, what background one has, and how diverse the behavioral context might be, among other aspects [7]. The same person might react differently to similar scenarios if the context is slightly different. As such, in psychology, to analyze human behavior, it is common to have extreme control of the action context. This approach usually requires human cooperation, and behavior interpretation is based on behavior-specific grids, filled with a set of aspects intended to analyze [8]. Furthermore, for better behavior interpretation, study participants might be asked to fill self-report surveys, with a set of questions designed to tackle a specific aspect [9].

With the advance of deep learning and computer vision technologies, good results have been reported in action and object recognition. Following this success, human behavior analysis has started to gain attention, with current approaches being mainly based on action recognition [10, 11, 12], which does not translate into complete human behavior understanding. Other approaches that evaluate human behavior from other perspectives do so in controlled environments and rely heavily on human cooperation [13, 14]. Given the advances and potential of vision-based methods, it is desirable to structure a different behavioral analysis approach that uses deep learning methods without requiring environmental constraints and human cooperation. This survey suggests an approach that combines behavior-specific grids' granularity, using behavioral influential features, and the non-cooperative aspect of computer vision-based methods. The key features proposed to analyze are presented in Figure 1.4, in Chapter 1. Furthermore, we support the selection of

each feature using psychology studies, demonstrating its importance/influence in human behavior, and discuss the most relevant deep learning areas of each feature to provide an overview of the work done and highlight directions for future works. In sum, the key contributions are:

- We introduce a new human behavioral approach based on behavioral features analysis via deep learning approaches, applicable in non-cooperative settings, which is distinctive from action-based and emotion-based approaches;
- We review psychology studies and present the main ideas that support the selection of behavioral features used for human behavior analysis;
- We provide reviews of deep learning areas related to each behavioral feature, highlighting the methods and datasets used, and discuss the limitations and future directions of each area towards behavior analysis.

The remaining of the chapter is structured as follows: Section 2.2 reviews the works and surveys of current human behavior analysis perspectives and compare them with the proposed approach; Section 2.3 refers to the considered behavioral features, each containing a subsection related to Psychology background, with work that supports the choice of the feature, and subsections related to deep learning methods, datasets and metrics; Section 2.4 reviews the main limitations and future directions of deep learning areas regarding the considered human behavioral analysis; and Section 2.5 summarizes the main topics of this chapter.

2.2 Human Behavior Analysis Approaches

Behavior is determined by psychological factors, named attitudes, divided into emotional, cognitive, and behavioral components [15]. According to Katz and Stotland [16], *attitude is a tendency or predisposition to evaluate an object or symbol of that object in a certain way*. Thus, although we can not see attitude, we can see behavior, the expression of an attitude. However, the way we analyze behavior may vary in the Deep Learning area, which we discuss in the following sections.

2.2.1 Behavior Analysis as Action Recognition

The most high-level approach to analyze human behavior is through action recognition. Hao *et al.* refer to their work as *Human Behavior Analysis Based on Attention Mechanism and LSTM Neural Network* [10] when they are evaluating specific actions such as hugging, sitting, handshaking, among others. Another example is the work done by Zhigang and his colleagues. They also label their work [11] as human behavior recognition when evaluating bending, squatting, and sitting, using pose estimation methods. Wang *et al.*, with their *Dynamic Human Behavior Pattern Detection and Classification* work [12], evaluate arms and legs patterns in sitting, walking, and standing situations. Additionally,

Table 2.1: Comparison between the current strategies and the considered approach for Human Behavior analysis.

Methodology	Non Cooperative	Intrinsic Analysis	Context Independent
Emotion-Based		×	×
Action-Based	×		×
Psychology		×	
This survey	×	×	×

there are video-based human behavior understanding surveys based on interaction and action recognition [17], as well as action, or activity, recognition based surveys with behavioral analysis references [18]. For more specific action recognition methods analysis, we refer to Zhang *et al.* [19] and Kong *et al.* [20] surveys.
Labeling action recognition as analyzing human behavior is a perspective shared by some authors. With such an approach, it is possible to evaluate what a given person is doing in a scene. However, **this fails to evaluate intrinsic motives, and context analysis is disregarded in some way**. Someone jogging or running away from something is still labeled as the "running" action, disregarding that both actions might have different behavioral meanings.

2.2.2 Behavior Analysis as Emotion Recognition

Analyzing emotions is not as often labeled as *human behavior analysis* but is a theme inherently associated with it. As previously mentioned, emotions are an attitude component and are the most low-level approach for behavior interpretation. Hessler *et al.* proposed, in their survey [21], the evaluation of *Deception*, *Emotion*, *Facial Recognition*, and *Stress Detection* for behavior analysis. Rouast and colleagues [22] summarize the modalities used, in the literature, for affect recognition using deep learning, with the more significant percentage being on facial expression, followed by text analysis. An overview of the current approaches for facial expression recognition is exhibited in Huang's survey [23]. Using Electroencephalogram (EEG) for emotion recognition is also described in several surveys [24, 25]. Using body information for emotion interpretation is also reported in the literature. Shen *et al.* [13] evaluated body movement and correlated it with emotional expression, Chen *et al.* [14] analyzed micro gesture expression in relaxed and stressful situations, while body movement and its emotion recognition association are also explored in other works [26, 27].
Using emotion recognition for behavior analysis allows for a more fundamental human evaluation. However, **this approach requires controlled setups and heavily relies on human cooperation**.

2.2.3 Considered Behavior Analysis

Given the limitations of the two main human behavioral analysis approaches, we suggest a more granular behavior evaluation, able to capture high-level information (like ac-

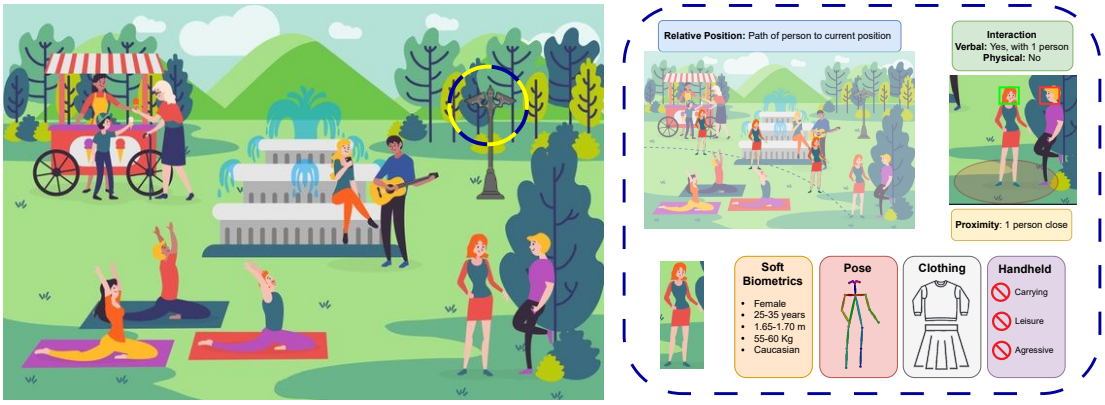


Figure 2.1: Overview of the use of behavioral features for subject profiling. Each feature is analysed from images/video of a surveillance camera in a non-cooperative manner.

tion recognition approaches) and intrinsic motives (similar to emotion recognition approaches), in a non-cooperative way. As such, our approach **focus on assessing specific behavioral features via surveillance cameras, using deep learning approaches to evaluate said features**. Then, we are able to analyze the behavior of individuals via subject profiling, in a non-cooperative manner. The key aspects of this approach are: 1) we are able to obtain subject behavior information without requiring their cooperation; 2) the profiling based on the analyzed features is context independent; and 3) by analyzing specific features to create a subject profile, we are closer to intrinsic analysis and behavioral understanding than via action recognition. An overview of the comparison between behavioral approaches is shown in Table 2.1.

To provide an example of profiling based on the features of the considered approach, we present Figure 2.1, where each of the features is analyzed regarding the female subject on the bottom right of the image. Note that the profiling is done via the surveillance camera and the analyzed features are: relative position, verbal and physical interaction, proximity, soft biometrics, pose, clothing, and presence of handheld devices. One major aspect of this approach is that the **features analyzed must be relevant to human behavior**. As such, we dedicate a subsection to psychology background to support the importance of features, in the sections of each feature.

2.3 Features Considered

The following section is dedicated to discuss the features considered, each in different subsections and structured into three key parts: *psychological background* to justify its choice, existing *deep learning methods* for feature assessment, and current *datasets, metrics and results* of the state-of-the-art.

2.3.1 Verbal Social Interaction

2.3.1.1 Psychology Background

Active Verbal Interaction. Social interaction is a vital aspect to characterize human behavior since human beings are, inherently, social animals [28], and, as such, there is a need to sustain significant interpersonal relationships [29]. Additionally, humans search for a sense of belonging by actively fostering social connections, linked to positive psychological and physical health outcomes [30]. Communication between people is also essential in the context of social engagement and the work environment. Effective communication helps eliminate uncertainty, build trust, and enhance human relation quality [31]. Furthermore, communication between humans has a reciprocal nature [32], and those involved in dialogue tend to engage in two-way communication, with the emitter being responsive to the receiver's needs [33]. Next, we explore the importance of verbal interaction between human, non-human, and inanimate objects.

Human. In social interactions, spoken words are a powerful tool of communication [34]. People use communication to express their ideas, thoughts, emotions, among others. Persuasion of others is also a potential intent of the communication. It has been reported that gestures, eyebrow lifts, and verbal fluency are key behavior factors that increase persuasion effectiveness [35]. Another important aspect is that such verbal persuasion behaviors are less expressed in shy children [36], illustrating an example of talking feature evaluation and its association with a given behavior. In mock interviews, innocent people who were questioned more guilt presumptive questions reported feeling more pressured to confess than those asked neutral questions [37]. Interestingly, listeners of said interviews rated innocent suspects as more guilty than guilty ones, even when both replied to guilt presumptive questions.

Non Human. Evaluating a human's active role in a non-human verbal interaction is mainly applicable to scenarios where people talk to other animals. The most common environment where such is seen is when people talk to their or other people's pets. Analysis of the pet visitation program's effect on human behavior and emotional state [38] enables evaluating non-human talking aspect. Although no significant differences came from pet visitation vs. no pet visitation, the feature *talking with animals* was evaluated. For 2/3 of the people, it was one of the reasons they enjoyed the pet visitation program. In a similar program, pet-assisted living for people with mild cognitive impairment was analyzed. Six behavior interactions were evaluated, with *talking to pets* being one of them [39]. *Looking at*, *talking to*, and *touching* the pet were the most frequently occurred activities, although only *looking at* was linked to mood change. Additionally, child and pet speech exhibit some resemblances [40], which suggests a behavior influence of talking with pets, similarly to what happens when people talk to children. The association of pets with humans is widespread, as many people have pets in their households [41], and such natural affinity for animals is seen in the early stages of life [42].

Inanimate. This trait of talking alone or to inanimate objects, more commonly linked to mentally unstable people, can also be of use to evaluate healthy people's behavior. A

closely related concept to inanimate object talking is anthropomorphism. It can be described as "imbuing the imagined or real behavior of nonhuman agents with human-like characteristics, motivations, intentions, and emotions" [43]. Interestingly, anthropomorphism is surprisingly common, and its presence might be associated with lacking a sense of social connection to other humans [43]. Another important consideration is that preoccupation with closeness and fear of abandonment are strong predictors of anthropomorphism, which further corroborates that anthropomorphism reflects an active search for human connection [44]. Another variation of "talking to inanimate objects' is talking alone. When performing a task, it is not uncommon to verbalize our intentions or the state of actions (what one is doing currently and what it is supposed to do next, respectively) in the form of self-direction [45]. This concept has been explored in working memory, where one of its components, sub-vocal rehearsal, assists in sustaining verbal information to be used in subsequent actions [46]. Verbalisation has shown to play a role in sustaining more than one active task in working memory [47], and the use of relevant verbal instructions promotes sustained concentration on task goals when performing multiple tasks [48]. As such, talking alone might not be only linked to mental issues and might even be seen in several humans, given its utility in task performing.

Passive Verbal Interaction. The concept of *passive* is to designate the listener of a conversation, highlighting that *listening*, by itself, is also a vital aspect of behavior analysis. Several studies have focused on the listening importance for social interaction and perception. Social perception, one of the four branches of emotional intelligence [49] and contributor to psychological wellbeing [50], is linked to listening comprehension [51]; the same authors also refer to listening comprehension as a *social skill and a form of achievement*. Regarding social interaction, a concept to look at is the use of active listening skills. Active listening can be defined as a performance of actions, in a conversation, for social engagement, such as: make empathetic comments, asking appropriate questions, paraphrasing, and summarizing for verification [52, 53]. Thus, effective communication makes use of active listening skills [54]. Listening has also been recognized as an essential communication component in healthcare [55], but the same idea can be translated into daily communications in other contexts. Furthermore, effective listening is a skill that requires both cognitive and behavioral processes to attend to verbal and non verbal cues of the speaker [56]. Another evidence of listening importance is that, in the context of human development, children are capable of listening to and talk about much more complex ideas that they can read about [57]. The assessment of complex ideas in the early stages of life requires oral input instead of a written one. Additionally, children's listening skills are a strong determinant of their scholastic success [58].

2.3.1.2 Deep Learning Methods

The main works in this topic relate to **ASD**, which aims to identify, from a set of potential candidates, active speakers on a given visual scene [59]. Currently, this assessment is done at the video frame level, based on facial cues and sound information, which is mainly due to the fact that state-of-the-art ASD datasets have good audio and face qual-

ity, yielding from controlled setups (movies, from AVA-ActiveSpeaker [59]) and cooperative settings (interviews, from AWS [60]). Based on the available data for ASD, current state-of-the-art heavily rely on face and audio data, with audiovisual combination using 3D architectures [61], hybrid 2D-3D models [62], large-scale pretraining [63, 64], feature embedding improvement [65], and attention mechanisms [66, 67, 68]. Various ASD works [69, 70, 71, 72] based their model on a two-step process, where the first focuses on short-term analysis (audio with face combination) and the second on multi-speaker analysis. ASC [69] focuses on multi speaker analysis via temporal refinement, ASDNet [70] uses a similar approach for inter-speaker relations with improved visual backbones, and UniCon [71] relies on audio-visual relational contexts with various backbones. The improvement of ASD performance by assessing contextual information via speaker relation using Graph Convolutional Networks (GCN) [73] has also been explored [72, 74, 75]. Diverging from two-step training, end-to-end models have also emerged for ASD [1, 75, 74, 76, 77]. TalkNet [1] focused on improving long-term temporal context with audio-visual synchronization, TS-Talknet [77] considered pre-enrolled speaker embedding to complement this synchronization, EASEE [75] included GCN to complement spatial and temporal speaker relations, and Light-ASD [76] proposed a lightweight model by splitting 2D and 3D convolutions for audio-visual feature extraction, and applied Bidirectional Gated Recurrent Units (BGRU) for cross-modal modeling. Recently, BIAS [4] has shown the importance of body inclusion for improved ASD in challenging conditions, while ASDnB [78] combined face with body cues for robust ASD, with state-of-the-art performance in cooperative and challenging data. An overview of the most commonly used approach for ASD is shown in Figure 2.2.

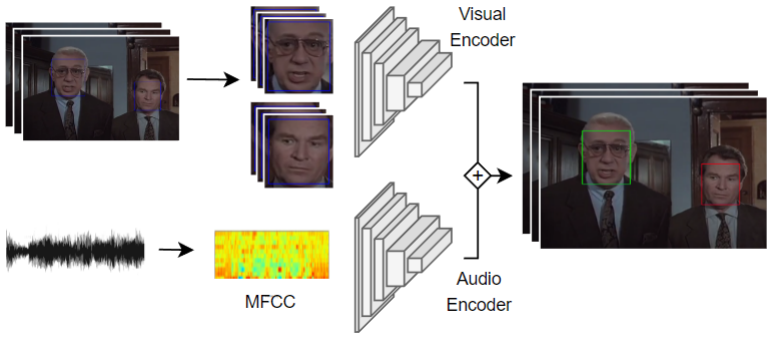


Figure 2.2: Overview of ASD models: combining face images with audio features to predict active speakers.

2.3.1.3 Datasets and Metrics

For specific verbal human interaction, we can use ASD datasets, which are mainly used in cooperative settings. The AVA-ActiveSpeaker dataset [59] is an audio-visual active speaker dataset from Hollywood movies, totalling almost 38 hours. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723 and 8,015 video utterances, respectively, ranging from 1 to 10 seconds. The main challenges of this dataset are related to language

diversity, FPS variation, the existence of faces with low pixel numbers, blurry images, noisy audio, and dubbed dialogues.

A more challenging ASD dataset is WASD [3] that compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*. With 30 hours of labelled data, WASD is divided into 5 categories with varying degrees of audio and face quality, grouped into categories: Optimal Conditions (OC), Speech Impairment (SI), Face Occlusion (FO), Human Voice Noise (HVN), and Surveillance Settings (SS). WASD contains 164 videos with varying FPS, averaging 28 second duration, with balanced demographics, and similar train/test division as AVA-ActiveSpeaker. Figure 2.3 contains representative frames from AVA-ActiveSpeaker and WASD.



Figure 2.3: Representative images from AVA ActiveSpeaker (top row) and WASD datasets (bottom row).

Metrics. To analyse and compare models performance, works use the official ActivityNet evaluation tool [59] that computes mAP. This metric relates Recall (R) and Precision (PR) of models predictions, expressed by:

$$mAP = \sum_{i=0}^{i=n-1} [R(i) - R(i+1)] \times PR(i), \quad (2.1)$$

where n is the number of thresholds, $R(n) = 0$, and $PR(n) = 1$. Recall and Precision are calculated by:

$$R = \frac{TP}{TP + FN}, \qquad PR = \frac{TP}{TP + FP}, \quad (2.2)$$

where TP referes to true positives, FN to false negatives and FP to false positives. As such, mAP summarizes the precision-recall curve into a single value representing the average of all precisions, expressed by the weighted sum of precisions at each threshold, where weight is the increase in recall.

State-of-the-Art Results. To provide an overview of the best performing models for ASD, we present the results on AVA-ActiveSpeaker (AVA) and WASD in Table 2.2. The best models in AVA are not able to maintain their performance on the harder settings of WASD, in particular with audio (HVN) and face access impairment (SS), highlighting the dependency of current approaches for cooperative settings and reliable face quality to perform.

Table 2.2: State-of-the-art results in AVA-ActiveSpeaker (AVA) and categories of WASD. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. Adapted from [3] and [4]. Best performance in bold and second best underlined.

Model	Body Data	AVA	WASD					
			OC	SI	FO	HVN	SS	Avg
AVA-Base [59]	×	82.2	-	-	-	-	-	-
ASC [69]	×	87.1	91.2	92.3	87.1	66.8	72.2	85.7
MAAS [72]	×	88.8	90.7	92.6	87.0	67.0	76.5	86.4
UniCon [71]	×	92.0	-	-	-	-	-	-
TalkNet [1]	×	92.3	95.8	97.5	93.1	81.4	77.5	92.3
BIAS [4]	✓	92.4	<u>97.8</u>	<u>98.4</u>	<u>95.9</u>	<u>85.6</u>	<u>82.5</u>	<u>94.5</u>
ASDNet [70]	×	93.5	96.5	97.4	92.1	77.4	77.8	92.0
TS-TalkNet [77]	×	93.5	96.8	97.9	94.4	84.0	79.3	93.1
EASEE-50 [79]	×	94.1	-	-	-	-	-	-
Light-ASD [76]	×	<u>94.1</u>	97.8	98.3	95.4	84.7	77.9	93.7
ASDnB [78]	✓	**94.6**	**98.7**	**98.9**	**97.2**	**89.5**	**82.7**	**95.6**

2.3.2 Physical Social Interaction

2.3.2.1 Psychology Background

The evaluation of physical interaction can be divided into human physical interaction, which means physical contact between two or more people, and non-human physical interaction, *i.e.*, a person interacting with a non-human entity (animal or object). Thus, the following subsections tackle the main points showing how human physical and non-human physical interaction analysis are relevant for human behavior analysis.

Human. Perhaps the most iconic example of human physical interaction is touch. The terminology *touch* is any contact between two people. Touch, as other non-verbal communicators, is thought to have a phylogenetic and ontogenetic primacy [80]. Phylogenetic means that, evolutionary wise, touch preceded language [81] and ontogenetic means that, in the early stages of life, the importance of touch supersedes that of verbal communication. Times of stress demonstrate the importance of touch and other types of non-verbal information. In these cases, humans tend to rely more on it than on verbal information, and our evolutionary heritage can explain such [80]. Regarding human development, touch plays a vital role in an infant's life. It is one of the first functional modalities to mature in the infant (and plays an essential role throughout his life) [82]. It regulates physiological states while also playing a decisive role in social development [83]. Touch has been shown to influence positive emotional communication and add positive reinforcement to other stimuli [84], generate positive emotions, and modulate negative ones [85]. Additionally, touch is responsible for generating negative emotions [86] and even contribute to behavior problems (later in life), in the context of harsh touch from mothers to their children [87]. In adults, the role of touch in communication is still a major one [80]. Touch is essential to communicate, namely, compliance, power, and emotional expression. The role of touch in compliance is one of the most studied phenomenons. It has been shown that customers who were touched are more inclined to taste new food samples [88], give more tipping to waitresses if touched while receiving change [89], and influence

the signing of petitions [90]. Touch is also a potential asset to communicate power and status [91]. In the context of emotional expression, quantity and quality of touch between couples are related to intimacy and happiness in the relationship [92]. Finally, people have more significant positive affective states and like a person more when touched [93], and men tend to respond less positively to touch than women [94] are aspects worth mentioning.

Non-Human. Non-human physical interaction refers to the interaction between a human and another non-human entity (object, animal, or pet) that are not possessed by said person, *i.e.*, are part of the environment. Shop scenarios, where customers' contact with selling objects is almost always present, demonstrate the importance of non-human physical interaction. In an era where online shopping is a possibility, almost 85% of people still do their shopping offline [95], claiming that the main reason for this is because they want "to see, touch, feel, and try out items" [96]. In this context, it has been shown that touch increases the feeling of ownership towards the product [97], increases confidence in product judgment [98], gives more favorable product evaluations [99], and contributes to unplanned purchases [100]. Manipulation of product touch accessibility, where customers were deprived of touching certain products, contributed to increased touching of other products (as a compensatory behavior) and increased purchasing during the shopping experience [101]. Another scenario where non-human interaction might occur is when entering in contact with pets. The literature has shown the influence of physical interaction with pets, particularly regarding health-promoting effects on oxytocin (social bonding related hormone) levels, blood pressure, and heart rate during soothing contact, petting, and holding [102]. In stressful events, physical interaction with pets reduces depression and anxiety levels [103]. In more dramatic scenarios, such as cancer, the support provided by non-human animals has shown to be as potent as that received from humans [104]. Human development also demonstrates the importance of non-human physical interaction. When presented with a novel object, children typically engage in *specific exploration* [105], where they intend to determine what the object does. Once they are familiar with the object, they engage in *diversive exploration* [105] to incorporate it into their play. Interestingly, interaction with object and play with it have a different demeanor. Object exploration has a flat or negative effect, increased heart rate, and the child is not easily distracted [106]. In contrast, object play has a positive effect, relaxed behavior, lower heart rate, with the child more easily distracted. Additionally, depriving children of the opportunity to explore objects may hinder their development. Prohibiting children from interacting with musical instruments may result in children *crippled musically* [107].

2.3.2.2 Deep Learning Methods

Although in psychology, humans interacting with objects from the environment (non-human physical interaction) or humans interacting with owned objects (handheld object use) have different meanings, this distinction is not easily retrieved from current computer vision approaches. As such, here we focus on human-human interaction while non-human physical interaction are included in Section 2.3.4. For human interaction, there

is no specific computer vision area, with **Action Recognition** being the most closely related topic. However, when referring to *action recognition*, works do not always have the same goal and can be typically divided into four major variations: 1) action recognition; 2) action segmentation; 3) temporal action localization; and 4) spatio-temporal action localization. The key differences between them are presented in Figure 2.4.

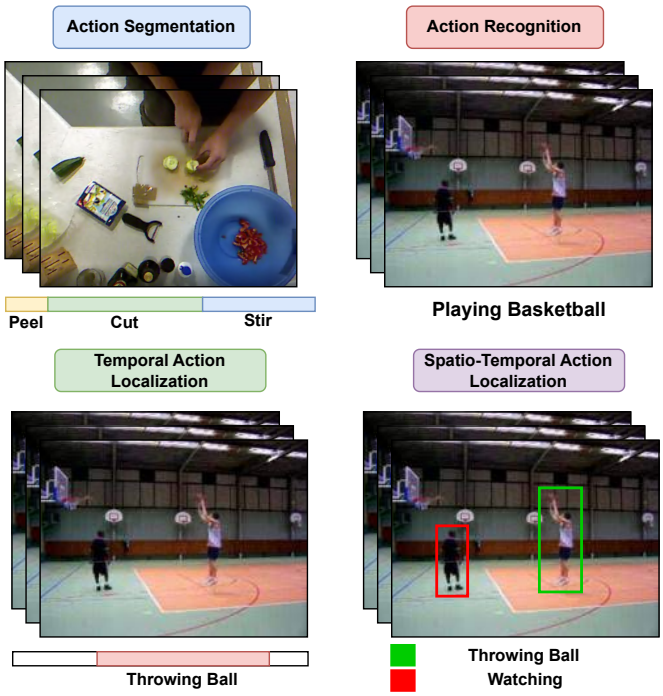


Figure 2.4: Variations of action recognition tasks.

For **action recognition/classification**, the goal of methods is to classify the action of a given video, with the datasets used for this task being **typically trimmed to the action and containing only one per video**. Various approaches complement additional data with RGB for action recognition, namely skeleton data, depth information, infrared, point cloud data, or complementing with movement acceleration analysis. Furthermore, aside architectural variations of CNN, RNN, or graph-based approaches, recently a greater focus on semi supervision [108, 109] (use of few labelled data) and few-shot [110, 111] (few data to generalize prediction for unseen categories) approaches has risen in attention.

Action segmentation methods aim to predict **frame by frame the division of different actions** of a given video. In this case, datasets contain videos of various **close-up actions**, related to simple tasks such as cutting items or moving objects. Contrasting with the previous variation, here, methods require a clear action division per frame, based on temporal convolutions by encoding-decoding [112], multi-stage architectures [113], graph-based convolutional blocks [114], and through global to local search schemes [115], improving over standard temporal convolutions. It can also be complemented via self or weakly supervised approaches [116, 117, 118], with context understanding [119], or even using a transformer-based architecture [120] to improve action segmentation.

Regarding **temporal action localization**, the goal is to identify the action of a video, and **predict when it starts and ends**. Dataset videos are not trimmed to the action

but typically only contain one. Relative to action segmentation, datasets for temporal localization require methods to analyze action context to accurately classify it and define its start and end (*e.g.* first frames of person running does not mean that the action is *running*, since the person can be running towards a football which means that the action to predict was *kicking a football*). In this area, the refinement of temporal action localization has been recently achieved through graph-based approaches [121, 122, 123], with weakly-supervised techniques [124, 125], considering temporal equivariance [126] and co-occorrence features [127].

For **spatio-temporal action localization**, methods have to **identify various actions** of a video, and **localize them via bounding box throughout the video duration**, thus being the most challenging task in action recognition. Instead of simply considering the context of the video to predict action, it is also necessary to consider how subjects interact since there are actions that depend on subject interaction (*e.g.*, *listening/talking to someone*). Initial approaches ranged from variations of action tubelets [128] to combining high with low frame temporal information [129], and most recently masked autoencoders [130, 131, 132, 133] for better model generalization, person tracker inclusion [134], and transformer-based architectures [135, 136].

For additional information on data modalities for action recognition and the most recent key works in this area, we refer to the works of Sun *et al.* [137] and Gui *et al.* [138].

2.3.2.3 Datasets and Metrics

For action recognition datasets with physical human interactions in less cooperative settings, we highlight the UT-Interaction dataset [139], which contains videos from an open space. It has 20 video sequences of a parking lot, whose length is around one minute, and each video contains at least eight executions of a specific interaction. Several participants were used with different clothing combinations, and videos were recorded on variable weather and luminosity conditions. Each video contains one of the six evaluated actions: shake-hands, point, hug, push, kick, and punch.



Figure 2.5: Representative images from UT-Interaction (top row) and AVA datasets (bottom row).

Regarding spatio-temporal action localization, AVA [140] is one of the more challenging datasets, with 15-minute segments from 432 Hollywood movies, where actions are localized in space and time, resulting in 1.62M action labels with multiple labels per human oc-

Table 2.3: State-of-the-art results in Kinetics 400 (K400) and AVA Actions (AVA) using Top-1 accuracy and the mAP metric, respectively. SSv2 refers to Something Something-v2 [5] dataset.

Method	Pre-training	Backbone	K400	AVA
SlowFast [129]	K600	ResNet	79.8	30.7
TubeR [136]	K400	CSN-152	-	33.6
MaskFeat [133]	K400	MViTv2	86.4	38.8
MAE-ST [132]	K400	ViT	84.8	39.0
VideoMAE [131]	SSv2	ViT	69.6	39.5
LART [143]	K400	ViT	-	42.3
VideoMAE V2 [144]	Various	ViT	88.6	42.6

curring frequently. It contains 80 classes, from which some are related to human-human interaction: watch a person, talk or listen to a person, fight, hit, hug, push, take an object from a person, among others. Kinetics 400 [141] has 400 human action classes, with each action having from 400–1150 clips, each lasting about 10 seconds from YouTube videos. The list of actions covers person actions (*e.g.*, drawing, laughing), person-on-person actions (*e.g.*, hugging, shaking hands) and person-object actions. Figure 2.5 contains examples of actions from AVA and UT-Interaction datasets.

Metrics. Given the relevance of spatio-temporal action localization for human-human interaction in surveillance conditions, we report the metrics used for AVA Actions Dataset [140] and Kinetics [141]. Intersection over Union (IoU) is done at frame level, using mAP on a IoU threshold of 0.5, following the protocol used by PASCAL VOC challenge [142], and at video-level, 3D IoU between ground truth tubes and prediction tubes is computed at threshold of 0.5. For action classification, mAP is computed by averaging over all classes for AVA Actions Dataset and fine-tuned Top-1 accuracy for Kinetics.

State-of-the-Art Results. We present the reported results on AVA Actions dataset and Kinetics 400 in Table 2.3 for spatio-temporal action localization, highlighting the importance of pre-training and Transformer-based architectures given its most recent success on this area.

2.3.3 Relative Position

2.3.3.1 Psychology Background

When considering a context, an important behavioral aspect is to evaluate where people go and define one's location in a given scene, expressed by the therm Relative Position (to a given camera). The interest in analyzing a person's position in a scene has its roots in what motivates a person to move to a specific location. In human development, there are two main theories regarding this topic: destination-directed locomotion [145] and peragration-driven locomotion [146]. In the first theory, locomotion is motivated by exciting destinations, engaging in social interactions, or interacting with objects [145]. Alternatively, the second theory proposes that an infant's locomotor exploration is motivated with no proximal goal other than the body's motion through space [146], not motivated by the allure of an immediate destination [147]. Furthermore, in children, even though the

presence of toys and caregivers may motivate locomotion, the simple act of locomotion may be self-motivating enough to explore a new room [148]. According to an embodied view of development, movement and motor control are essential features for cognitive development. In this sense, cognition comes from the interaction between one's physical body and the physical world [149]. Children do not passively perceive information but, rather, actively interact with the world, thus learning increasingly more about it [150]. A way to explore and interact with the world, both in infants and adults, is space exploration via locomotion. Another concept associated with the human position in a scene is exploration. In line with a previously mentioned locomotion theory, it is feasible to think that exploring regions of interest motivates humans to move. In fact, in human development, exploration is linked to advances in spatial cognition [145]. Evaluation of exploratory behavior, assessing gender differences, has been done [151], showing that women present a more cautioned approach to exploring, reflected in increased amounts of pausing and revisiting of previously visited locations.

2.3.3.2 Deep Learning Methods

The area more closely related to assessing the relative position of a person in a scene, over a period of time, is **MOT**. The goal of methods in this area is to relate all detected objects (people) bounding boxes of a video, such that all boxes belonging to one person contain the same identifier (ID), even in the presence of occlusions/overlap. As such, there are two key tasks in this area: 1) Detecting objects; and 2) Re-ID of detected objects. An overview of the key steps for MOT is shown in Figure 2.6.

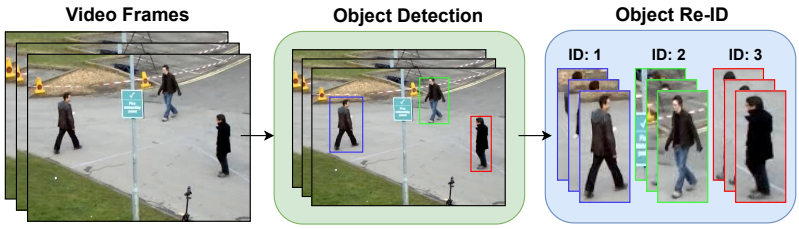


Figure 2.6: Overview of the two key steps for MOT.

Given the objective of MOT, some approaches obtain the object bounding boxes from off-the-shelf object detectors, focusing the novelty on the (Re-Id) part [152, 153, 154] (**Patch-based** methods). Although this has various advantages such as ease of combination with other methods/detectors (*i.e.*, separate training) and easier training (since its inputs are cropped data, and not the entire video frame), the main drawback is the lack of temporal context for Re-ID. By the same token, **Single-Frame** methods [155, 156, 157] focus on better feature extraction for the detection portion, without requiring extra computation and storage for detection boxes (from external object detectors), while having similar limitations regarding Re-ID (temporal information).
To convey a better idea of video information, emphasis on temporal notion to aid in detection and, particularly, Re-ID is the focus on several MOT works. They jointly learn the appearance and temporal features across multiple frames (**Cross-Frame**), with variations

on feature extraction being on multi-frame level [158, 159, 160] or frame-independent feature extraction, with shared weight extractor [161, 162, 163]. There are also works that focus on improving target location by combining feature maps of frame and object bounding box [164, 165, 166] (**Correlation-Features**), while others incorporate recurrent models for sequential modeling, relating motion and appearance variation for better target localization [167, 168, 169] (**Sequential**). To combine local (object) and temporal information, several methods take a sequence of detections as input (instead of individual object bounding boxes), typically denominated as **Tracklets**, with variations on considering tracklets from one person [170, 171, 172] or assessing pair-wise tracklets [173, 174, 175]. Similarly, evaluating other tracklets related to a subject (**Cross-Tracklet**) can also be useful for MOT tasks, either via attention-based methods [168, 176] or graph representation [177, 178].

For a better overview of MOT methods, works of Wang *et al.* [179] and Ciaparrone *et al.* [180] summarize the key concepts and works of this area.

2.3.3.3 Datasets and Metrics

For MOT evaluation, MOT Challenge is the most commonly used benchmark, with all variations of MOT dataset contain videos from different locations, crowded areas, variable light conditions, with videos duration of up to two minutes. In particular, **MOT15** [181] is composed of 22 videos (11/11 for training and testing), collected from older datasets, with 11 283 frames at various resolutions, 1 221 different identities and 101 345 boudning boxes. **MOT16** [182] has higher pedestrian density, with object detection from Deformable Part-based Model (DPM) [183], with 14 videos (7/7 train and test), 11 235 frames, 1 342 identities and 29 2733 boxes in total. **MOT17** has the same videos of MOT16, but each sequence is provided with 3 sets of detections: DPM, Faster-RCNN [184], and SDP [185]. **MOT20** [186] contains 8 videos (4/4 train and test), in unconstrained environments with extremely high pedestrian density (up to 245 per frame), composed of 13 410 frames, 6 869 tracks, and 2 259 143 bounding boxes. Figure 2.7 exhibits six frames from videos of MOT datasets.

Another dataset commonly used for MOT is **KITTI MOT** [187]. The main difference for MOT datasets is that KITTI can be used for tracking of people and vehicles. The data is collected by driving a car arround a city, consisting of 21 training videos, 29 test ones, totalling of about 19 000 frames (32 minutes), and object detection is obtained from DPM and RegionLets [188].

Metrics. Although there are various metrics for MOT evaluation, the most recent and commonly used are Multiple Object Tracking Accuracy (MOTA) [189], ID-based metrics [190], and (Higher Order Tracking Accuracy (HOTA) [191]. MOTA is based on the matched pairs, misses, false positives, and mismatches, expressed by:

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \quad (2.3)$$



Figure 2.7: Representative images from MOT20 (top row) and KITTI MOT datasets (bottom row).

where t is the frame index, GT refers to ground-truth, FP to false positives in the whole video, FN to false negatives, and IDSW to the total number of ID switches. For ID-based metrics, the focus is to assess for how long the tracker correctly identifies targets, by association of ground-truth trajectory with predicted one. Typically, IDF_1 score is the most commonly used metrics given its rank of all trackers on a single scale that balances identification precision and recall through their harmonic mean:

$$\text{IDF}_1 = \frac{2\text{IDTP}}{2\text{IDTP} + \text{IDFP} + \text{IDFN}}, \quad (2.4)$$

where IDTP refers to true positive identification, IDFP to false positive identification and IDFN to false negative identification. Given that **MOTA overemphasizes detection performance** and **IDF_1 overemphasizes association rather than detection**, HOTA has been recently proposed to balance accurate detection, association, and localization. It is composed of sub-metrics that are able to evaluate each of five basic error types (detection recall, detection precision, association recall, association precision, and localization accuracy), expressed by:

$$\text{HOTA}_\alpha = \sqrt{\frac{\sum_{c \in \{TP\}} \mathcal{A}(c)}{|TP| + |FN| + |FP|}} \quad (2.5)$$

$$\mathcal{A}(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \quad (2.6)$$

where TPA, FNA, and FPA refer to associations of true positives, false negatives, and false positives, respectively. Details regarding the association function can be seen in [191].

State-of-the-Art Results. To provide an overview of the best performing models for MOT, we present the reported results on MOT17 in Table 2.4, based on method grouping. The overall best performing methods regarding tracking and detection balance (HOTA) are linked to single-frame evaluation, while the worse performing ones relate to tracklet assessment.

Table 2.4: State-of-the-art results in MOT17 dataset.

Mode	Method	MOTA	IDF1	HOTA
Patch	SimpleReID [192]	61.7	58.1	46.9
	Semi-TCL [193]	73.3	73.2	59.8
	ReMOT [152]	77.0	72.0	59.7
Single-Frame	OUTrack [155]	73.5	70.2	58.7
	FairMOT [194]	73.7	72.3	59.3
	RelationTrack [195]	73.8	74.7	61.0
Cross-Frame	TransCenter [196]	73.2	62.2	54.5
	GSDT [197]	73.2	66.5	55.2
	PatchTrack [163]	73.6	65.2	53.9
Correlation	QDTrack [166]	68.7	66.3	53.9
	TransTrack [198]	75.2	63.5	54.1
	CorrTracker [164]	76.5	73.6	60.7
Sequential	TrajE [167]	67.8	61.4	50.0
	PermaTrack [199]	73.8	68.9	55.5
	MNT [169]	74.9	75.0	62.0
Tracklet	Visual-Spatial [175]	56.8	58.3	46.4
	Dai *et al.* [171]	59.0	66.8	51.5
	Split-Connect [170]	61.5	63.3	50.5
Cross-Tracklet	Brasó *et al.* [200]	58.8	61.7	49.0
	GSDT [197]	73.2	66.5	55.2
	TPAGT [201]	76.2	68.0	57.9

2.3.4 Handheld Devices

2.3.4.1 Psychology Background

This section regards humans' interaction with handheld devices that belong to said person, such as phones, personal computers, books, among others. Since not all handheld objects have the same meaning regarding human behavior, here we divided them into: *Phone*, *Leisure Items*, *Aggressive Items*, and *Carrying Items*.

Phone. Pedestrian crossing behavior analysis report evaluates phone usage. Phone usage contributes to many unsafe crossing behaviors such as sudden stops, disobeying traffic laws, not looking both ways before crossing [202]. In this context, pedestrians who use a phone for texting are less able to maintain walking speed and pose a higher threat to public safety by ignoring their surroundings than pedestrians who used phones for other purposes other than texting (*e.g.*, listening to music) [203]. In simulation scenarios, drivers, while texting and driving, might compensate for their reduced context-awareness by driving more slowly and with a more significant distance from other cars [204].

Leisure Items. A context that one might consider is the fact that usage of a handheld device is related to leisure activities. Leisure behavior is self-determined activities outside of obligatory time, where one experiences freedom, positive emotions, a sense of control, and intrinsic motivation [205]. Thus, handheld devices such as books or personal computers can be linked to leisure activities. Some perceived experiences from leisure are known as resources to resist stress [206]. Similarly, leisure could be applied to generate engagement and enhance one's mood [207], and leisure generated positive emotions promote subjective well-being and happiness [207]. In adverse life events, turning to leisure appears to be protective and restorative [208]. In cases of depression and other chronic

conditions, it seems to be amenable to the positive influence of leisure [209]. Additionally, people in stressful conditions that are actively involved in leisure have stress' negative impact on health and well-being reduced [210].

Aggressive Items. The presence of aggressive items, such as weapons, guns, knives, or clubs, may give some information regarding one's intentions and, potentially, predict a behavior. It is a significant warning to see a person carrying a weapon, especially if the country's legislation does not allow one to do so. An aggression scale associated with fighting and weapon carrying exists in the literature [211], which indicates the importance of such aspect in human behavior categorization. Many weapon carriers face repercussions in schools, such as suspension or expulsion, and seriously sanctioned weapon carriers are more likely to be delinquent [212]. Weapon carrying is associated with substance trade or abuse [213, 214] and is linked to deviant peer groups (*i.e.*, gangs) [215]. Additionally, a history of delinquency is positively related to future intentions of knife carrying [216].

Carrying Items. The terminology "Carrying Items" refers to items that can carry other items in it, such as baggage, backpacks, and bags. The importance of these types of items may vary, depending on the environment. A backpack or bag might be ordinary in a classroom but, for airport security screeners who, within a few seconds, have to judge whether a bag should enter in an airplane or not [217], it might represent security issues. In x-ray screening, there is a focus on visual skills in airports and prohibited items performance evaluation [218, 219], which exhibits the importance of bag screening and bags' association to security issues. In 2010, about 9% of high schools banned book bags or required bag clearance for weapon carrying detection in America [220]. When inquiring legal gun holders if, in the past year, they carried a weapon on them (in a bag, pocket, or holster) or in their vehicles, 12.5% said they carried on their person [221]. Regarding drug trafficking, it has been shown that drug luggage carriage is the predominant way (45%) when compared to other possible approaches such as internal concealment (18%) or in clothing (14%) [222].

2.3.4.2 Deep Learning Methods

The deep learning area focused on assessing the interaction between humans and objects is **HOI**. The objective of methods in this area is to find, given an image, what kind of interaction exists, outputting a prediction composed of a triplet of human, verb, and object, with verb expressing the relation between human(s) and object(s) of the given image. An overview of the key steps for HOI is shown in Figure 2.8.

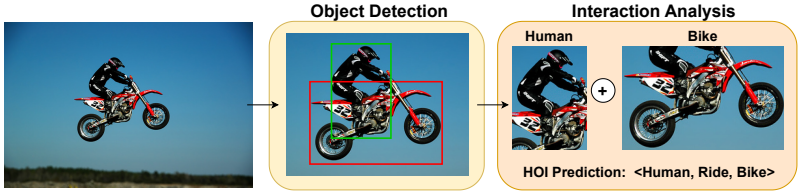


Figure 2.8: Overview of a two-stage approach for HOI.

The most commonly used approach by HOI methods is based on **two-stage**: detection,

followed by interaction. For detection, most work rely on off-the-shelf object detectors, such as Faster-RCNN, with the novelty being centered in the interaction prediction stage. Typically, following the object detection stage (which also include human detection), object bounding boxes are inputted to a Convolutional Neural Network (CNN), for feature processing and HOI prediction. The approaches for better feature processing, methods may opt to use various/custom architectures [223, 224], focus on spatial relation between object bounding boxes [225, 226], complement image features with pose information [227, 228], translate object detection to semantic word representation [229, 230], or use Graph Neural Networks to map object relation in graphs for more accurate HOI predictions [231, 225].

Other approaches rely on **one-staging**, where detection and interaction are estimated with two parallel branches. These methods tend to be more efficient than two-stage ones by extracting integrated features of object pairing (*vs.* individually processing of objects in the two-stage approach). The two key variations for one-stage approach are point-based [232, 233], which focus on the area between the object pair center points, and anchor-based [234, 235], which rely on a set of predefined bounding boxes of a certain height and width.

With the increase of popularity for other visual tasks, the use of **Transformers** have also been reported for HOI [236, 237]. The self-attention mechanism of transformers allow for better contextual relationship analysis between humans and objects, promoting better HOI predictions. These methods start with feature extraction to input on a Transformer Encoder-Decoder, and output a prediction for human an object bounding boxes (and class labels) to use on HOI triplet prediction. Other approaches for HOI consist of zero-shot learning [238], *holistic* and *zoom in* combination [239], and contextual object re scoring [240].

Given the psychology background, it may be relevant to distinguish what type of object a person interacts with. Although HOI focus on interaction for object in general (and can be used for specific uses), there are other works worth mentioning that focus on particular sets of objects such as abandoned items [241], aggressive item carrying and gun identification [242, 243]. For a better overview of HOI methods, Antoun and Asmar reviewed the key concepts and works of this area in a recent survey [244].

2.3.4.3 Datasets and Metrics

The two most commonly used datasets for HOI are VCOCO [245], a subset of the COCO dataset, and HICO-Det [246].

The HICO-Det is a large-scale and generic HOI detection dataset, containing 47 776 images, covering 600 categories of HOI, containing 117 verbs (*e.g.*, carry, hold, inspect, read, talk on, text on, wield), and 80 object categories (*e.g.*, cell phone, book, knife, laptop, baseball bat). HICO-Det is divided into 38 118 train/val and 9 658 test set, and each person in image can have multiple interaction with different objects. Typically, HICO-Det is evaluated on two different modes (Default and Known Object) and different categority sets (based on the presence of interaction categories), which we discuss in the metrics section.

Although this dataset is extensively used for HOI, it has some category balance and interaction label issues [244]. Examples of some interactions seen in this dataset are presented in Figure 2.9.
The VCOCO is build off the COCO dataset [247], and contains 10 346 images, 69 object categories, and 29 actions, with 25 having interactions with objects (*e.g.*, cut, hit, eat) and the remaining related to body motion (*e.g.*, walk). VCOCO has on average 1.57 people annotated per image, which corresponds to 16 199 human instances, having 2 000 and 800 images with two and three people annotated, respectively.



Figure 2.9: Representative images from HICO-Det (top row) and VCOCO datasets (bottom row).

Metrics. The performance of both datasets is assessed via mAP. In HOI, a detection is considered as true positive (TP) if both the location of human/objects and the interaction label are accurately predicted, and considered as false positive otherwise. For accurate human and object location, the IoU ratio between the predicted bounding box and ground truth needs to be greater than 0.5.
Following the Humans Interacting with Common Objects (HICO) classification benchmark [248], there are two modes typically used for HOI evaluation: Default (DT) and Known-Object (KO) mode. In KO settings, for each HOI category (*e.g.*, "riding a bike"), detection is only done on images containing the target object category (in this case, "bike"), while in DT settings HOI prediction is done on the full test set. Furthermore, mAP is also reported over three different HOI category sets: 1) all 600 HOI categories in HICO (Full); 2) 138 HOI categories with less than 10 training instances (Rare); and 3) 462 HOI categories with 10 or more training instances (Non-Rare).
State-of-the-Art Results. To provide an overview of the best performing models for HOI, we present reported results on HICO-Det, using the Default and Known-Object modes and its groupings, in Table 2.5. Overall, transformer-based methods are superior to two-stage, with one-stage being the approach with worse performance for all variations of HICO-Det.

2.3.5 Proximity

2.3.5.1 Psychology Background

The terminology *Proximity* expresses the idea of how many people are close to a given person and analyze if said person is in a group or not. Regarding pedestrian behavior, people have stronger intentions to cross a road against a traffic signal when surrounded by

Table 2.5: State-of-the-art results in HICO-Det. TS, OS, and T indicate Two-Stage, One-Stage, and Transformer, respectively. F, R, and NR refer to Full, Rare, and NonRare, respectively.

	Method	HICO-Det					
		Default			Known Object		
		F	R	NR	F	R	NR
TS	VCL [249]	23.63	17.21	25.55	25.98	19.12	28.03
	DRG [250]	24.53	19.47	26.04	27.98	23.11	29.43
	IDN [251]	26.29	22.61	27.39	28.24	24.47	29.37
	SCG [226]	**31.33**	**24.72**	**33.31**	**34.37**	**27.18**	**36.52**
OS	PPDM [232]	21.94	13.97	24.32	24.81	17.09	27.12
	DIRV [235]	21.78	16.38	23.39	25.52	**20.84**	26.92
	GG-Net [252]	**23.47**	**16.48**	**25.60**	**27.36**	20.23	**29.48**
T	CDN [253]	32.07	27.19	33.53	34.79	29.48	36.38
	STIP [254]	32.22	28.15	33.43	35.29	31.43	36.45
	UPT [255]	32.62	28.62	33.81	36.08	31.41	37.47
	QAHOI [256]	**35.78**	**29.80**	**37.56**	**37.59**	**31.66**	**39.36**

pedestrians doing so [257]. Human's unmatched ability to imitate [258], a trait frequently cited as vital for our success [259], explains the pedestrian crossing scenario. Grouping influence is seen in human development where children tend to imitate in-groups more than out-groups regardless of whether the group engages in pro-social or antisocial behaviors [260]. Additionally, group behavior influences self-control in children. Using the marshmallow task [261], children tend to wait longer when their in-group chose delayed gratification (and out-group chose immediate gratification) [262]. When in-group chose immediate gratification, children also opted to decide accordingly. In the context of gratification, leisure activities, where positive interactions with other people occur, contribute to stress coping [263]. Grouping's influence on human behavior has been evaluated in several famous experiments. One of the most iconic ones is Sherif's Autokinetic Effect Experiment [264], where it was shown that people tend to conform to group norms when put in ambiguous situations. Furthermore, the influence of social pressure from a majority group was also explored in a given person. In Ash's Conformity Experiment [265], one person was grouped with people who occasionally gave an obviously wrong answer regarding a stick size. The critical point here was that the naive person believed that no group existed, and all people gave their honest opinion regarding a stick size. The results show that, throughout the experiment, about 75% of participants conformed at least once with a clearly wrong answer.

Evaluating if one is alone is another aspect of human behavior analysis. The isolation of a person, in a given scenario, from other people does not necessarily mean loneliness. However, the literature mainly focuses on the behavior of single isolated individuals in the context of loneliness. Loneliness, or perceived social isolation, has been reported to have negative consequences for mental and physical health while also reducing motivation for engaging in healthy behaviors [266]. Additionally, loneliness is linked to more sedentariness, less engagement in physical activities, and more emotional eating [267, 268]. Targeting loneliness for psychosis treatment has also been proposed with positive results [269], indicating that interventions to decrease loneliness motivate the lonely person to participate in healthy eating and activity. These two components are negatively affected

by loneliness, suggesting that loneliness and social isolation are features worth analyzing since they can be used, by targeting it, to improve one's health.

2.3.5.2 Deep Learning Methods

Proximity conveys the idea of how many people are close to a given person and analyze if said person is in a group or not. Although there is not a major area in deep learning related to this, we highlight **Depth Estimation** as a topic of interest regarding proximity analysis. The goal of methods of this are is to estimate distance of each image pixel relative to the camera, based on a single view (monocular) or various (stereo) of the scene. As such, even if it can not be used directly on proximity-related tasks, methods estimating distance from camera can be useful to infer relative object proximity. Examples of depth estimation are shown in Figure 2.10.



Figure 2.10: Examples of depth estimation in indoor and outdoor scenarios.

For standard depth estimation methods, works are mainly divided into monocular and stereo depth estimation. For **stereo-based** methods, which consider different views of a scene for depth estimation, recent approaches include change of problem solving strategy [270, 271], improvements based on geometric considerations [272, 273], dynamic temporal representation [274], focus on events for depth detail [275], and unsupervised learning [276]. For **monocular**, where a single view is used for depth estimation, there has been a greater emphasis on self-learning based methods, exploring feature variation [277, 6], multi-scale approaches [278, 279], encoder-decoder [280, 281, 282], geometric constraints [283], and resolution reduction [284]. For additional information of depth estimation topics, we refer the works of Ming *et al.* [285] and Laga *et al.* [286] for stereo and monocular depth estimation, respectively.
Regarding other methods, not depth-related and more targeted toward people proximity, there are some occasional works reported in the literature. These are mainly based on object detection approaches, containing different strategies for proximity analysis such as centroid comparison or pose estimation complement, with some GitHubs emerging for this task given its importance in COVID era. Other important concepts towards proximity are *group formation* and *facing formation* [287] (*i.e.*, socio-spatial formation where people maintain a convex space where everybody in the gathering has direct, easy, and equal access), which were the basis of some works [288, 289, 290]. These works are highlighted here for its topic relevance but their technical content is not highly innovative nor its comparison with other methods is easy since they were proposed for particular uses. As such, given the depth estimation topic, we highlight the monocular-based methods as a reference, given their importance and prevalence on the current literature, with the works of Ming *et al.* [285] and Laga *et al.* [286] summarizing the key concepts of the area.

2.3.5.3 Datasets and Metrics

The NYU-Depth V2 dataset [291] is comprised of video sequences from a variety of indoor scenes recorded by both the RGB and Depth cameras from the Microsoft Kinect, containing 1 449 densely labeled pairs of aligned RGB and depth images, 464 scenes from 3 cities, image size of 640×480, 407 024 unlabeled frames, with each object being labeled with a class and an instance number. Train-test division may differ from work to work, and NYU-Depth V2 contains a labelled and raw (unlabelled) version.
KITTI [292] is an autonomous driving dataset consisting of 61 outdoor scenes of different types (*e.g.*, city, residential, road), with RGB images of 1224x368, with up to 15 cars and 30 pedestrians visible per image, totalling 6 hours of traffing scenarios.
For person proximity-tasks, Oxford Town Center [293] is a commonly used dataset. It has around 5 minutes of footage of people walking in a busy street where partial occlusion occurs, with several different people walking by at the same time. It contains information regarding bounding boxes for people and their heads. Each person is associated with a number for tracking purposes. Frame examples of this dataset are seen in Figure 2.11.



Figure 2.11: Representative images from KITTI (top row) and NYU Depth (bottom row) dataset.

Metrics. The standard depth estimation metrics are Root Mean Square Error (RMSE), its log variant ($\text{RMSE}_{\log}$), Absolute Relative Error (A.Rel), Squared Relative Error (S.Rel), and percentage of inlier pixels within 1.25 (δ_1), as defined below:

$$
\begin{aligned}
\text{A.Rel} &= \frac{1}{n}\sum_{p}^{n}\frac{|y_p - \hat{y}_p|}{y}, \\
\text{S.Rel} &= \frac{1}{n}\sum_{p}^{n}\frac{\|y_p - \hat{y}_p\|^2}{y}, \\
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{p}^{n}(y_p - \hat{y}_p)^2}, \\
\text{RMSE}_{\log} &= \sqrt{\tfrac{1}{n}\textstyle\sum_{p}^{n}\|\log y_p - \log \hat{y}_p\|^2}, \\
\delta_1 &= \%\text{ of } y_p \text{ s.t. } \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta < 1.25,
\end{aligned}
\tag{2.7}
$$

where y_p is a pixel in depth image y, $\hat{y}_p$ is a pixel in the predicted depth image $\hat{y}$, and n is the total number of pixels for each depth image.

Table 2.6: State-of-the-art results in KITTI at a depth of 80m. Results extracted from [6].

Method	A.Rel	S.Rel	RMSE	$RMSE_{log}$	δ_1
DORN [294]	0.072	0.307	2.727	0.120	0.932
VNL [283]	0.072	-	3.258	0.117	0.938
AdaBins [280]	0.067	0.278	2.960	0.103	0.949
P3Depth [6]	0.071	0.270	2.842	0.103	0.953
BTS [279]	0.059	0.245	2.756	0.096	0.956
DPT-Hybrid [295]	0.062	-	2.573	0.092	0.959

State-of-the-Art Results. To provide an overview of the best performing models for depth estimation, we present the reported results on KITTI in Table 2.6.

2.3.6 Clothing

2.3.6.1 Psychology Background

Clothing is a vital behavior aspect for human understanding since people use clothing to reinforce their identity and portray social status [296]. Giving appearance a high degree of importance is linked to higher appearance investment, influencing clothing choice [297]. Furthermore, innovative and creative people are more likely to play with aesthetics to differentiate themselves from others [298]. The importance of clothing for one's self-esteem is also worth noting. Fashioning the body in clothing is a practice that helps the person feel better about themselves and allows this feeling's sharing with others [299]. Interestingly, clothing therapy has been reported to aid rehabilitation for cancer [300] and mental illness patients [301]. Another aspect is the effect that clothing can have on others. People tend to judge competence, confidence, and credibility in the first 12 seconds of an interaction, with clothing having an influential role [302]. Formally dressed students receive higher presentation ratings from fellow students [303], and women who appear "sexy" are labeled as less competent than those who dress appropriately are examples of such influence [304]. Another one is casually dressed teachers are seen as more approachable but less knowledgeable [305]. Interestingly, gender and social status might influence other's perceptions. Higher status women are judged more harshly when using provocative clothing [306], including rating them as less intelligent if they are in the manager's position [307]. For comparison, people do not generally view informally dressed professional men as something negative [308]. Adam and Galinsky proposed the theory of enclothed cognition [309], where it is mentioned that clothing has various implications and affects the wearer's social cognition. In the same study, participants who whore a medical coat performed visual search tasks better than those who did not. However, their performance deteriorated when they were told that it was a painter's coat. Wearing revealing clothing has been shown to impair women's performance in mathematical tasks [310]. Clothing attributes, namely color, can also have behavioral meaning. Individuals are more aggressive when wearing a black sports kit [311] or a hood and cap [312]. Women's usage of red or pink-colored clothing was linked to ovulation periods since they are motivated to enhance attractiveness during this period [313], and men tend to perceive red as sexually

attractive [314]. However, such a concept has been recently contested [315].

2.3.6.2 Deep Learning Methods

For clothing detection/recognition, the area of deep learning more closely related uses **Clothing** or **Fashion**-based methods. However, while the area of other features are targeted towards specific tasks, in clothing deep learning we consider four key groups related to clothing assessment: 1) landmark estimation; 2) clothing segmentation; 3) item retrieval; and 4) (surveillance) clothing detection. Figure 2.12 summarizes the key differences between the four groups.



Figure 2.12: Clothing related tasks.

For **landmark estimation**, the goal of methods is to predict clothing keypoints, similar to pose estimation approaches, landmarking the image regions linked to clothing, capturing the design, pattern, and category of clothes from a given image. From the most widely used approaches for this task, we highlight architecture customization [316, 317], multi-stage evaluation [318], pose estimation complement [319, 320], global-local embedding [321], graph representation [322], and attention-based approaches [323, 324].
Regarding clothing segmentation, the focus of methods is to segment parts of an image related to different classes of clothes. Relative to object segmentation, this task requires finer-grained image evaluation and higher-level judgment [325] given the clothing semantic and deforming of different pieces of clothes. Key approaches for this task relate to multi-branch architecture [326], adding semantic information [327, 328], coarse-to-fine granularity [329], pose estimation complement [330], and clothing-relation [331, 332].
When considering **item retrieval**, the goal is to propose the more closely related items (from a item gallery) based on the item(s) from a given image. The main real-world application of these methods is retrieving the most closely looking item(s) from an online shopping image given a real-world image. The key challenges is the misalignment of clothing from real-world to item gallery, leading methods to focus on (body) part-based alignment [333], pose estimation complement [334, 335], attribute-guided learning [336], siamese networks to relate item similarity [337], graph-based similarity [338], and using semantic information [336, 339].
For **clothing detection**, the goal is to draw bounding boxes for different cloth pieces of

an image, accompanied by its classification. Approaches based on body poses [340] or object keypoints [341, 342, 343] are among viable strategies for this task, as well as treating clothing detection as an object detection task (we refer to [344] for a survey on object detection). When considering the most challenging condition for clothing evaluation, we have surveillance settings, which is done in **PAR datasets**. The goal of PAR methods is to assess the presence/absence of attributes in a given image. These attributes range from soft biometrics, objects, to clothing. As such, there are two key different of these methods regarding the previously mentioned: 1) their predictions takes into consideration multiple attributes and not only clothing; 2) the presence of clothing is not accompanied by outline of its location in the image. As such, even if its application to fine-grained clothing localization is not entirely suited, these methods are trained in most challenging sets (surveillance), contrasting with the cooperative settings of the previously mentioned approaches. More details regarding these methods are discussed in Section 2.3.8.
For a better overview of clothing methods, works of Cheng *et al.* [325] and Gong *et al.* [345] summarize the key topics of this area. Another clothing-related topic is recommender systems, whose main concepts are discussed in the survey of Deldjoo *et al.* [346].

2.3.6.3 Datasets and Metrics

DeepFashion-C [347] is a universal dataset for clothing classification and attribute recognition, which divides the clothes into 50 fine-grained categories, with corresponding images. The attributes are split into five groups: *texture*, *fabric*, *shape*, *part*, and *style*. There are 289 222 images in this dataset, of which 209 222 are used for training, 40 000 for verification, and the remaining 40 000 are test samples.
For various clothing tasks, DeepFashion2 is the largest fashion dataset to date [348]. It contains 491 000 images, with 801 000 items in total, with 43 800 clothing identities of interest, each with 12.7 items with different styles, such as color and printing. This dataset provides annotations for cloth detection and classification, landmark and pose estimation, instance segmentation and clothes retrieval, with data collected according to four properties, scale, occlusion, zoom-in, and viewpoint. The clothing items are grouped into three sets, such as small, moderate, and large, according to their proportion relative to image size. Although this dataset contains various clothing-related tasks, it is still on cooperative settings, with subject cooperation and good image quality. Figure 2.13 contrasts the examples of this dataset, with images from PAR datasets. Further details regarding PAR datasets are discussed in Section 2.3.8.
Metrics. We gather the main metrics used for DeepFashion2 [348], given its application for three key groups of clothing methods: landmark detection, clothing detection, and segmentation. **Clothing detection** uses mAP to evaluate the prediction of labelled bounding boxes (for different IoU, Section 2.3.4) and **landmark detection** employs the evaluation metric for human pose estimation (Equation 2.8), both following COCO [247]. **Segmentation** is calculated by average precision over the masks of each item, similar to clothing detection.
State-of-the-Art Results. To provide an overview of the best performing models for

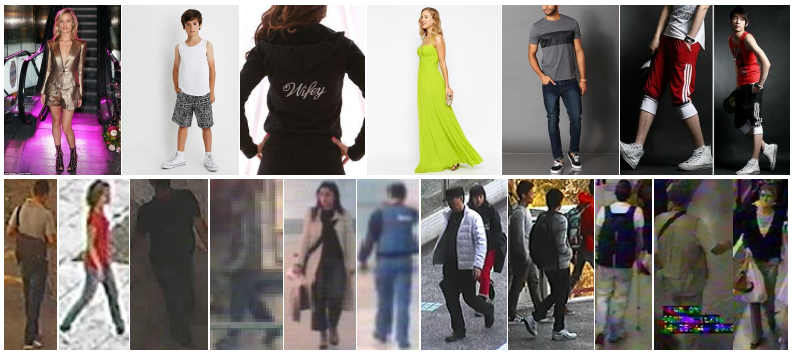


Figure 2.13: Representative images from DeepFashion (top row) and PETA (bottom row) datasets.

Table 2.7: Clothes Detection (mAP_{box}) and Landmark Detection (mAP_{pt}) in DeepFashion2.

Method	mAP_{box}	$mAP_{box}^{IoU=0.5}$	mAP_{pt}	$mAP_{pt}^{OKS=0.5}$
DeepMark++ [349]	65.2	-	50.7	-
Kim *et al.* [350]	68.6	79.7	48.4	74.1
Baseline [348]	66.7	81.4	64.1	82.0
DeepMark [351]	72.3	82.7	60.1	79.3

DeepFashion2, we present the reported results on Table 2.7 for landmark and clothing detection. DeepMark is the most resilient approach, relating the number of channel with clothing classes, and using features of object centers and keypoints offsets in a coarse-refined way.

2.3.7 Pose Estimation

2.3.7.1 Psychology Background

How a person behaves in a scene is also dictated by its pose or posture. We divide *pose* into three components for more nuanced pose evaluation: head, arms/hands, and legs/feet.
Head. The way people look, where they look, or some head gestures can help infer a person's feelings or intentions. Head nodding is a nearly universal sign of agreement, similar to bowing, and is a form of submission that affirms the speaker's status in a conversation [352]. In addition to nodding, eye gaze and smiling are other examples of non-verbal behaviors during social interaction [353]. Eye gaze and head movements are associated with directness, honesty, and friendliness [354, 355]. Another aspect of nodding in social interaction is related to social status as both male and female students tend to nod equally to professors speaking and less to their peers [356]. In social interactions, namely physician-patient dyads, patient satisfaction was linked to physician expressiveness, which, among other aspects, included more nodding and gazing [357]. Even more critical, distance behavior (such as looking way) was strongly related to short and long-term decreases in patients' physical and cognitive functioning in the physician-patient relationship. In contrast, smiling and nodding usage was associated with functioning improvement, both in the short and long-term [358]. The head orientation is also influential for social interactions. Tilting head downwards has been shown to increase dominance perception, even with a neutral face. Such an effect is due to lowered and V-shaped eyebrows' illusory ap-

pearance [359], which increases social rank perceptions [360]. Head posture and gaze interactions have been shown to modulate status-related traits such as dominance, strength, and anger [361]. Another important concept related to head movement is gaze bias. It is a phenomenon with a proven role in attention and decision-making where, when people face a decision, their gaze tends to be biased towards the option they will, ultimately, choose [362].

Arms and Hands. Hands or arms' movement is something inherently present in humans and is something associated with human interaction [363]. Most people move their hands spontaneously even when talking with someone on the phone or to a blind person [364], suggesting that hand movement serves a purpose for the speaker, independent of the listener. Blind children also exhibit hand gestures similar to sighted children [364], which corroborates hand movement's influence on the speaker communication. The purpose of hand movements focuses on the idea that, even though they aid in speaking, such as emphasizing particular words or highlighting key ideas, they might also convey information that is not explicitly encoded in speech. As such, they provide additional information about the speaker's intentions or thoughts [365]. Information conveyed in gesture is also incorporated, by the listener, into the speaker's message. When adults observed a series of children explaining their responses to a given talk, they incorporated information obtained from the child's gesture to justify the children's reasoning [366]. Similar to head movements, hand gestures, in a social context, have been associated with directness, consciousness, and friendliness [354, 355]. Also analogous to head movements, patient satisfaction correlates with physician expressiveness (*i.e.*, more gesture usage) in a physician-patient dyad [357]. A more explicit usage of arm movement is pointing. Usually, in adults, pointing is not considered a communicative act unless "the social intention is recoverable from it" [367]. In such cases, verbal or other non-verbal communicators accompany pointing. Interestingly, in children, there have been proposed two cooperative motives for pointing: helping and sharing [368].

Legs and Feet. Evaluation of how humans position themselves in an environment and how they interact with others through legs and feet communication might give information regarding one's behavior or attitude towards a given scenario. In the context of deception detection, there are three types of cues observers seem to rely on: non-verbal visual (*e.g.*, eye contact, head, hand, foot, and leg movements), paraverbal (*e.g.*, voice pitch, unfilled pauses), and verbal content (*e.g.*, usage of first-person pronouns) [369]. Although we carefully monitor speech, hands, and facial movements while lying, foot, and leg movements are out of conciseness and portray lying related emotions more accurately [370]. Additionally, both genders have marked changes in foot and leg movement when lying: either moving more or less while lying [371]. Another concept linked to this behavioral aspect is proxemics. Proxemics is a socially and culturally influenced space distance usage in social interaction [372]. Such distance varies with the type of interaction, but the overall general conformable distance is unconsciously assumed. In this context, people are drawn towards liked targets and move away from disliked ones [373]. As such, it is feasible to consider that invasion of such interpersonal space will promote

leg and foot movements. Another vital aspect of proxemics is the predictive information it might contain. The evaluation of proxemic behaviors is capable of predicting aggressive and hostile behaviors towards black people [374]. Regarding interpersonal space invasion, when an adult makes a series of approaches and withdrawals, in child-adult dyad, it correlates with the child performing also approaches and withdrawals, regulating their personal space distance [375]. This corroborates Halls' proposed interpersonal distance model [372, 376] and shows that such is seen from early stages in life. On the other hand, when an interviewer stepped further than the expected interpersonal distance space, interviewees tried to re-established a new conformable distance, which was a compromise of the original one and the distance assumed by the interviewer [377].

2.3.7.2 Deep Learning Methods

To analyze a person's pose, the deep learning area more closely related is **Pose Estimation** (or **Keypoint Detection**), where methods output human keypoints, associated with different parts of the human body. Typically, these approaches can identify up to 17 human body keypoints: nose, right and left eye, ear, shoulder, elbow, wrist, hip, knee, and ankle. This evaluation can be done at frame and video level, and methods can output pose estimation in 2D or 3D. An overview of two key approaches for 2D pose estimation is shown in Figure 2.14.

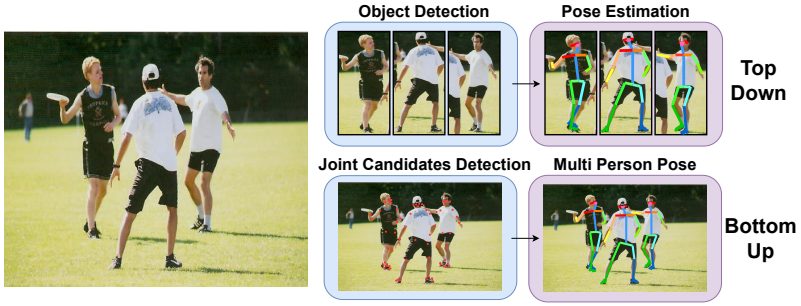


Figure 2.14: Overview of top-down and bottom-up approaches for multi-person 2D pose estimation.

For 2D pose estimation, we can start by considering **single-person** approaches, where the goal is to find keypoints given as input a cropped image of a person. In this setting, methods can predict keypoints via an end-to-end approach to learn a mapping from the input image to body joints of human body models [378, 379, 380] (**regression-based**), or focus on predicting approximate locations of body parts and joints with supervision of heatmaps [381, 382, 383] (**heatmap-based**). Further improves to heatmap creation can be based on architecture changes, namely through feature representation at different levels, *i.e.*, high-to-low [384, 385] and low-to-high [386, 387] variations.
When considering images containing multiple people, 2D pose estimation requires the detection of people and their keypoint estimation (**multi-person**). The main approach in this context focus on first extracting people bounding boxes, followed by single-person methods for keypoint detection (**top-down**). In these cases, the novelty relies on improving human pose estimation [388, 389, 385] and improving estimation of partially

occluded/truncated people [390, 391, 392]. Alternatively, methods can start by detecting body joint candidates to assemble a pose estimation (**bottom-up**). The main variations come from first selecting joint candidates, followed by joint assembling (**two-stage**), or obtain pose detection and group assignments simultaneously (**one-stage**). Regarding **video settings**, the inclusion of temporal notion (Recurrent Neural Network (RNN) models) [393, 394], movement (optical flow) [395, 396], and combining with pose tracking [397, 398] has been reported for improving pose estimation.
For 3D pose estimation, instead of outputting a stick figure (2D), methods can output 3D stick figures (**Skeleton-only**) or 3D mesh of a person's body (**Human Mesh**). For single-view multi person pose estimation, works can also be grouped in **top-down** and **bottom-up**, similarly to 2D pose estimation. While top-down approaches are more computationally expensive, especially in crowded areas, and disregard global information when inferring human skeleton, its applicability for body mesh prediction is more reliable than bottom-up approaches, given the easiness to incorporate 3D single-person human mesh recovery method to top-down methods. The inclusion of **multi-views** has also been reported [399, 400, 401], namely to deal with occlusions, and depth and point could sensors [402, 403], inertial measurement units [404, 405], and radio frequency [406, 407] have also been considered as alternative inputs for 3D pose estimation.
For easy to use and available methods, Detectron2 [408], AlphaPose [409], and OpenPose [410] are available on GitHub. Regarding a better overview of pose estimation topics, the surveys of Zheng *et al.* [411] and Chen *et al.* [412] summarize the key concepts and works of this area.

2.3.7.3 Datasets and Metrics

The most commonly used dataset for human keypoints detection is the person subset of COCO [247], which contains over 328 000 images and 200 000 person instances labeled with 17 keypoints, where methods are required to localize person keypoints in challenging and uncontrolled conditions. There are multiple versions of COCO, which differ in the train, validation, and test-dev sets, with **COCO 2017** being the most recent one that contains 118 000, 5 000, and 41 000 images for these three sets, respectively. Examples of images used in the COCO validation set are seen in Figure 2.15.



Figure 2.15: Representative images from COCO (top row) and MPII datasets (bottom row).

Other widely used dataset is **MPII** Human Pose dataset [413], containing around 25 000

Table 2.8: COCO.TD and BU refer to Top-down and Bottom-up approaches, respectively.

	Method	AP	AP.5	AP.75	AP(M)	AP(L)
TD	MSPN [414]	77.1	93.8	84.6	73.4	82.3
	DARK [415]	77.4	92.6	84.6	73.6	83.7
	PRM [385]	78.6	94.3	86.6	75.5	83.3
BU	HGG [416]	67.6	85.1	73.7	62.7	74.6
	PersonLab [417]	68.7	89.0	75.4	64.1	75.5
	HigherHRNet [418]	70.5	89.3	77.2	66.6	75.8

Table 2.9: MPII.The table is divided by the type of approach used, the top refers to Heatmap and bottom to Regression.

Method	Hd	S	E	W	Hp	K	A	M
Zhang [419]	98.6	97.0	92.8	88.8	91.7	89.8	86.6	92.5
MSPN [414]	98.4	97.1	93.2	89.2	92.0	90.1	85.5	92.6
PRM [385]	98.5	97.3	93.9	89.9	92.0	90.6	86.8	93.0
TFPose [379]	98.0	95.9	91.0	86.0	89.8	86.6	82.6	90.4
FPD [420]	98.3	96.4	91.5	87.4	90.0	87.1	83.7	91.1
Luvizon [421]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2

images from a variety of real-world activities, such as recreational and household. There are over 40 000 person instances with annotated body joints, among which 12 000 instances are used for testing and the remaining for training. The dataset covers 410 human activities and each person annotation is composed of 16 keypoints ids.

Metrics. For 2D pose estimation, Percentage of Correct Keypoints (PCK) and Average Precision (AP) are the *standard* metrics. PCK is used to measure the accuracy of localization of different keypoints within a given threshold. Typically, the threshold is set to 50% of the head segment length of each image, represented as PCKh@0.5. The AP metric is based on Object Keypoint Similarity (OKS) [247] that evaluates the similarity between predicted and ground-truth keypoints, and is represented under different OKS thresholds N as AP@N, with OKS expressed by:

$$\text{OKS} = exp(-\frac{d_i^2}{2s_i^2k_i^2}), \quad (2.8)$$

where i is the keypoint index, d_i is the euclidean distance between predicted and ground-truth, s_i is the object's scale, and k_i is a constant for a specific keypoint. Regarding their use, PCKh@0.5 is commonly used to report results on MPII and AP (at different thresholds) for COCO, incremented by distinction of AP for medium ($32^2 <$ area $< 96^2$) and large objects (area $> 96^2$).

State-of-the-Art Results. To provide an overview of the best performing models for keypoint detection, we present the reported results on COCO and MPII in Tables 2.8 and 2.9, respectively. The overall best performing methods in COCO relate to top-down approaches, while heatmap-based approaches are superior in MPII.

2.3.8 Soft Biometrics

2.3.8.1 Psychology Background

Gender. Women and men behave differently. Such can be seen from infancy when gender expression is exhibited. Gender expression refers to how a child expressed his behavior stereotypically, either masculine or feminine, through the choice of clothes, interests, and friendships [422]. The absence of gender expected social norms is linked to poor mental health [423], mental illness, and suicide, both in adolescence and adulthood [424]. Another area where gender models behavior is in social relationships. In male children, those who exhibit high aggression levels, primarily against other boys, tend to be recognized as central members in their classroom [425]. In the same context, girls who exhibit similar behavior will tend to face rejection [426] and attain less central roles in their social groups. The expression of internal problems is, in some cases, translated into aggression. This aggressive expression is more prevalent in boys than girls, even though the prevalence of such problems, such as anxiety, is more prevalent in girls [427]. Furthermore, depressive symptoms are considered a significant risk factor for aggression and delinquency, more so in girls than for boys [428]. Behavior gender differences have also been seen in some specific areas. Regarding exploration, women present a more calculated approach to explore, which results in increased pausing behaviors [151]. In the context of physical human interaction, men respond less positively to touch when compared with women [94]. Women tend to be more empathic than men [429] and more reactive to unpleasant events [430], which may trigger different behavioral responses to the same scenario. Regarding comfortable interpersonal distance, which is the space between people during an interaction, height was positively associated with said distance for men but not for women [431]. In the context of what type of weapon people carry, men tend to carry proactive weapons, like guns or clubs, while women tend to carry reactive ones, such as pepper spray [432].

Age. Age is another factor that modulates the way a human behaves. One aspect where age might affect behavior is through decreased muscle strength and lower limb power, which affects mobility factors such as stepping frequency and gait velocity [433]. The reduced mobility might affect human exploration and thus condition human behavior. This aspect is corroborated by older adults preferring to perform slower-paced activities [434] and do them accompanied by other people [435]. On a similar note, only a small percentage of older people (10-15%) meet recommended guidelines for engaging in moderate physical activity levels [436]. Attitudes, which have an emotional component, influence human behavior [15]. As such, it is feasible to consider that emotional variance might contribute to behavioral differences. One aspect where age might influence behavior is, precisely, through emotional variance. The literature reports that older adults show decreased negative affect arousal [437] while paying more attention to positive emotions [438]. Similarly, older adults are better at emotion control [437] and more effective in emotion regulation [439] since they tend to spend less time with negative stimuli than younger adults [440]. In word fragment completion tasks, older adults show bias towards

neutrals words, while younger adults favor positive ones [441].

Body Mass. Evaluation of one's body mass might give some information regarding his behavior or those around them. The unprecedented spread of obesity and overweightness has created the need to raise awareness from the World Health Organization [442] to find meaningful ways to tackle this issue. Obesity and overweightness are complex health issues with personal, social, and economic impacts [443]. In addition to health-related issues, such as diabetes, cardiovascular diseases, or kidney failures, obesity negatively impacts stigmatization and well-being [444]. Furthermore, experiencing such stigma early in life is linked to an adverse eating disorder, with health consequences extended into adulthood [445]. Another concept linked to behavior and body mass is depression. Depression might model human behavior since its symptoms cause significant distress or impairment in social interactions [446]. The association between body mass and depression has been shown, where depressed individuals tend to be less physically active and have worse dietary habits, contributing to a high Body Mass Index (BMI) [447]. However, the link between body mass and depression might even be stronger since there is evidence that obesity is a causal risk factor for depression [448], and depression, in itself, may lead to obesity [449]. It is important to note that not only overweight is, potentially, influential in human behavior. Underweight, especially in children, is associated with nutritional deficiencies, psychiatric disorders, and negatively impacts adulthood's health [450, 451]. Some differences might be observable between overweight and underweight people. Extroversion and psychoticism (aggressiveness) are more associated with overweight and neuroticism (associated with more likeliness to experience anxiety, worry, fear) with underweight [452].

Height. The relevance of the height can be seen, primarily, when interacting with others. Humans tend to seek an optimal range of distance in interaction, and a violation of said space may trigger reactions oriented to re-establish said space [453]. However, the range of distance is changeable, depending on what person they are interacting with. Body size and social status influence this range, and there is a positive relation between height, dominance, and social status [454]. Similarly, masculinity, dominance, and height seem to be associated with each other [455]. It has been hypothesized that height influences social status perception, linking taller people with higher social status [456]. The height, in particular, is positively associated, for men, with comfortable interpersonal distance [431]. Height is also influential in the perception of others. In cases of dwarfism, people tend to associate little people as less likely to be hired for a job, less attractive, and less likely to be in a relationship [457]. Another height associated perception is that shorter men tend to be more sensitive to dominance in voice cues [458].

Race and Ethnicity. The race is generally based on physical characteristics, such as skin color or facial features, while ethnicity is often defined based on nationality [459]. The effect of ethnicity on human behavior can be seen in children's development since their environment, and cultural contexts mold their behavior [460]. An example of this is the difference seen in spanking, where, for European Americans, it is predictive of maternally anti-social behavior. In contrast, for African Americans, such is not seen [461]. In the

same context, Hispanic mothers' higher authoritarian parenting approach translates into increased parenting stress [462]. Similar to body mass, ethnicity might influence not only one's behavior but also others. For opioid users, stigma expression is more significant for white people than black people [463]. In America, black people have higher incarceration rates and higher rates of violent crime victimization than white people [464]. Regarding emotion recognition, race appears to affect face memory and face emotion identification similarly in controls and schizophrenic patients [465]. Race and ethnicity can also influence people's interactions. In dyadic interactions, same race dyads exhibited significantly more friendly non-verbal behavior and had higher performances than those in interracial dyads [466]. Similarly, East and West people have different communication styles [467] expressed by Americans being less attentive to indirect language cues than East Asians. Race and ethnicity also influence facial expression recognition [468].

2.3.8.2 Deep Learning Methods

Given the settings considered for Human Behavior (surveillance), the most suited area is **PAR**. Their methods are evaluated on surveillance setting images, analysing various pedestrian attributes (and not only soft biometrics), with the most commonly soft biometrics included in PAR being gender and age.
One of the main approaches for PAR is to analyze the relation of attributes and different body parts [469, 470, 471], or relating pose information with body part localization [472, 473, 474]. Another key approach in PAR method is considering the relation of various attributes: learning the relationship, dependency, and correlation between attributes [475, 476], considering a hierarchical feature embedding [477], enhancing the localization of attribute-specific areas [478], representing attribute relation via graphs [479], or considering image quality and face availability [480]. Other viable strategies to improve PAR are based on attention maps at different scales [481], analyzing relevant image patches [482], focusing on different feature levels [483], losses based on heatmaps [484], feature augmentation [485], using semantic information [486, 487], and unified learning to improve PAR generalization [488].
For specific soft biometrics, the work done is more centered on cooperative settings, with closed up faces. For **gender**, work are mainly based on variations of layer or block arrangement [489, 490, 491], or grouping various CNN and compute similarity [492, 493, 494]. For **age**, works tend to follow a similar pattern [495, 491, 490, 494, 493], and the same occurs for **race/ethnicity** evaluation, as some age prediction methods also display this capability [494, 493, 495]. Regarding **height** estimation, we have assessing head and toe estimation [496], considering viewpoint position or ground plane location [497], using depth information [498], and pose estimation [499]. For **weight** estimation, work are based on face features [500], full-body and relating with body parts [501], or via silhouette analysis [502]. However, all the methods here discussed (aside PAR) work on cooperative settings, and are not designed to reliably perform in unconstrained scenarios (surveillance settings).
For additional information on the soft biometrics related topics, we refer to surveys on

soft biometrics [503], ethinicity [504], and PAR [505].

2.3.8.3 Datasets and Metrics

For PAR, the most commonly reported datasets are PETA [506], PA-100K [507] and RAP [508]. PETA contains 19 000 images, divided into three subsets (9 500 for training, 1 900 for validation, and 7 600 for testing), with data from various datasets, with varying resolution, indoor/outdoor scenarios, and illumination, containing 61 binary and 4 multi-class attributes. RAP dataset is collected from shopping mall surveillance cameras, with 15 frames per second, resulting in 33 268 training and 8 317 test images, with 69 binary and 3 multi-class attributes. RAPv2 [509] contains 84 928 images, from 2 589 people, with body part location annotations, maintaining the same attributes as the original RAP. PA-100K is the biggest PAR dataset, composed of 100 000 images from indoor surveillance cameras, with 26 binary attributes, from almost 600 scenes, and with over 18 000 person tracklets.
Considering more cooperative settings, we highlight datasets that can be used for specific soft biometrics or a combination of several. For **gender estimation**, ChaLearn Looking at People - Fair Face Recognition [510] is a potentially usable dataset, with around 153 000 facial images from 6 100 unique identities, containing men and women with bright and dark skins. For **age prediction**, VGGFace2 [511] has images with faces in different viewpoints and pictures with different ages, from the same people. The whole dataset is composed of 3.31 million images of 9 131 subjects, averaging 362.6 images per subject. Combining **gender and age estimation**, Image Movie DataBase (IMDB)-Wiki [512] has over 500 000 images of men and women, with age mean distribution at around 40 years. LFW [513] has 13 233 images, from 5749 people, and can be used for **gender and ethnicity estimation**, although its primary purpose is to be used for face verification. UTKFace [514] is a dataset for **gender, age, and ethnicity evaluation**, composed of 20 000 face images with variation in pose, facial expression, illumination, occlusion, and resolution. For **BMI estimation**, Face In the Wild for BMI analysis Face In the Wild for BMI analysis (FIW-BMI) [515] and MORPH-II [516] are valid options: FIW-BMI contains 7 930 wild face images of 4 881 individuals along with the labels of gender, height, and weight, while MORPH-II dataset is a composition of mugshots from 16 to 77 years of age, averaging four images per person. It contains 55 000 images taken over five years. Examples of surveillance (PAR) and cooperative settings are presented in Figure 2.16.



Figure 2.16: Representative images from PA-100K (top row, first four images), LFW (top row, last 3 images), and UTKFace datasets (bottom row).

Table 2.10: State-of-the-art results in multiple PAR datasets.

Method	PETA	PA-100K	RAP	RAPv2
MsVAA [481]	84.6	-	-	-
VAC [484]	-	79.1	-	-
ALM [483]	86.3	80.7	81.9	-
JLAC [479]	87.0	82.3	83.7	-
VFA [485]	86.5	81.3	81.8	-
MSCC [517]	-	82.1	82.2	80.2
SB [518]	85.2	81.6	80.8	77.4
DAFL [487]	83.5	83.5	83.7	81.0
SSC_{soft} [486]	86.5	81.9	82.8	-
UPAR [488]	88.4	84.8	-	79.9

Metrics. To deal with unbalanced distribution of attributes, mA is the most common metric used in PAR. For each attribute, mA relates the accuracy classification of positive and negative samples, averaging over all attributes, expressed by [508]:

$$\text{mA} = \frac{1}{2N}\sum_{i=1}^{L}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right), \quad (2.9)$$

where L is the number of attributes, TP_i and TN_i refer to correctly predicted positive and negative examples, respectively, and P_i and N_i are the number of positive and negative examples, respectively. Even though mA deals with unbalanced distributions, it treats attributes independently, ignoring inter-attribute correlation. As such, four different metrics tend to complement mA to capture example-based evaluations (*i.e.* consistency of predictions for a given pedestrian image) [508]:

$$\begin{aligned} \text{Acc} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|}, \quad \text{Rec} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|}, \\ \text{Prec} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|}, \quad \text{F1} = \frac{2 * Prec * Rec}{Prec + Rec}, \end{aligned} \quad (2.10)$$

where N is the number of examples, Y_i is the ground truth of positive label, and $f(x)$ refers to predicted positive labels.

State-of-the-Art Results. To provide an overview of the best performing models for PAR, we present the reported results on PETA, RAPv2, and PA-100K in Table 2.10, with semantic information use and unified learning for improved PAR generalization being the best performing approaches overall.

2.4 Current Limitations and Future Directions

After reviewing the state-of-the-art for different areas related to the behavioral features considered to analyse human behavior non-cooperatively, several limitations arouse.

When considered the analysis of **verbal interactions**, the area of ASD is mainly focused

on datasets with cooperative settings, with good audio quality and guaranteed face access, which motivates the development of models based on these premises: combination of audio with face image. However, given the importance of ASD in surveillance settings for this survey, where face and audio may not be reliably accessed, similar to WASD [3] **more work should be done on creating challenging datasets and models that perform without assuming cooperative settings for ASD**.

Regarding **physical interactions**, the area we focus on to refer to related work is action recognition. From this, two issues can be highlighted: 1) **there is not a clear area related to physical human interaction**, aside occasional sporadic works; and 2) **action recognition is not a single area and can be divided into (at least) 4 different parts**, as we discussed in Section 2.3.2. This shows the range of work that can be explored regarding datasets/work more focused on **specific human physical interaction** (and relate to their importance in human behavior) as well as creating **surveys more focused on the different types of action recognition areas that exist**. To the best of our knowledge, there is not yet surveys/works that highlight such division, with our survey being the pioneer towards clarifying existing divisions in action recognition.

For **proximity** analysis, our discussion is centered on depth estimation, which is not an area with direct applications for the considered behavioral analysis. Ideally, we would have works that focus on perceiving where people are in a given scene and how close/far they are from a relative reference point. As such, works on **surveillance/non cooperative settings with physical distance annotations to motivate the development of models to identify (subject) proximity** are worth exploring. Note however, that these type of works require a fixed camera position with calibration for real physical distance estimation, which may justify the reduced interest in this topic: models may have reduced scalability and portability to perform in different (but similar) settings. Nevertheless, we highlight here this topic as a area of improvement given the goal of our behavior analysis approach.

In the **clothing** analysis, there is a clear area with state-of-the-art datasets and models: clothing/fashion. However, datasets are more related to online shopping usage, meaning that the conditions for clothing assessment is done in close-up images with cooperative conditions. Given the premise of the considered behavioral analysis, these datasets are not entirely suited, which led us to discuss alternatives such as PAR datasets, in Section 2.3.6. For these reasons, **assessing the resilience of standard clothing/fashion methods in more challenging sets (such as PAR), and proposal of new models that can perform in such settings** is an area worth exploring. Furthermore, even in the context of fashion datasets, a **standardization of metrics and datasets** used across fashion models is necessary to compare their performances and ease the identification of areas of improvement towards future works.

On a more general analysis regarding all behavioral features, the vast majority is **still evaluated in cooperative settings, with the exceptions being pose estimation and relative position (MOT)**. Although we mention alternatives to analyse behavioral features in more wilder (surveillance) conditions, such as PAR datasets, the goal of these is

on multi-attribute classification, thus PAR models are not specialized at classifying specific features. For this reason, focus on **creating datasets/models more targeted towards behavioral analysis in surveillance conditions** is a area worth exploring. Finally, we also highlight that using PAR datasets is a viable strategy for immediate assessment of the resilience of current models in surveillance conditions for various behavioral features.

2.5 Conclusion

Analyzing human behavior is a difficult task that requires action, motive, and context interpretation, with several works proposing human behavioral analysis from different perspectives. This survey discusses these perspectives and introduces a new one that focuses on assessing specific human features, without requiring human cooperation, based on psychology studies. We discuss the deep learning areas capable of evaluating said features, highlighting their limitations and future directions, in particular methods/datasets that specifically evaluate them in wilder conditions, representing areas worth exploring in the future.

Chapter 3

Is Gender "In-the-Wild" Inference Really a Solved Problem?

Soft biometrics analysis is seen as an important research topic, given its relevance to various applications. However, even though it is frequently seen as a solved task, it can still be very hard to perform in wild conditions, under varying image conditions, uncooperative poses, and occlusions. Considering the *gender* trait as our topic of study, we report an extensive analysis of the feasibility of its inference regarding image (resolution, luminosity, and blurriness) and subject-based features (face and body keypoints confidence). Using three state-of-the-art datasets (PETA, PA-100K, RAP) and five Person Attribute Recognition models, we correlate feature analysis with gender inference accuracy using the Shapley value, enabling us to perceive the importance of each image/subject-based feature. Furthermore, we analyze face-based gender inference and assess the pose effect on it. Our results suggest that: 1) image-based features are more influential for low-quality data; 2) an increase in image quality translates into higher subject-based feature importance; 3) face-based gender inference accuracy correlates with image quality increase; and 4) subjects' frontal pose promotes an implicit attention towards the face. The reported results are seen as a basis for subsequent developments of inference approaches in uncontrolled outdoor environments, which typically correspond to visual surveillance conditions. The work presented in this chapter has been published in [480].

3.1 Introduction

Soft biometrics analysis has been gaining increased interest over the last years, given its applicability to aid in person identification while requiring little to no cooperation from observed subjects and being robust to low quality data [519, 520]. This interest is also enhanced by the availability of massive amounts of data in wild conditions, provided by surveillance cameras and footage gathered by hand-held devices.

Even though soft biometrics inference is frequently seen as a relatively easy task, this work reports results that point to the opposite. In particular, we focus our extensive analysis on the *gender* trait, posing the following research questions:

- **How well solved is *in-the-wild* gender inference?**
- **What features have the most relevance for gender inference accuracy?**
- **Given the face importance in gender inference, what conditions justify its use?**

Gender information can be retrieved from an image containing the human silhouette and face. As such, the conclusions yielding from the answers to the above questions might apply to other soft biometrics traits with similar characteristics, such as *age*, *body mass*, or *ethnicity*.

Gender inference *in-the-wild*, such as in surveillance environments, faces various challenges (*e.g.*, occlusions, subjects' pose, image resolution, and lighting conditions) that strongly decrease the state-of-the-art deep learning-based models' accuracy. When comparing the conditions between cooperative and *in-the-wild* environments, subjects' pose and image quality are key discriminating factors that might influence inference accuracy. As such, in this work, we correlate image and subject-based features' importance with models' accuracy to perceive the importance of each feature in gender inference. To achieve this goal, we define quantifiable image and subject-based features to categorize the datasets' quality. We define *resolution*, *luminosity*, and *blurriness* as the most influential image-based features and the Keypoints (KP) *confidence* of the *face*, *upper*, and *lower body* parts as the subject-based features. The KP data is obtained from Alphapose [521, 522, 523], a state-of-the-art pose estimator.

Given the intent to evaluate gender inference in wild conditions, we use the most challenging PAR datasets in our experiments. Furthermore, to obtain a generic deep learning model performance evaluation, we gather five state-of-the-art PAR inference models and combine their predictions to obtain a model-independent gender inference. To objectively perceive features' importance, we correlate them with the models' combined inference and analyze it using the Shapley value [524], a game-theoretic approach to explain machine learning outputs.

Another important analysis in our work regards the usability of the facial region for gender inference. Given that the face is intuitively one of the most important attributes for gender inference, several models have been proposed using only the face for this task. However, the availability of the facial region in wild situations is not guaranteed, and a subject pose might influence its importance for face-based gender inference. To evaluate this, we start by creating challenging (wild) face sets, yielding from frontal images of various PAR datasets; frontal images are obtained from a pose-based image division, using Alphapose. Then, we compare a face-based model performance (trained in Adience) and evaluate it on the created face sets. Furthermore, we assess the effect of pose in face-based gender inference, by training the model in *any* or *frontal* pose full-body images and evaluating its performance in face datasets.

According to the above points, the main contributions of this work can be summarized as follows:

- We define the most relevant image and subject-based features for gender inference and relate their importance with image quality;
- We describe an approach to obtain a novel wild face dataset from various PAR sources, using a pose estimation method for head detection and ROI definition;
- We evaluate the importance of facial regions for *in-the-wild* gender inference, relat-

ing it with image quality;

- We report an implicit face attention mechanism for gender inference, via frontal pose image training.

The remainder of this paper is organized as follows: Section 3.2 summarizes the most relevant face and body-based methods used for gender inference; Section 3.3 describes the methodologies used for our analysis and Section 3.4 discusses the results obtained. The main conclusions and future work are presented in Section 3.5.

3.2 Gender Inference State-of-the-Art

3.2.1 Face-Based Methods

The facial region is commonly used for gender inference, with several datasets announced and models designed for this task. Levi and Hassner [491] proposed a simple CNN model, with five layers, given the nature of age and gender inference (small number of classes). Zhan *et al.* [490] used Residual Networks of Residual Networks (RoR), outperforming other CNN architectures. Ranjan *et al.* [525] developed HyperFace, a multi-task framework based on Alexnet [526], for face detection, fused with a CNN for face landmarks detection and gender recognition. Althnian *et al.* [527] compared the performance of VGG-16 and Support Vector Machine (SVM) in gender recognition, using hand-crafted, deep-learned, and fusion features.
The usefulness of the facial region in gender inference has been explored using real-time prediction [489], minimizing the number of parameters by combining *Conv2D* and *BatchNorm* blocks and avoiding *Fully Connected (FC)* layers. Lee *et al.* [528] described a lightweight inference approach that uses depthwise separable convolution layers to reduce the model size and save inference time.
The use of facial attributes have also been reported. Liu *et al.* [529] combined a SVM for face attribute inference with two CNN: LNET for ROI definition and ANET for distinguishing face identities, using global/local convolutions for feature extraction. Ryu *et al.* presented InclusiveFaceNet [530], using race and gender for face attribute detection. Ge *et al.* [531] used a classifier based on face attributes' correlation, dividing the face into several regions to aid correlating attributes. Makeup influence on automatic gender inference for transgender people is also discussed in the literature [532].
Several datasets have been announced [529, 533, 511, 510], all of them representing relatively controlled environments, being Adience [534] the most challenging one.

3.2.2 Body-Based Methods

It is important to consider that the facial region might not be available in surveillance scenarios, which supports the use of body-based models for gender inference. In this context, PAR is a topic broadly discussed in the literature, aiming to identify attributes

(*e.g.*, clothing, accessory usage, gender, and age) under challenging covariates such as occlusion, pose, image resolution, and luminosity.
Wang *et al.* [469] proposed a model based on a RNN encoder-decoder framework, using Long Short-Term Memory (LSTM) as a recurrent neuron. They performed joint recurrent learning by using their framework in images divided into six horizontal strips. Zhao *et al.* [470] used a similar partition strategy, combining LSTM and Batch Normalization (BN)-Inception [471] networks. Barra *et al.* [535] assessed the applicability of gait analysis for gender inference, using pose information.
Attention-based approaches are also highly popular: Sarafianos *et al.* [481] combined primary and attention classifiers outputs to learn effective attention maps at multiple scales. Tang *et al.* [483] focused on a feature pyramid structure-based model, each level with an Attribute Localization Module (ALM). ALM is inspired by Spatial Transformer Network (STN) [536], which discovers the discriminative regions for each attribute in a weakly supervised manner. Guo *et al.* [484] defined an attention consistency loss, translated by the distance between transformed attention heatmaps of original and flipped images.
The analysis of each attribute importance for pedestrian classification is also reported in the literature. Li *et al.* [475] performed global image analysis, exploring how one attribute contributes to the representation of others. They introduced DeepMAR, a learning framework that recognizes multiple attributes, exploring the relationships among them. Lin *et al.* [476] developed a re-identification model that considers the attribute importance for this goal, introducing an Attribute Re-weighting Module (ARM) that corrects attribute predictions based on the learned dependency and their correlation. Jia *et al.* [478] showed that enhancing the localization of attribute-specific areas may not be beneficial for performance optimization.
For additional information on this topic, Wang *et al.* [537] review the main divisions of PAR approaches, providing a brief summary of the state-of-the-art.

3.3 Methodologies

An overview of our processing chain is shown in Fig. 3.1. We divide three well known PAR datasets (PETA, PA-100K, and RAP) into three image quality levels, based on image-based feature analysis. Then, we correlate image and subject-based features with the gender inference accuracy of a meta-model, in order to perceive the image/subject-based features that favor/penalize overall accuracy. The used meta-model is a combination of five state-of-the-art PAR models. Furthermore, we analyze the importance of the facial region in gender inference, evaluating a face-based model performance in wild conditions and examining the importance of pose in face-based gender inference.

3.3.1 Methods and Datasets

For gender inference analysis, we use five models with publicly available implementations. We consider DeepMar [475] and StrongBase [478], which aim to infer pedestrian

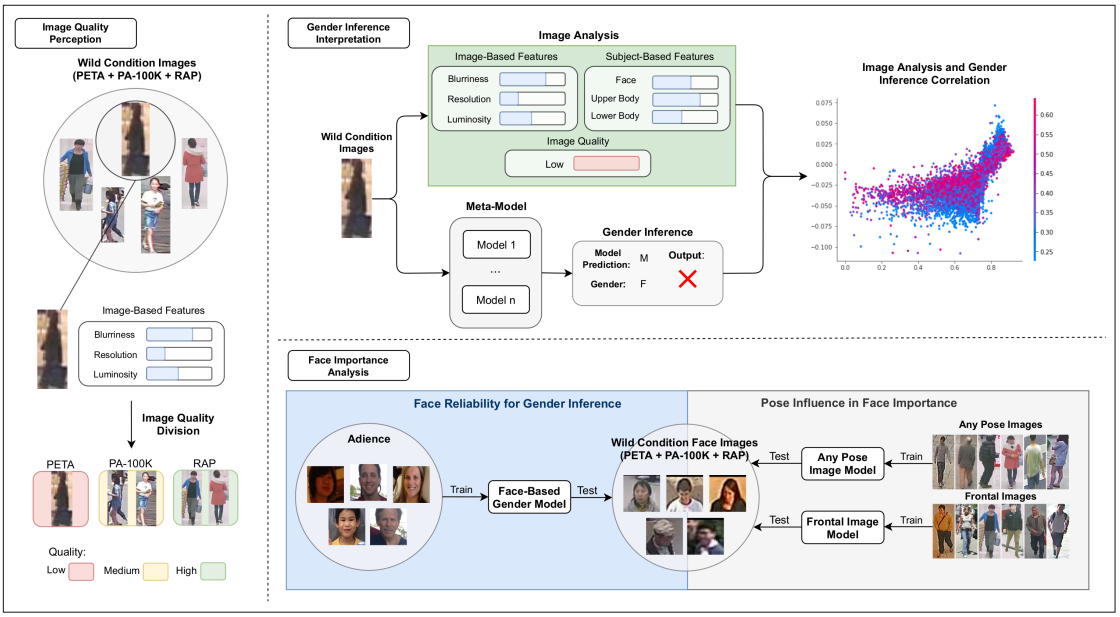


Figure 3.1: Overview of the used approaches for image and subject-based features correlation with gender inference. We divide the work into three major phases: 1) image quality perception; 2) gender inference interpretation; and 3) face importance analysis. We divide the PAR datasets considering image-based features and correlate image and subject-based features to gender inference accuracy of a meta-model. This model is a combination of five PAR models. The face importance is analyzed in two ways: by evaluating its reliability in wild conditions and assessing pose influence in face-based gender inference.

attributes based on global image analysis, and VAC [484] and ALM [483], which consider attention mechanisms for this task. Additionally, we consider the PAR component of the re-identification method ARP [476]. Regarding the models backbones, all methods are based on ResNet50 [538], except for ALM, which is based on BN-Inception [471].

Regarding the selection of datasets, our goal was to use a set of images with a wide variability range to reproduce, as much as possible, *in-the-wild* conditions. For this reason, we use PETA [506], PA-100K [507], and RAP [508], three state-of-the-art PAR datasets. The chosen datasets clearly display different levels of quality, as discussed in our experiments.

3.3.2 In-the-Wild Major Variability Factors

Our primary goal was to perceive what is a *good quality* image, in wild conditions, for gender inference. This requires measuring the factors that typically degrade gender inference. We propose a division in image and subject-based features, described in the following subsections. Examples of image and subject-based features analyzed are shown in Fig. 3.2, displaying contrasting examples of each feature.

3.3.2.1 Image-Based Features

We consider *blurriness*, *luminosity*, and *resolution* as the major image-based factors for gender inference. We retrieve image resolution by multiplying its width and height. For luminosity, we obtain the values of red, green, and blue channels to measure the perceived brightness [539]. Blurriness yields from the convolution of the images with a Laplacian



Figure 3.2: Dataset examples of the evaluated features, divided into image (*luminosity* and *blurriness*) and subject-based (*head*, *upper body*, and *lower body* availability) features. Each row corresponds to one of the three PAR datasets. Each column contains opposite cases of each feature, with the left one representing *bad quality* (with feature value closer to its minimum) and the right one representing *good quality*. For the blurriness feature, high feature value represents *bad quality*. All images are resized to the same resolution for visualization purposes.

kernel, taking the variance as result.

3.3.2.2 Subject-Based Features

We use Alphapose [521, 522, 523], an accurate multi-subject pose estimator, to extract subject-based features. Given that the used datasets contain subjects' ROI, we change the Alphapose implementation to ignore its You Only Look Once (YOLO) [540] detector input. This change translates into a more coherent pose estimation. In exceptional cases, where the image was not trimmed to the subject, we keep the original implementation. The Alphapose outputs 51 KP for each image, corresponding to the image coordinates x and y, and the confidence score for each of the 17 COCO [541] KP: nose, left and right eye, left and right ear, left and right shoulder, left and right elbow, left and right wrist, left and right hip, left and right knee, and left and right ankle. We group the KP into three regions: *face*, *upper*, and *lower body*. *Face* confidence is the mean of nose, eyes, and ears KP confidences. *Upper body* confidence is based on shoulders, elbows, and wrists, while *lower body* confidence uses the values of hips, knees, and ankles.

Regarding subject poses, we consider three values: *frontal*, *sideways*, and *backside*. We differentiate a subject facing forward or backward using pose estimation information. If the rightmost shoulder, from the top left of the image, is the left shoulder, the subject faces forward; otherwise, the subject is facing backward. A subject is considered *sideways* if the shoulder length with respect to the upper body height is less than 0.5. Shoulder length is the absolute difference between the right and left shoulders, and the upper body height

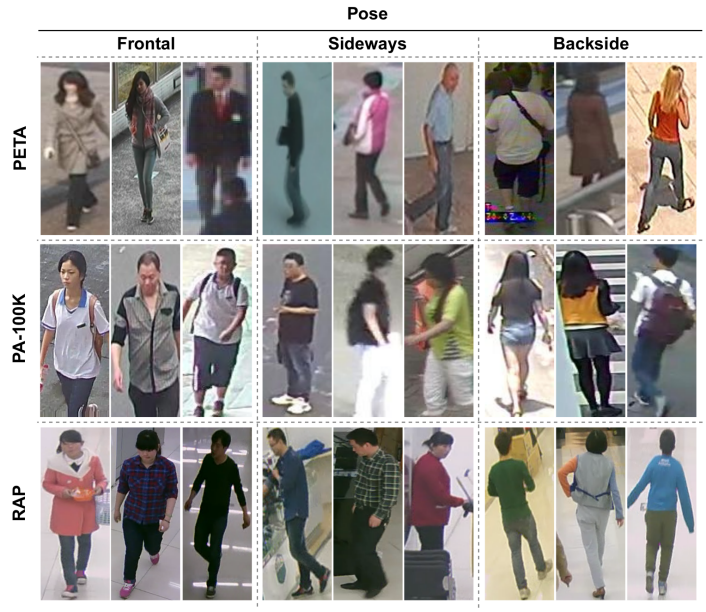


Figure 3.3: Examples of images considered as *frontal*, *sideways* and *backside*. Each row is related to one of the three PAR datasets and each column regards examples of the respective pose.

is the difference between the shoulder and hip. Some examples of images considered *frontal*, *sideways*, and *backside* are presented in Fig. 3.3.

3.3.3 Shapley Value

The Shapley value, coined by Shapley in 1953 [524], is a cooperative game theory-based method used for assigning payouts to players, depending on their contribution towards the total payout. The use of Shapley values in our experiments is linked to our interest in analyzing how each feature contributed to gender inference and if it contributed positively or negatively. The higher the absolute Shapley value is, the higher the feature influence. We associated a value of 1 to instances where all models correctly inferred the gender and 0 otherwise. As such, positive values represent cases that influence correct gender inference, while negative values contribute to gender misinference. We correlate the proposed feature values with gender inference using the SHAP framework [542] and the *TreeExplainer* model [543], from a publicly available implementation [544].

3.3.4 Head Detection

Given the varying image quality, we use pose estimation to aid in head detection. Based on the proposed subjects' pose division, we crop the head regions from frontal images. First, we obtain the coordinates x and y of the right and left ears and assume that the head's central point is given by the arithmetic mean of such coordinates. Then, head ROI are drawn using the top-left and bottom-right bounding boxes coordinates, centered in the previously calculated head's central point, always considering the head ROI height as $\frac{2}{9}$ of the whole body silhouette height. Head bounding boxes drawing is based on Detectron [408]. We provide the steps involved in head detection in Fig. 3.4, illustrating pose



Figure 3.4: Phases of the head detection process, displaying pose estimation and head bounding box size. Faces are obtained exclusively from the frontal images of each dataset.

estimation and head ROI definition.

3.3.5 Face Importance Analysis

We evaluate face importance in gender inference by training a face-based gender model in full-body images and obtaining the mA [508] in faces of frontal images of the corresponding dataset. The mA criterion is the mean classification accuracy of the positive and negative samples for each attribute, averaged over all attributes:

$$mA = \frac{1}{2N} \sum_{i=1}^{M} \left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i} \right), \quad (3.1)$$

where N is the number of examples, M is the number of attributes, P_i is the number of positive examples and TP_i is the number of correctly predicted positive examples for the i^{th} attribute; TN_i and N_i are defined analogously.
To obtain an objective measure of Face Importance (FI), we transform the mA values of frontal image faces (mA_f) as follows:

$$FI = \frac{mA_f - 50}{mA_{max} - 50}, \quad (3.2)$$

where the value 50 represents the model inference randomness, and mA_{max} is the mA value obtained when the model is evaluated in full-body images. FI value is enclosed in the $[0, 100]$ interval, directly corresponding to the face importance in each evaluated context.

3.4 Experiments

3.4.1 Dataset and Evaluation Metrics

Our experiments are carried out in three state-of-the-art PAR datasets: PETA [506], PA-100K [507], and RAP [508]. The PETA dataset has 19,000 images, each with 61 binary attributes. It is divided into three subsets: 9,500 for training, 1,900 for validation, and

7,600 for testing. Based on this dataset's protocol, we report the results on the 35 attributes with a ratio of positive labels higher than 5%. The PA-100K dataset is composed of 100,000 images from outdoor surveillance cameras. It is split into 80,000 images for training, 10,000 for validation, and 10,000 for testing. Each image has 26 attributes. The RAP dataset contains 41,585 images, each annotated with 72 attributes. Following the official protocol [508], we split the dataset into 33,268 training images and 8,317 test images and report the results on the 51 binary attributes with a positive ratio higher than 1%.

For face-based gender inference, we use Adience [534], a dataset composed of face images obtained from cameras, smartphones, and tablets. This dataset contains varying quality factors such as facial expressions, blurred samples, partial occlusions, and head pose variations. It is composed of 26,580 unconstrained images of 2,284 subjects. We use the original in-plane aligned version of the faces [534].

To present our results, we use the label-based metric mA [508]. For state-of-the-art comparison purposes, we use accuracy when evaluating our face-based gender model (described in Section 3.4.2.2) in the Adience dataset.

3.4.2 Implementation Details

3.4.2.1 Baseline Methods

The models used were VAC [484], ARP [476], ALM [483], StrongBase [478], and DeepMar [475]. The VAC model train/test transformation for PETA and RAP is based on the WIDER dataset [545] transformation. Since the attribute weights for weighted sigmoid cross entropy [475] loss were only defined for PA-100K, we retrieve the attribute weights for PETA and RAP from the same source the authors used to obtain the PA-100K ones. To allow gender inference comparison in our experiments, we change the PETA dataset evaluation for the StrongBase evaluation implementation, adjusting the output-related loss without modifying attention-related ones. For ARP, we only consider attribute loss, ignoring the re-identification one. As the available implementation did not consider PETA, RAP, and PA-100K dataset evaluation, we introduce it following the original paper's guidelines.

The remaining models are evaluated using the original implementations. All the gender models used in our experiments are specifically trained for gender attribute only. For all models' implementations that did not consider the evaluated datasets, we incorporate StrongBase dataset reading into their implementation. For consistency purposes, all metric evaluations are also based on StrongBase's implementation.

3.4.2.2 Face-Based Gender Model

Our model is based on the StrongBase implementation [478]. The ResNet50 [538], pretrained in ImageNet [546], is used as backbone to extract image features. The model is separated into feature extraction and classification. The feature extractor is composed of ResNet50 without the *Average Pool* and *FC* layers, which were incorporated in the

classifier. The classifier has its *FC* layer replaced by a *Linear* layer, followed by a *Batch Normalization* one. Images are resized to 256×192 with random horizontal mirroring as inputs. Stochastic Gradient Descent (SGD) is used for training, with momentum of 0.9, and weight decay of 0.0005. The initial learning rate is set to 0.01 for the feature extractor and 0.1 for the classifier. Plateau learning rate scheduler is used with a reduction factor of 0.1 and a patience epoch number of 4. Batch size is set to 64 and the total epoch number of training is 30, using Binary Cross Entropy with Logits (BCELogits) as loss function.

3.4.3 Cross-Domain Performance

We evaluate the gender inference performance of each model, when trained for all pedestrian attributes, in the *within* and *cross-domain* settings. This analysis allows us to perceive the potential robustness of each model. Additionally, we train each model specifically for the gender attribute and compare its performance with the corresponding model trained for all dataset attributes. We report our results in Table 3.1.
From this experiment, we can observe that models trained exclusively with the gender label tend to have slightly better gender inference accuracy than those trained for all attributes. This is confirmed in both *within* and *cross-domain* scenarios. Therefore, in all the remaining experiments, we resort to models trained exclusively for gender.
Regarding the obtained models' accuracies, we conclude that no model is significantly better than the others, regardless of the dataset where it was trained/evaluated. However, when examining model performance within and across datasets, some differences can be observed. First, models trained in PETA or PA-100K and evaluated in other datasets tend to perform well, achieving results close to those they present when evaluated in the same dataset they were trained on. The exception is RAP-trained models, where they perform considerably worse if evaluated in other datasets. Second, considering the *within-domain* setting, RAP is the dataset where models perform the best, achieving the highest accuracy out of all the three evaluated datasets. Finally, models trained in PETA or PA-100K achieve better performances in RAP than in any other dataset, aside from the one they were trained on. PETA is the dataset where models, if trained on other datasets, perform the worst. These observations indicate that accurate gender inference is easier in RAP images and that PETA is the most challenging dataset for this task.

3.4.4 Dataset Analysis

To perceive how image-based features influence gender inference, we obtain the corresponding values for each feature/dataset. We present the normalized mean and standard deviation values in Table 3.2.
Given the low variability of luminosity across the datasets, we group them into categories exclusively based on resolution and blurriness: *low*, *medium*, and *high*. The higher quality dataset is RAP, given the high resolution value and sharpness. PETA is the dataset with lower quality, based on the low resolution values and high blurriness. Since PA-100K dataset feature values are typically between those of PETA and RAP, we consider

Table 3.1: Gender mA of different models trained and evaluated on various PAR datasets. Models named with *Gender* suffix refer to models trained solely for the gender attribute. The outperforming methods for each dataset are shown in bold.

Train \ Eval	Methods	PETA	PA-100K	RAP
PETA	StrongBase	92.80	78.98	79.47
	StrongBase-Gender	**93.13**	79.65	80.44
	ALM	91.36	76.95	80.62
	ALM-Gender	92.28	78.27	82.66
	DeepMar	91.08	72.83	71.76
	DeepMar-Gender	92.33	77.63	74.78
	VAC	92.85	78.16	78.50
	VAC-Gender	92.76	78.22	77.92
	APR	92.84	78.83	78.95
	APR-Gender	92.34	75.49	78.49
PA-100K	StrongBase	77.28	90.97	86.93
	StrongBase-Gender	75.01	90.77	88.38
	ALM	75.48	88.34	85.87
	ALM-Gender	76.59	90.34	88.55
	DeepMar	76.17	88.52	85.73
	DeepMar-Gender	74.50	90.39	85.72
	VAC	75.20	90.17	87.79
	VAC-Gender	76.85	**91.05**	88.98
	APR	76.53	89.91	87.24
	APR-Gender	75.26	90.05	86.29
RAP	StrongBase	67.67	72.25	**96.74**
	StrongBase-Gender	71.41	76.97	96.34
	ALM	76.81	81.24	95.69
	ALM-Gender	76.03	80.28	95.41
	DeepMar	69.29	75.76	95.52
	DeepMar-Gender	70.22	76.89	96.40
	VAC	66.11	68.84	96.52
	VAC-Gender	72.89	76.67	96.59
	APR	68.39	75.41	96.20
	APR-Gender	69.24	72.93	95.90

Table 3.2: Dataset image-based feature analysis. Values are normalized for the combination of the three dataset values, for each feature.

Dataset	Resolution	Luminosity	Blurriness
PETA	0.037 ± 0.024	0.432 ± 0.100	0.120 ± 0.148
PA-100K	0.062 ± 0.063	0.449 ± 0.126	0.095 ± 0.089
RAP	0.131 ± 0.083	0.407 ± 0.107	0.022 ± 0.013

this dataset as of medium quality. We present a representative quality bar of the three datasets in Fig. 3.5, illustrating some dataset examples.

3.4.5 Methods Evaluation

We evaluate the image and subject-based features of the correctly inferred cases, in comparison to all test set images, to perceive the features that contribute the most to gender

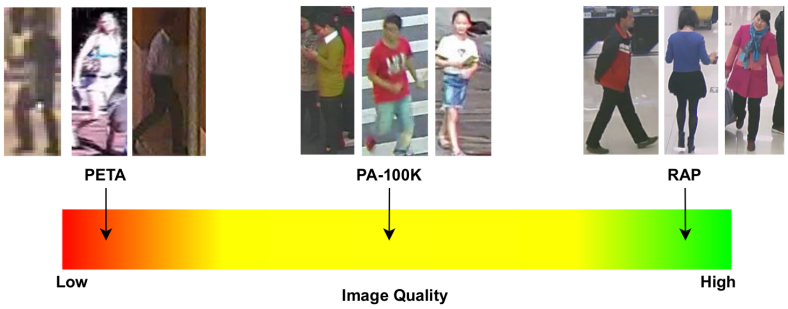


Figure 3.5: PAR dataset quality division, taking into account image-based feature values. Each dataset illustrates examples of the corresponding quality levels.

Table 3.3: Image and subject-based feature values for PETA, PA-100K, and RAP datasets. *Correct* represents the correctly inferred images (when all models inferred the correct gender value) and *All* refers to all test data. Subject pose values represent the portion of test images with the given pose.

Features		**PETA**		**PA-100K**		**RAP**	
		Correct	All	Correct	All	Correct	All
Pose	Frontal	0.387	0.383	0.363	0.355	0.350	0.348
	Sideways	0.223	0.250	0.356	0.368	0.309	0.315
	Backside	0.390	0.366	0.282	0.277	0.341	0.337
KP	Face	0.823	0.821	0.820	0.813	0.839	0.835
	Upper Body	0.781	0.775	0.773	0.765	0.820	0.817
	Lower Body	0.764	0.759	0.768	0.761	0.786	0.783
Image	Resolution	0.211	0.207	0.066	0.054	0.223	0.227
	Luminosity	0.417	0.415	0.452	0.462	0.478	0.475
	Blurriness	0.148	0.143	0.114	0.110	0.115	0.114

misinference. Here, we consider the *within-domain* setting. Given the predictions of the five methods used, we consider *correctly classified* cases exclusively when all models correctly infer the corresponding value. Results are displayed in Table 3.3, with values normalized per dataset.

The results suggest no significant differences between images where gender is correctly inferred and all test set images, for all the evaluated features. This experiment points out that a general feature evaluation of the correctly gender-inferred images might not be enough to identify how each feature contributes to gender inference accuracy.

3.4.6 Interpreting Gender Inference

With the intent to interpret feature values of different quality images and their relation to gender inference, we grouped all models' predictions, for all datasets, and analyze their correlation with gender inference, as described in Section 3.3.3.

By analyzing the bar plot for each feature mean absolute SHAP values, in Fig. 3.6, we conclude that image resolution and lower body KP confidence are the most important characteristics. Bars in blue are positively correlated values, while the red bar denotes a negatively correlated feature. In this case, high values of blurriness contribute to gender misinference. The resolution importance was expected *a priori*, given the difficulties in discriminating gender in poor resolution data. The lower body importance might be related to hip or footwear influence, which is likely to occur when the face is fully occluded,

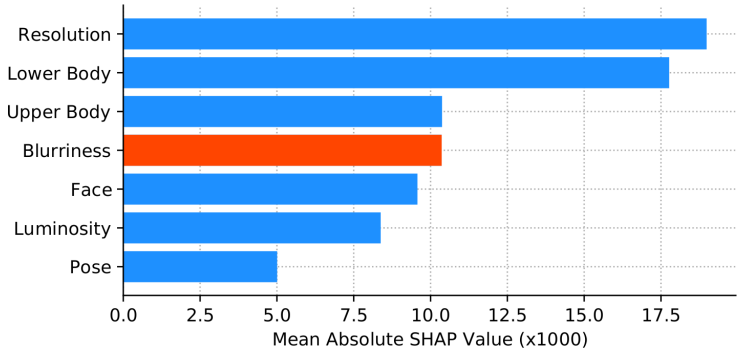
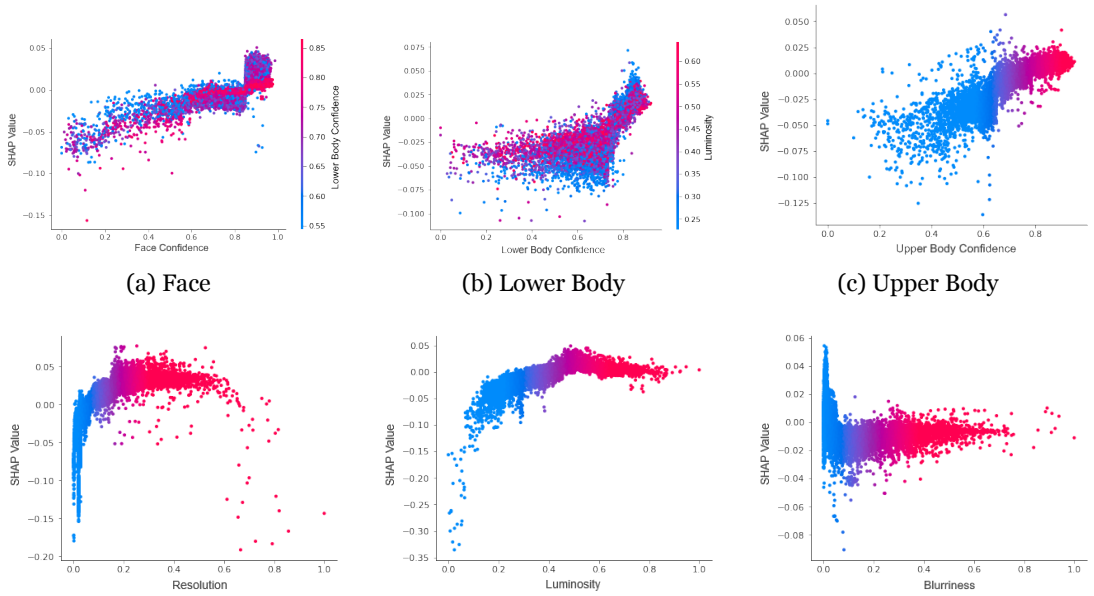


Figure 3.6: Bar plot of the mean absolute SHAP values for image and subject-based features. Blue colored bars correlate with gender correct inference, while the red one correlates with gender misinference. Features are ordered by its influence in gender inference accuracy.

forcing models to analyze different parts for gender inference. Subjects' pose is the least influential factor among those evaluated.



(a) Face (b) Lower Body (c) Upper Body

(d) Resolution (e) Luminosity (f) Blurriness

Figure 3.7: Dependence scatter plots of Face (a), Lower Body (b), Upper Body (c), Resolution (d), Luminosity (e), and Blurriness (f) SHAP values. Face and Lower Body plots present a correlation with the named feature and Lower Body and Luminosity, respectively. These plots contain a right y axis with color variation, representing correlated feature value variance. Color variation from blue to red in the other plots represent increasing values of the feature. Image-based features are normalized.

To better understand each feature's importance in gender inference, we evaluate the dependence plots for some features. Fig. 3.7 shows the results for the face, lower, and upper body KP confidence (subject-based factors) and resolution, luminosity, and blurriness dependence plots (image-based features). Negative SHAP values are linked to gender misinference, while positive ones are linked to correct inference. The face and lower body KP confidence plots correlate each corresponding feature with lower body and luminosity, respectively. For the plots with no feature correlation, the color variation from blue

to red represents an increase in feature values. We conclude that higher KP confidence values translate into gender correct inference. For lower and upper body, KP confidence values above 0.7 are linked to correct inference, while face requires KP confidence values above 0.8 for the same effect. Low values of resolution heavily negatively impact gender inference and are positively influential in higher values. Luminosity only influences gender inference when it has low values and tends to have a neutral role with value increase. Blurriness has a positive influence on low values and a negative influence when its value arises. An increase in blurriness does not translate into more incorrect gender inference.

Regarding the feature correlation plots, for face confidence values lower than 0.8, higher confidence of lower body tends to translate into higher SHAP values. This indicates that models tend to focus on the lower body portion to perform inference when the face is not visible. Regarding the lower body, increased luminosity generally improves the accuracy, which is more evident when KP confidence is between 0.4 and 0.7. This suggests that when a subject's lower body is in uncooperative poses or minor occluded, better illumination might dissipate the difficulty in interpreting such poses for gender models.

The observed results suggest that higher resolution images with good illumination, low blurriness, and cooperative poses (translated in high KP confidence values) are important factors for correct gender inference. Furthermore, the face might not be reliable for *in-the-wild* gender inference since it requires higher confidence values than those of the lower and upper body for accurate inference.

In addition to the overall analysis of features influence, we further investigate if image quality can change the importance of each image/subject-based feature. For this, we analyze feature importance in each dataset, translated in low (PETA), medium (PA-100K), and high (RAP) quality images. We present the bar plots of our analysis in Fig. 3.8. All SHAP values refer to the mean absolute values of each feature.

We can conclude that image-based features are highly influential in low quality images, given the high SHAP values they present in these conditions. This analysis also suggests that subject-based features are not reliable for gender inference in scenarios where image quality is low. If we examine image-based features across image quality variance, we conclude that they tend to have less importance as quality increases. Subject-based features increase their importance from low to medium quality images. Additionally, in this quality range, subject-based features are more important than image-based, illustrating a shift of feature importance for gender inference. In high quality images, the SHAP values are all lower than in other quality range images. This indicates that each feature loses importance in high quality data, where the models can analyze different features reliably. The pose is the least influential feature, presenting an insignificant influence in high quality images.

These findings can be applied to improve model inference. Our results suggest that, if the inference is done using low quality images, it is advised to collect new data with improved image-based features. However, assuming medium quality data usage, image quality increase (such as having better illumination or resolution) is not as important as improving subject-based features. In high quality images, gender inference accuracy improvements

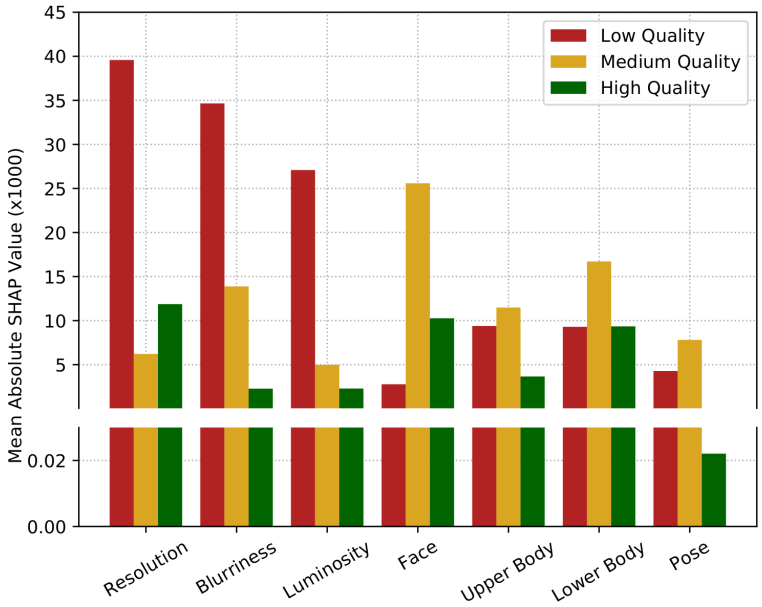


Figure 3.8: Bar plots of the evaluated features, presenting the SHAP values for low, medium, and high quality image datasets. Datasets are represented by different colors. High SHAP values represent high feature relevance for gender correct inference. The bar plot is broken in the vertical axis for better visualization.

Table 3.4: Comparison of our face-based gender model, in the Adience dataset.

Methods	**Acc (%)**
Lee *et al.* (2018) [528]	85.16
Levi and Hassner (2015) [491]	86.80
Hung *et al.* (2019) [547]	89.50
Zhang *et al.* (2017) [490]	92.43
Face-Model	**93.79**

might not be achieved through data changes, given the relatively low influence of the analyzed features for gender inference. Redefining models' architecture might be a better strategy to improve overall accuracy.

3.4.7 Face Importance in Gender Inference

The results described in section 3.4.6 suggest that face importance for gender inference might be directly related to image quality. To confirm this idea, we analyze the reliability of face-based gender inference given different data quality. We start by training our face-based gender model in Adience, one of the most challenging face gender datasets. We present the accuracy obtained for our model in Table 3.4, in comparison to results reported in the literature (5-fold cross-validation [534] values).

The obtained accuracy is better than the originally reported ones, confirming that our model is representative of a good face-based model. To assess its applicability in wild conditions, we evaluate its *cross-domain* performance using Adience, PETA, PA-100K, and RAP datasets. To keep the Adience values' comparison as fair as possible, we train and evaluate the models in PETA, PA-100K, and RAP using only face images of frontal subjects of each dataset. We report the results in Table 3.5.

Table 3.5: Gender mA of face-based gender model, trained and evaluated in within and cross-domain settings. Only frontal face images of PETA, PA-100K, and RAP datasets were considered.

Train \ Eval	Adience	PETA	PA-100K	RAP
Adience	94.83	65.56	55.55	77.85
PETA	64.96	91.16	74.60	82.43
PA-100K	75.62	72.52	84.71	79.72
RAP	75.49	68.09	68.32	95.31

Table 3.6: Gender mA of face-based gender model, trained for *Any* and *Frontal* pose images, and evaluated in frontal image faces (*Face*) and full-body images of corresponding pose (*Body*). *Any Pose* refers to all dataset images and *Frontal Pose* images refer to frontal images. `FI` stands for Face Importance.

	Any Pose			**Frontal Pose**		
Dataset	Face	Body	FI (%)	Face	Body	FI (%)
PETA	57.60	93.24	17.58	61.09	92.52	26.08
PA-100K	52.52	91.06	6.14	57.03	91.86	16.79
RAP	73.63	96.09	51.27	75.31	95.98	55.05

We can observe that the Adience-trained model performs poorly in other datasets, pointing that face-based training is not adequate for *in-the-wild* inference. However, the best results were observed for RAP, characterized as a high quality set. Similarly, RAP-trained models achieve higher accuracy performance in Adience than in PETA or PA-100K. These results indicate that Adience and RAP face dataset have evident similarities, which corroborates that face importance is closely related to increased image quality. Regarding the remaining *cross-domain* performance values, the major discrepancy was observed in the PETA/RAP combination. The performance gap of the RAP-trained model in PETA (27.22%) is superior to the observed in the mirrored case (8.73%). This suggests that the PETA dataset may be more generalizable than RAP, which may derive from facial region variability. PETA is composed of images with varying image conditions and viewpoints, from 10 different locations, in opposition to RAP, where all 26 cameras regard the same location, under similar viewpoints. Furthermore, the *cross-domain* underperformance of RAP-trained models (also seen for full-body inference, in Table 3.1) might also be linked to image quality disparity relative to the other datasets.

Next, we assess image quality and the subject pose effect on face importance. For this, we train our face-based gender model in all images (full-body) and evaluate it in face images of the correspondent dataset. This procedure is repeated when training the model only in frontal images. We present the obtained results in Table 3.6.

The results observed support that higher quality images directly influence the importance of face in gender inference. Furthermore, when we train the model with frontal data and evaluate using only facial images, the face importance increased for all three datasets, suggesting that the face importance factor is also linked to its availability. This finding suggests an implicit attention towards the face, motivated solely by guaranteed face access, via training in frontal images.

3.5 Conclusions

Soft biometrics inference in wild conditions is hindered by factors such as partial occlusions, subjects' pose, and varying levels of image quality. Considering, in particular, the gender trait, this work analyzes the image and subject-based features that are the most/least linked to correct inference of state-of-the-art models. In our analysis, we assess the combined gender inference of five state-of-the-art PAR models in well known PAR datasets. We evaluate the effectiveness of face-based gender inference models in surveillance scenarios, also considering the effect of pose. Our results suggest that image-based features are more important in low-quality images, while the importance of subject-based features is linked to an increase in image quality. Moreover, the relevance of the facial region to correctly infer gender is directly related to high-quality data. Finally, we show an implicit face attention approach, linked to its guaranteed availability, via frontal image training.

Our findings provide the basis for designing more robust gender inference models, particularly suited to work *in-the-wild*, where image and subject-based features highly vary. The observed improvements in face-based gender inference via an implicit face attention mechanism paves the way to incorporate this kind of approach in future works.

Chapter 4

YinYang-Net: Complementing Face and Body Information for Wild Gender Recognition

Soft biometrics inference in surveillance scenarios is a topic of interest for various applications, particularly in security-related areas. However, soft biometric analysis is not extensively reported in *wild conditions*. In particular, previous works on gender recognition report their results in face datasets, with relatively good image quality and frontal poses. Given the uncertainty of the availability of the facial region in wild conditions, we consider that these methods are not adequate for surveillance settings. To overcome these limitations, we: 1) present frontal and *wild* face versions of three well-known surveillance datasets; and 2) propose YY-Net, a model that effectively and dynamically complements facial and body information, which makes it suitable for gender recognition in wild conditions. The frontal and *wild* face datasets derive from widely used PAR sets (PETA, PA-100K, and RAP), using a pose-based approach to filter the frontal samples and facial regions. This approach retrieves the facial region of images with varying image/subject conditions, where the state-of-the-art face detectors often fail. YY-Net combines facial and body information through a learnable fusion matrix and a channel-attention sub-network, focusing on the most influential body parts according to the specific image/subject features. We compare it with five PAR methods, consistently obtaining state-of-the-art results on gender recognition, and reducing the prediction errors by up to 24% in frontal samples. The announced PAR datasets versions and YY-Net serve as the basis for *wild* soft biometrics classification and are available in here. The work presented in this chapter has been published in [548].

4.1 Introduction

Soft biometric cues are particularly useful in security-related issues and have increased in popularity in recent years. This is mainly motivated by their applicability in person identification, while requiring no cooperation from the subjects and being robust to low quality data [519, 520]. Furthermore, the availability of massive amounts of data acquired in uncontrolled conditions, provided by surveillance cameras and hand-held devices, has also contributed to enhancing the interest in this topic.

Gender is a particularly relevant soft label and its classification has been extensively reported in the literature, with various datasets announced [529, 533, 511, 510, 534]. However, we consider that gender recognition in *wild conditions* is not yet extensively explored, with most of the previous methods frequently linking it to face analysis in relatively good quality data, without the challenges found in surveillance scenarios. Alternatively, gender information can be retrieved in video surveillance settings, using PAR datasets,

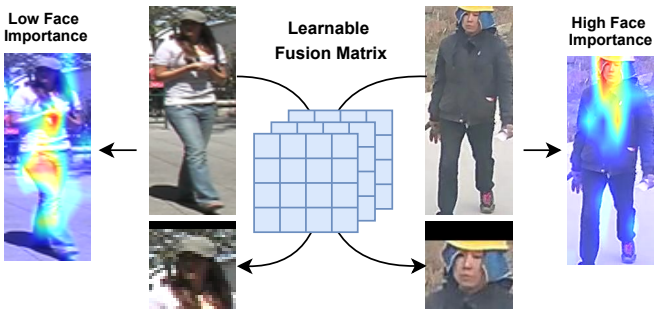


Figure 4.1: Illustration of the effect of the learnable fusion matrix, a major component of YY-Net, as an attention mechanism with respect to the (non)availability of facial information.

with methods analyzing up to 51 attributes simultaneously and not the gender alone. This promotes the use of weighted losses for the different attributes and is an evident obstacle towards the optimization of results for each one.

Given that the face is intuitively a discriminative attribute to classify gender, our main goal is to develop an integrated framework that uses facial information, whenever possible, in surveillance conditions. This is useful for scenarios where analysis is carried out mostly from frontal poses, such as airports, casinos, and border services. However, reliably retrieving the facial region from surveillance images can not be considered a solved problem. In this context, we start by creating *frontal* versions of PAR datasets (PETA, PA-100K, and RAP) using pose data from Alphapose [521, 522, 523]. Then, we create *wild* face datasets using a pose-based approach to define the head ROI, from frontal PAR images. This approach is applicable in contexts with varying image quality and uncooperative subjects, outperforming state-of-the-art face detectors. Based on the varying image/subject qualities and face variability factors, the created *wild* face datasets contrast with the state-of-the-art, showing its applicability for wild soft biometrics classification.

To effectively classify gender, using the created datasets, we propose YY-Net, a model that combines face and body-based features through a learnable fusion matrix, promoting weighted conjunction of both types of information. Fig. 4.1 displays the effect of that matrix, with respect to the availability of face information. Moreover, we apply a residual link with body features, using a channel attention sub-network, to promote a weighted face-body combination. This contributes to better focus on the most influential body parts for gender recognition.

According to our experiments, the proposed model improves gender recognition results in all frontal PAR datasets, when compared to five state-of-the-art PAR methods. Furthermore, in non-frontal images, where the facial region is unavailable, YY-Net uses body information and not the weighted combination of face-body features. In this context, it also outperforms the baseline PAR methods, in all datasets.

In summary, the major contributions of this paper are:

- We present frontal and *wild* face versions of well-known PAR datasets (PETA, PA-100K, and RAP), released as meta-data, yielding from a pose processing approach. Given their face variability factors, these *wild* face datasets contrast with the state-of-the-art, representing more challenging sets to soft biometrics applications;

- We propose YY-Net, an attention-based model for *wild* gender recognition, complementing face and body information, that surpasses the state-of-the-art in PAR datasets, particularly in frontal samples.

The remainder of this paper is organized as follows: Section 4.2 summarizes the most relevant gender recognition works; Section 4.3 describes the proposed model; Section 4.4 describes the proposed datasets, and Section 4.5 discusses the results obtained. The main conclusions and future work are presented in Section 4.6.

4.2 Related Work

4.2.1 Face-Based Gender Recognition

The face is one of the most discriminative aspects to classify gender, which culminated in several datasets designed for this task [529, 533, 511, 510]. Given the nature of gender and age estimation, simple CNN models (with only five layers) can be used for gender recognition [491]. However, more complex architectures have recently been used, such as those based on RoR [490], which have outperformed other CNN architectures.
Some works have focused on complementing the face images with facial attributes analysis. Han *et al.* [549] proposed a Deep Multi-Task Learning (DMTL) framework for joint estimation of face attributes, considering their correlation and heterogeneity for various demographic attributes (gender, age, race). Ranjan *et al.* [525] developed HyperFace, a multi-task framework based on Alexnet [526] for face detection, fused with a CNN for face landmarks detection and gender recognition. Liu *et al.* [529] combined a SVM, for face attribute classification, with a CNN for appropriately resized face localization and another for distinguishing face identities.
The complement of particular face attributes for gender classification is also a topic of research. Datcheva and Brémond [550] analyzed smile-face dynamics in gender estimation, relating its importance with the subject's age and demonstrating the complement of dynamic and appearance-based features. Lapuschkin *et al.* [551] examined which facial attributes actually impact age and gender prediction, using a Layer-wise Relevance Propagation. In the same topic, Yaman *et al.* [552] used ear appearance to aid gender estimation in profile face images, employing domain adaptation. Ryu *et al.* presented InclusiveFaceNet [530], proposing the inclusion of race and gender for face attribute detection.

4.2.2 Body-Based Gender Recognition

Regarding wild gender recognition, where the focus is on the whole body and not on the facial region alone, we present works using PAR datasets [506, 507, 508], a context that more closely relates to surveillance scenarios. Some works opt to analyze the relation of different body parts: Wang *et al.* [469] proposed a model based on a RNN encoder-decoder framework, using LSTM as a recurrent neuron, with joint recurrent learning via

images division into six horizontal strips; Zhao *et al.* [470] also used a similar partition strategy, combining LSTM and BN-Inception [471] networks.

The importance of various attributes is also reported in the literature, namely on the relation of attributes towards pedestrian attribute recognition. Learning the relationship, dependency, and correlation between attributes and using this information to accurately predict attributes have been the focus of some works [475, 476]. Similarly, using a hierarchical feature embedding framework [477] has proven to promote finer-grained clustering, translating into increased attribute recognition accuracy. However, enhancing the localization of attribute-specific areas, typically adopted by state-of-the-art methods, may not necessarily improve performance [478]. Image and subject-based features are also aspects to consider for accuracy improvement, given that their importance in gender classification varies with image quality and face availability [480].

Pose information and body part localization are also influential for pedestrian attribute recognition. Li *et al.* [472] used pose data to aid in attribute body part localization, fusing features at multiple levels for attribute recognition. Zhang *et al.* [473] proposed a Deep Template Matching method to capture body parts features, complemented by pose keypoints to guide discriminative cues learning. Liu *et al.* [474] used body ROI and assigned attribute-specific weights based on extracted ROI proposals and attribute localization. Yaghoubi *et al.* [553] used body keypoint estimation to define attribute region of interest, while also evaluating pose effect. Finally, Barra *et al.* [554] explored the usage of body pose information for gender classification in setups with subjects and camera in movement.

4.2.3 Attention-Based Gender Recognition

Various works base their approaches on attention mechanisms. Effective attention maps at different scales [481], analyzing relevant image patches [482], or focusing on different feature levels [483] are all reported methodologies to discover discriminative regions for attribute classification. Alternatively, using an attention consistency loss [484] based on heatmaps difference of original and flipped images can yield good performances.

Attribute relation was also the basis of attention approaches [555, 556, 557, 558, 559]. Li *et al.* [555] used Convolutional LSTM (ConvLSTM) for spatial and semantic correlation between attributes, combining it with channel attention to adjust the weight of relevant channel features adaptively. Zhao *et al.* [557] proposed the use of a Recurrent Attention Model (also based on ConvLSTM), mining the correlations among different spatially related attributes. Ji *et al.* [558] announced an encoder-decoder with multiple LSTM, aiming to exploit more contextual knowledge, mining deeper relations between images and attributes. Gao *et al.* [559] proposed a two-stage training, complementing semantic segmentation with attribute recognition, relating attributes in an attention-based manner between both stages.

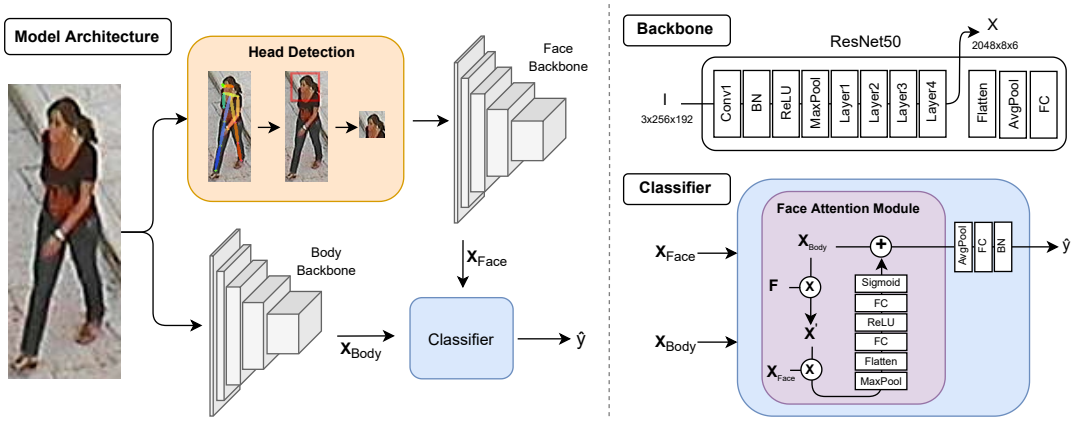


Figure 4.2: Cohesive perspective of the YY-Net, divided into three parts: *1) Model Architecture*; *2) Backbone*; and *3) Classifier*. The *Model Architecture* displays our processing flow, where the input data feeds two different backbones: 1) for the body; and 2) for facial information. The facial regions are obtained according our head detection approach, using pose information. The backbones derive from ResNet50, with classifier-related layers ignored, as shown in the *Backbone* portion. The combination of the face and body inputs is done through a classifier, which uses the proposed Face Attention Module to fuse both types of information effectively.

4.3 YinYang-Net

An overview of YY-Net is shown in Fig. 4.2. The input is processed in two ways: 1) retrieving body information; and 2) retrieving face information. We obtain facial information by first processing the image to obtain pose data and then using a head ROI detection approach, described in Section 4.4.1. We use Alphapose [521, 522, 523] to collect pose information. Both body and face information are retrieved from a modified ResNet50 [538], via feature maps at *Layer 4*. Finally, we pass both feature maps to a classifier for gender recognition.

4.3.1 Network Architecture

The key concept of our approach is to combine face and body information to classify gender. The facial region is intuitively a discriminative attribute for gender recognition, and its importance is linked to the subject pose. To achieve our goal, we use a ResNet50, pre-trained in ImageNet [546], for feature extraction of both face and body images. To better control features combination, we retrieve features from a modified ResNet50 model, with *Flatten*, *Pooling*, and *FC* layers removed. We refer the modified ResNet50 as *backbone*. We use two different backbones for face and body image processing to fine-tune extraction based on the body region inputted.

Given an input image $\boldsymbol{I}$, we denote the features extracted as $\phi_i(\boldsymbol{I}) \in \mathbb{R}^{H_i \times W_i \times C_i}$, $i \in \{\text{face}, \text{body}\}$. For a $3 \times 256 \times 192$ input image, we obtain a $H_i \times W_i \times C_i$ of $2048 \times 8 \times 6$, for both values of i. We represent our features as $\boldsymbol{X}$:

$$\boldsymbol{X}_i = \phi_i(\boldsymbol{I}), \quad i \in \{\text{face}, \text{body}\}, \quad (4.1)$$

where ϕ_i is the backbone for a given input type. Given that our framework aims to use facial information effectively, Fig. 4.2 displays the steps involved for gender recognition in frontal images. However, we are able to adapt our framework to consider any pose images through pose data processing, following our head detection approach. Representing θ as said approach, we define gender recognition of the n^{th} input as:

$$\hat{y}_n = \begin{cases} f(\rho(\boldsymbol{X}_{\text{face}_n}, \boldsymbol{X}_{\text{body}_n})), & \text{if } \mathbb{1}_{\theta(\boldsymbol{I}_n)=\text{Frontal}} \\ f(\boldsymbol{X}_{\text{body}_n}), & \text{otherwise,} \end{cases} \quad (4.2)$$

where f is a sequence of *Average Pooling*, *FC*, and *Batch Normalization* layers, ρ is our Face Attention Module (FAM), $\mathbb{1}$ is the indicator function, and $\boldsymbol{X}$ are the features extracted from a given body part.

4.3.2 Face Attention Module

Having the body and face features extracted, we combine them efficiently by promoting the focus of YY-Net in the most relevant regions of the body. We design a weighted face-body feature combination by using a learnable weights matrix: $\mathbf{F} \in \mathbb{R}^{H_i \times W_i \times C_i}$. We denote the features combination $\boldsymbol{X}_c$ as:

$$\boldsymbol{X}_c = \boldsymbol{X}_{\text{body}} \odot \boldsymbol{F} \odot \boldsymbol{X}_{\text{face}}, \quad (4.3)$$

where $\odot$ represents the Hadamard product, $\boldsymbol{F}$ is our fusion matrix and $\boldsymbol{X}$ are the extracted features. The importance of each body part is expected to vary depending on the image quality, person pose, and partial occlusions. For this reason, we add a channel-attention sub-network, based on the Squeeze-and-Excitation block [2], to modulate inter-channel dependencies. The combined input $\boldsymbol{X}_c$ passes through a series of linear and nonlinear layers, as shown in the *Classifier* part of Fig. 4.2, yielding a weight vector for the feature importance across channels. Then, we induce a residual link of this vector with body features to promote information complementary. We express the FAM output as:

$$\rho = \psi(\boldsymbol{X}_c) + \boldsymbol{X}_{\text{body}}, \quad (4.4)$$

where ψ is the sequence of linear and nonlinear layers previously described, $\boldsymbol{X}_c$ refers to the result of (4.3), and $\boldsymbol{X}_{\text{body}}$ corresponds to the extracted body features.

4.3.3 Implementation Details

Images are resized to 256×192 with random horizontal mirroring as inputs. The SGD algorithm is used for training, with momentum of 0.9 and weight decay of 0.0005. The initial learning rate is set to 0.01 for the backbones and classifier. Plateau learning rate scheduler is used with a reduction factor of 0.1 and a patience epoch number of 4. Batch

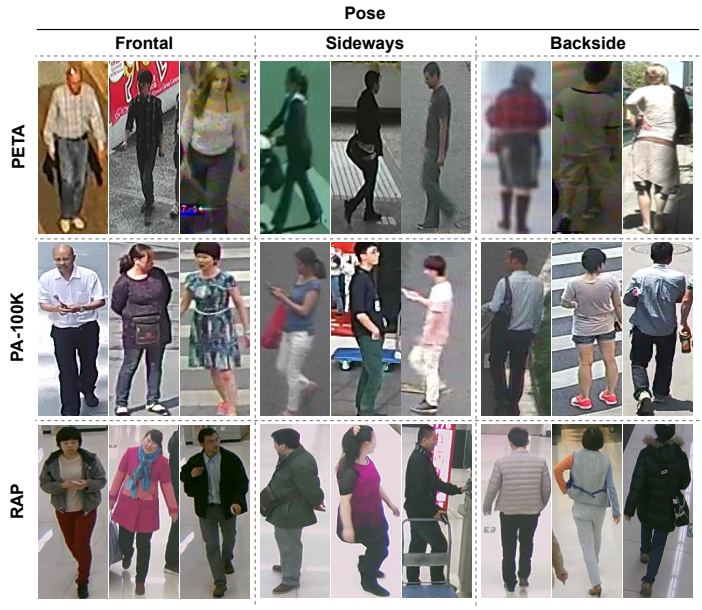


Figure 4.3: Examples of images considered as *frontal*, *sideways* and *backside*. Each row is related to one of the three PAR datasets and each column regards examples of the respective pose.

size is set to 32 and the number of training epochs is set to 30, using BCELogits as loss function.

4.4 Frontal PAR and Face Datasets

We create frontal and *wild* face versions of three well-known PAR datasets. The frontal images are obtained from a division of three possible values: *frontal*, *sideways*, and *backside*. We differentiate these values based on pose information from Alphapose [521, 522, 523]: if the rightmost shoulder, from the top left of the image, is the left shoulder, the subject faces forward (*frontal*); otherwise, the subject is facing backward (*backside*). Regarding *sideways* classification, we attribute this label to subjects with a shoulder length, with respect to the upper body height, less than 0.5. Examples of the pose classes considered are presented in Fig. 4.3.

4.4.1 Head Detection

Head ROI detection also uses pose information, cropping head regions from frontal images. Initially, we obtain the coordinates x and y of the right and left ears, considering the head's central point as the arithmetic mean of such coordinates. Then, head ROI are drawn using the top-left and bottom-right bounding boxes coordinates, centered in the head's central point, with the head ROI height as $\frac{2}{9}$ of the whole body silhouette height. Head bounding boxes drawing is based on Detectron [408]. To ensure the versatility of our approach, head ROI detection is applied to the original image, before resizing.
Based on the pose data, this head ROI detection approach can detect the head of multiple people in an image. However, although various people may appear in PAR dataset images,



Figure 4.4: Comparison of the head ROI detection results between MTCNN, EXTD, RetinaFace, and our approach, in images with varying quality, head pose and blurriness. Green boxes denote cases where the method accurately locates the face, and red boxes denote the opposite. Dotted red boxes refer to cases where the model over-predicted the number of existent faces. All images are resized to the same resolution for visualization purposes.

pedestrian attributes are typically related to the main/central cropped person. Therefore, given this setting, this approach only detects the head of the main person by considering, for each image, the pose data outputted from Alphapose with the highest confidence. This allows it to retrieve head regions from frontal images with varying quality and/or partial occlusions. Other face detectors, such as MTCNN [560], EXTD [561], and RetinaFace [562], fail to detect head region as reliably as the used approach in the considered (wild) image conditions. To show the versatility of this method, Fig. 4.4 presents some examples of head ROI detection in frontal PAR images. In our head detection comparison, we use the default implementation of all methods.

We observe that MTCNN only detects faces in higher quality images, failing to detect in lower quality ones. This is expected since MTCNN was not designed to target "small faces" specifically, unlike EXTD and RetinaFace. Nonetheless, these methods fail to detect faces in low quality images reliably. EXTD was able to detect faces in more images but outputs more incorrect detections as well. RetinaFace accurately detects faces in more images than MTCNN but is not able to detect faces when quality is subpar, with a background color similar to the person (second and fourth column) or with blurry faces (third and fifth column). Our approach detects the facial region in all presented scenarios, regardless of the image quality. This is an important aspect, given that we intend to provide our model with facial images from surveillance scenarios, with varying image quality. Furthermore,

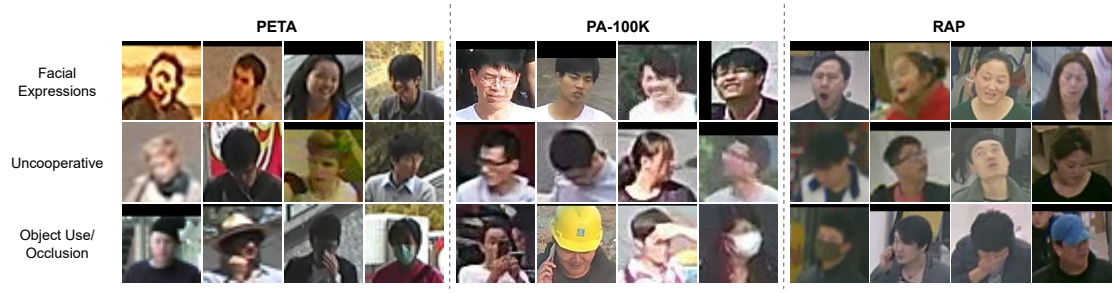


Figure 4.5: Examples of image quality variability in *wild* face datasets. The datasets contain subjects with varying facial expressions, levels of cooperativeness, and partial occlusions. Each row refers to one variability factor, and each group of four columns regards examples of a PAR dataset.

Table 4.1: Dataset image-based feature analysis. Values are normalized for the combination of all dataset values, for each feature. Datasets named with *Frontal* suffix refer to datasets composed solely of frontal images.

Dataset	Resolution	Luminosity	Blurriness
PETA	0.037 ± 0.024	0.432 ± 0.100	0.120 ± 0.148
$\text{PETA}_{\text{Frontal}}$	0.037 ± 0.023	0.437 ± 0.101	0.140 ± 0.158
PA-100K	0.062 ± 0.063	0.449 ± 0.126	0.095 ± 0.089
$\text{PA-100K}_{\text{Frontal}}$	0.061 ± 0.064	0.459 ± 0.127	0.100 ± 0.090
RAP	0.131 ± 0.083	0.407 ± 0.107	0.022 ± 0.013
$\text{RAP}_{\text{Frontal}}$	0.138 ± 0.090	0.416 ± 0.108	0.023 ± 0.013

we use the described approach to create *wild* face datasets, representing a challenging set for soft biometrics classification. The variability encountered in these datasets is displayed in Fig. 4.5.

4.4.2 Data Analysis

To prove that frontal datasets preserve the main characteristics of the original one, we analyze the major image-based factors that influence gender recognition. Note that train/test sets follow the set divisions of each PAR dataset (*i.e.*, we did not alter the train and test set division). The frontal datasets sets are images, from the original train/test sets, that match the defined frontal criterion (defined in 4.4). We consider *resolution*, *luminosity*, and *blurriness* as the main image-based features: *resolution* is retrieved from image width and height multiplication; for *luminosity*, we use the red, green, and blue value channels to measure the perceived brightness [539]; and *blurriness* yields from the convolution of the images with a Laplacian kernel, taking the variance as result. The image-based feature analysis of the used PAR datasets, and respective frontal versions, is shown in Table 4.1.
The obtained results suggest that frontal datasets preserve the characteristics of the original version. As such, they are not easier for gender recognition, and the accuracy variance of models on these datasets (when compared to the original ones) derives from pose and face access. Furthermore, given the resolution and blurriness values, RAP can be considered the dataset with the highest quality, while PETA is the lowest quality one.
Another conducted experiment was the evaluation of image number ratio between *test* and *all* (train + test) images. Furthermore, we assess the image number ratio of frontal

Table 4.2: Number of images, test/all image number ratio for each dataset, and frontal/PAR image number ratio, for train and test set. Frontal datasets are represented with the subscript term *Frontal* and test/all ratios are represented by Dataset Ratio.

Dataset	**Train**		**Test**		**Dataset**
	Images	Ratio	Images	Ratio	**Ratio**
PETA	11400	0.379	7600	0.383	0.400
$\text{PETA}_{\text{Frontal}}$	4318		2918		0.403
PA-100K	90000	0.379	10000	0.354	0.100
$\text{PA-100K}_{\text{Frontal}}$	34128		3539		0.093
RAP	33268	0.339	8317	0.346	0.200
$\text{RAP}_{\text{Frontal}}$	11275		2880		0.203

and PAR datasets, for the test and train set, to verify if the pose division was balanced. We present the ratios analyzed in Table 4.2. The test/all images ratios are very similar between all PAR datasets and their frontal versions. Additionally, frontal/PAR ratios are similar between datasets and around $\frac{1}{3}$, which is expected given that the dataset pose division is based on three values (*frontal*, *sideways*, and *backside*).
Regarding the *wild* face datasets, all the ratios presented for frontal datasets are analogous for the face ones. The resolution, luminosity, and blurriness of face datasets are similar to the frontal dataset values, given that the facial regions were retrieved from frontal images.

4.5 Experiments

4.5.1 Datasets, Methods, and Evaluation Metrics

Our experiments are carried out in three state-of-the-art PAR datasets: PETA [506], PA-100K [507], and RAP [508]. The PETA dataset has 19,000 images, divided into three subsets: 9,500 for training, 1,900 for validation, and 7,600 for testing. The PA-100K dataset is composed of 100,000 images from outdoor surveillance cameras. It is split into 80,000 images for training, 10,000 for validation, and 10,000 for testing. The RAP dataset contains 41,585 images. Following the official protocol [508], we split the dataset into 33,268 training images and 8,317 test images. Furthermore, we carried out experiments in the proposed frontal datasets, consisting of frontal images of the described PAR datasets.
The PAR methods used in our experiments are VAC [484], APR [476], ALM [483], StrongBase [478], and DeepMar [475]. All models are evaluated using the original implementations with minor adaptations to only classify gender and not all pedestrian attributes.
To present our results, we use the label-based metric mA [508], which is the mean classification accuracy of the positive and negative samples for each attribute, averaged over all attributes:

$$\text{mA} = \frac{1}{2n}\sum_{i=1}^{m}\left(\frac{\text{tn}_i}{p_i} + \frac{\text{tp}_i}{n_i}\right), \quad (4.5)$$

where n is the number of examples, m the number of attributes, p_i is the number of pos-

Table 4.3: Gender mA of different models on frontal PAR datasets. The outperforming method for each dataset is shown in bold.

Methods	$\text{PETA}_{\text{Frontal}}$	$\text{PA-100K}_{\text{Frontal}}$	$\text{RAP}_{\text{Frontal}}$
ALM [483]	91.27	91.23	95.71
APR [476]	92.49	91.63	95.80
DeepMar [475]	91.09	90.75	95.96
VAC [484]	91.17	90.15	94.49
StrongBase [478]	92.62	92.05	96.14
YY-Net	**93.45**	**92.79**	**97.07**

Table 4.4: Gender mA of different models on PAR datasets. The outperforming method for each dataset is shown in bold.

Methods	PETA	PA-100K	RAP
ALM [483]	92.28	90.34	95.69
APR [476]	92.84	90.05	96.20
DeepMar [475]	92.33	90.39	96.40
VAC [484]	92.85	91.05	96.59
StrongBase [478]	93.13	90.77	96.74
YY-Net	**93.39**	**91.20**	**96.86**

itive examples and tp_i is the number of correctly predicted positive examples for the i^{th} attribute; tn_i and n_i are defined analogously. We adapt this metric to only consider the gender attribute, from the PAR datasets.

4.5.2 Gender Recognition

Given that our focus is to improve wild gender recognition, combining facial and body information, we start by assessing the performance of YY-Net in frontal PAR datasets. We compare it with five state-of-the-art PAR methods, presenting our results in Table 4.3.

Our approach is better than all considered methods, which corroborates the importance of effectively combining face and body images in gender classification. When comparing with the best result for each dataset, we improve gender accuracy by 0.83%, 0.74%, and 0.93% in PETA, PA-100K, and RAP, respectively. The improvement in all datasets demonstrates the versatility of YY-Net, corresponding to an error reduction of 11.25%, 9.31%, and 24.09%, respectively. The increased error reduction displayed in RAP could be linked to its higher image quality, relative to the other PAR datasets. High quality images might benefit more from effective facial and body information conjunction, giving the increased body details, culminating in improved relevant body portion focus.

With the intent to analyze the adaptability of our model to deal with unconstrained pose scenarios, we evaluate its performance in PAR datasets. We present, in Table 4.4, the accuracy of the analyzed state-of-the-art PAR methods, for each dataset.

Our results show that the combination of face and body information, even in the context of unconstrained poses, contributes to slight improvements. This effect was observed for all three datasets. Compared to the second best method, YY-Net improved by 0.26%, 0.15%, and 0.12% in PETA, PA-100K, and RAP, respectively, which corresponds to an

Table 4.5: Gender mA of YY-Net in frontal datasets, when trained in PAR datasets and in its frontal version. Frontal datasets are represented with the subscript term *Frontal*. The highest evaluation accuracy for each dataset is shown in bold.

Train \ Test	$PETA_{Frontal}$	$PA\text{-}100K_{Frontal}$	$RAP_{Frontal}$
PETA	**93.88**	78.67	81.47
$PETA_{Frontal}$	93.45	77.38	78.09
PA-100K	79.02	92.70	89.51
$PA\text{-}100K_{Frontal}$	77.98	**92.79**	89.10
RAP	73.06	75.16	96.95
$RAP_{Frontal}$	73.51	77.94	**97.07**

error reduction of 3.79%, 1.68%, and 3.68%. The results suggest that YY-Net is reliable in frontal pose images and applicable in wilder contexts, with varying poses.

4.5.3 Within and Cross-Domain Performance

To assess if the performance improvements in PAR datasets derive essentially from more accurate results for frontal images, we evaluate YY-Net in frontal datasets, using the weights from PAR training and compare it to YY-Net trained solely on frontal images. We report our results in Table 4.5, exploring *within* and *cross-domain* settings.

This experiment shows that models trained on different datasets have distinct performances, linked to dataset length and quality differences. If we analyze the *within-domain*, only PETA has a significant difference between PAR and frontal training models, with the PAR model obtaining better performance. Models trained in PAR datasets do not interact with more frontal images than those trained in frontal datasets. However, they are trained in more images and with more varying body poses. This contributes to better body information processing which, in conjunction with our face combination approach, translates into a more accurate gender recognition. In this scenario, PETA training is more beneficial than only $PETA_{Frontal}$, given that it is the dataset with the lowest number of training images. Regarding the RAP and PA-100K, their frontal versions contain enough training images that PAR training does not translate into increased *within-domain* performance.

When we analyze the *cross-domain* performance, we observe that PAR training is more beneficial to PETA and PA-100K, with RAP having better performance with frontal training only. These differences could be associated with image quality, where RAP is the highest quality dataset, given our results in Table 4.1. This is particularly relevant, given that high-quality images have increased face importance, particularly in frontal images. As such, training solely in frontal images may promote a more effective face data processing, translating into a better face-body combination. In this case, training for various poses might dissipate this aspect leading to a poorer generalization, expressed by worse *cross-domain* performance in RAP. For PETA and PA-100K, more training data (PAR) promotes better generalization, as the *cross-domain* results demonstrate.

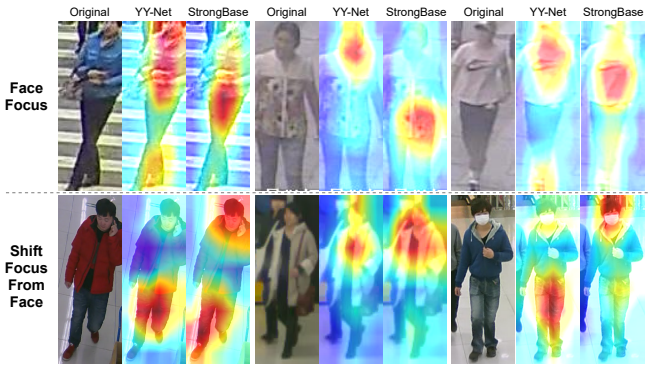


Figure 4.6: Comparison of StrongBase and YY-Net attention focus from images where YY-Net correctly classifies gender and StrongBase does not. Each row is related to one of the categories of diverging focus.

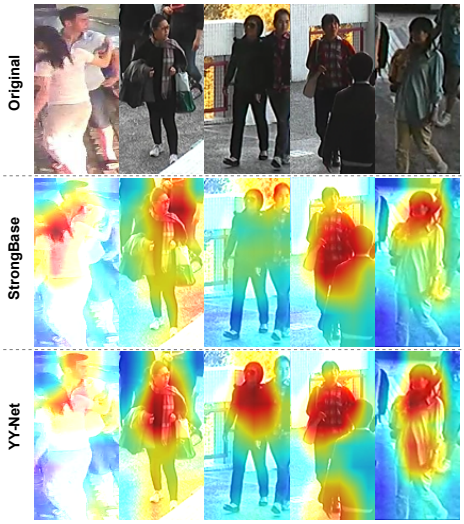


Figure 4.7: Crowd effect on attention focus from StrongBase and YY-Net. Each column displays an example where YY-Net accurately classifies gender and StrongBase misclassifies it.

4.5.4 Face Attention Influence

Given that YY-Net uses face information to aid in gender recognition, we assess whether this approach would, implicitly, focus more on faces to perform its task. To achieve our goal, we compare CAM [563] from our model with the StrongBase's ones, from images where StrongBase exclusively misclassifies. This is the chosen method since it outperformed the remaining ones. We group all three datasets and present the CAM of representative images in Fig. 4.6.

The displayed examples suggest that our FAM does not force a predominant face focus but does contribute to a more coherent face-body combination. In the first row (*Face Focus*), we observe a hip/upper body focus from StrongBase. In these cases, the facial region is a more discriminative factor (relative to the hip/upper body region), which is the focus of YY-Net. However, there are instances where the facial region may difficult/misguide gender recognition. In the second row, we observe cases where the facial region is not easily disclosed or presents unisex qualities. In these cases, analyzing other body portions may be beneficial, which is the approach taken by our model. StrongBase opts to focus more on the neck/facial region, contributing to gender misclassification. Another scenario where

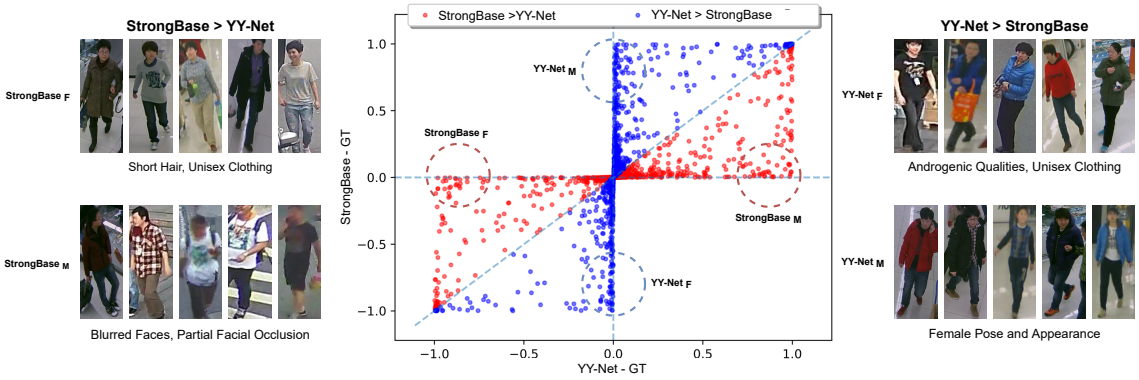


Figure 4.8: Scatter plot of the difference between YY-Net and StrongBase predictions, and GT labels, in RAP and PA-100K images. Points in blue represent cases where YY-Net outperforms StrongBase, and points in red denote the opposite. Four different regions are highlighted in the plot, displaying the cases where a model outperforms the other the most, for male (M) and female (F) recognition. Representative images, from the four regions, are presented sideways of the plot, accompanied by descriptive labels.

YY-Net has better performance is in crowded situations, where it is able to focus on one subject, while StrongBase underperforms in the same context. We display images with multiple subjects where our model accurately classifies gender and StrongBase does not, in Fig. 4.7.

Influenced by FAM, YY-Net is capable of focusing on the main subject in the image. By contrast, StrongBase has performance issues when other people appear in the front/ background, dispersing its attention to other subjects in the image. This translates into a more scatter focus on the main subject, contributing to incorrect gender classification. In the image of the rightmost column, YY-Net complements facial region information with hip one, contrasting with the StrongBase, where the facial region was the main (sole) focus. The approach of YY-Net is justifiable given the background "noise" of other people, namely an arm near the facial region, and the androgenic qualities of the subject.

The analysis of Figs. 4.6 and 4.7 focuses on the model attention reasoning regarding its face/body balance importance and crowded scenarios, respectively. Note, however, that said reasonings might not be extensible to all dataset images. To further compare YY-Net and StrongBase gender classification performance, we process the output of each model and assess the difference between each prediction and the ground truth label. We perform this comparison in a set of RAP and PA-100K images, where label 0 refers to male and 1 to female, and create a scatter plot, displayed in Fig. 4.8. We focus our analysis on four different regions, representing cases where a model outperformed the other by the largest margins. Furthermore, we display representative images of each region, which will be the basis of our analysis.

In the cases where YY-Net outperforms StrongBase, our model efficiently distinguishes females and males with androgenic/unisex appearance. For both genders, the pose, clothing, and face appearance misguides StrongBase in its classification, with YY-Net being more resilient. If we examine the cases where StrongBase was the outperforming model, we observe that shorter hair females, with unisex clothing, are situations where the performance of our model is subpar. Additionally, males with faces blurred, partial occluded,

Table 4.6: Performance comparison, on $PETA_{Frontal}$ dataset, when gradually adding each component to the baseline model. Variants of the same key parts lie in the same group. Components in bold are the adopted approach in YY-Net.

Component	mA
Baseline (Face Input)	91.13
Baseline (Body Input)	92.49
Face-Body Combination (FBC)	92.65
Weighted FBC (WFBC)	93.10
WFBC + Body	93.26
Attention + WFBC	93.31
Attention + WFBC + Body (YY-Net)	**93.45**

or with slightly uncooperative (misguided) poses might represent a challenge for YY-Net. This suggests that, although our model can handle subjects with androgenic/unisex qualities, these are also troublesome scenarios, particularly for females. Based on the typical appearance of females in PAR datasets (longer hair without unisex clothing), YY-Net might have prioritized the facial region over body information, thus contributing to gender misclassification. For males, and females to a lesser extent, face blurriness might be a major contributing factor for misclassification. Given our subjacent approach of combining face with body information to classify gender, this could justify the underperformance in the stated conditions.

4.5.5 Ablation Studies

To validate the sensitivity of the obtained results with respect to the major components of YY-Net, we evaluate the influence of each one on gender recognition accuracy. As *Baseline*, a ResNet50 network with the modifications described in Section 4.3.1 and a classifier without the FAM are used. Furthermore, we analyze the influence of increment/variance of the components used for two key parts: 1) combination of face and body information; and 2) focus in body part attention mechanisms. We present our results in Table 4.6.

To obtain an initial baseline, we start by evaluating the importance of YY-Net components by assessing gender recognition using only face or full-body images. The results obtained by both baselines denote that full-body images (which also contain the facial region) promote a higher gender recognition accuracy than facial images. Nevertheless, neither facial nor body information by themselves are enough to obtain YY-Net performance.

We assess the effect of combining face and body information through the Hadamard product, denoted as *Face-Body Combination* (FBC) in Table 4.6. This approach translates into higher mA than the baseline with body input, corroborating the idea that both inputs are valuable for gender recognition in frontal images. However, this was only a slight improvement, suggesting that the combination of the input is inefficient. Adding the learnable fusion matrix F to influence facial and body information combination, denominated as *Weighted FBC* in the Table, promotes a mA increase.

The evaluation of the attention approach (*Attention*) and the preservation of complementary information (*Weighted FBC + Body*) is also a focus of our experiments. *Attention* refers to the sequence of linear and nonlinear layers (described in Section 4.3.2) to modu-

late inter-channel dependencies. The residual approach of complementing the combined face-body information with body input contributes to higher gender recognition accuracy. This notion is linked to contextualizing face-body information, while using all the available information (full-body). In practice, this strategy intends to deviate the model from heavy facial region focus, regardless of the subject appearance. The attention method, along with the weighted face-body combination, provide an efficient focus on the critical parts of the body to classify gender, obtaining a similar performance to only using complementary information. Finally, *YY-Net* combines the weighted FBC with attention, while also using complementary body information, which translates into an aggregated improvement when compared with all other evaluated components. This approach promotes a balanced equilibrium between face and body information, while focusing on the most influential body portions for gender recognition.

4.6 Conclusions

Subjects pose is known to hinder gender recognition in wild conditions, influencing the focus of models on different subject features. In this context, frontal images typically bias the focus towards the facial region, while the body silhouette is the main focus in non-frontal data. Given that the state-of-the-art usually evaluates gender in good quality face datasets, we present frontal and *wild* face versions of well-known PAR datasets, created through pose data processing. These subsets supply the learning data to our proposed model, YY-Net, designed to effectively complement face and body information, making it suitable for gender classification *in-the-wild*. In our proposal, the key is to use a learnable fusion matrix and channel-attention sub-network, which enables a dynamic focus on the most relevant body regions relative to specific image/subject features, achieving state-of-the-art results in PAR datasets.

The announced versions of the *wild* face datasets differ from the existing ones, presenting a more challenging set to classify soft biometrics. Furthermore, the robustness displayed by YY-Net to classify gender in wild conditions supports its usability as a basis for subsequent developments in soft biometrics analysis. Extending our approach to be applied in the context of pedestrian attribute classification in video settings [564] or classifying attributes of multiple people in a given scene are potential improvements for future work. Finally, given the limitations of YY-Net for specific group cases, the implementation of an ensemble model based on holistic gender analysis (StrongBase) and face-body balance (YY-Net) is also a direction for future work.

Chapter 5

WASD: A Wilder Active Speaker Detection Dataset

Current ASD models achieve good results on cooperative settings with reliable face access using only sound and facial features, which is not suited for less constrained conditions. To demonstrate this limitation of current datasets, we propose a Wilder Active Speaker Detection (WASD) dataset, with increased difficulty by targeting the key components of current ASD: audio and face. Grouped into 5 categories, WASD contains incremental challenges for ASD with tactical impairment of audio and face data, and provides a new source for ASD via subject body annotations. To highlight the new challenges of WASD, we divide it into Easy (cooperative settings) and Hard (audio and/or face are specifically degraded) groups, and assess state-of-the-art models performance in WASD and in the most challenging available ASD dataset: AVA-ActiveSpeaker. The results show that: 1) AVA-ActiveSpeaker prepares models for cooperative settings but not wilder ones (surveillance); and 2) current ASD approaches can not reliably perform in wilder settings, even if trained with challenging data. To prove the importance of body for wild ASD, we propose a baseline that complements body with face and audio information that surpass state-of-the-art models in WASD and Columbia. All contributions are available at `https://github.com/Tiago-Roxo/WASD`. The work presented in this chapter has been published in [3].

5.1 Introduction

Active Speaker Detection (ASD) aims to identify, from a set of potential candidates, active speakers on a given visual scene [59]. Currently, this assessment is done at the video frame level using facial cues and sound information. Despite its application in several topics such as speaker diarization [565, 566, 567], human-robot interaction, or speaker tracking [568, 569], its applicability in wild conditions is still an open issue.
The state-of-the-art dataset for ASD is AVA-ActiveSpeaker [59], composed of several Hollywood movies, with diversity in languages, recording conditions, and speaker demographics, totalling in 38 hours and over 3 million face images. Although AVA-ActiveSpeaker has some challenging aspects, it still is not a perfect representation of *in-the-wild* data [59], since it assesses ASD in movies, a setup with controlled (scripted) action and speaking, with adequate audio and image quality. This motivates state-of-the-art models to identify active speakers solely based on audio and face data, disregarding other informations such as speaking context or body expressions. This is particularly problematic since ASD in wild conditions can not assume face availability, subject cooperation, and good audio quality, as shown in Figure 5.1. To overcome these limitations, we propose a Wilder Active

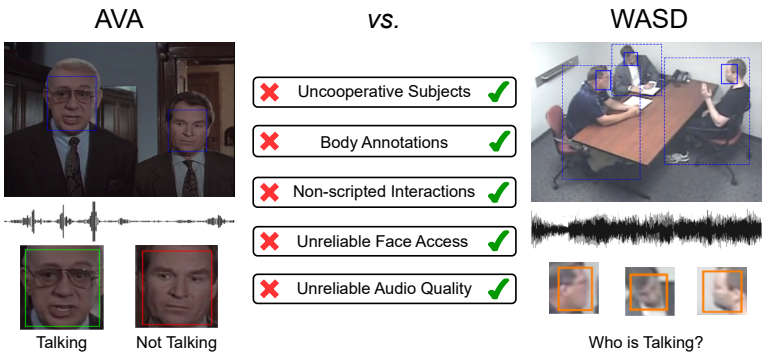


Figure 5.1: AVA-ActiveSpeaker (AVA) state-of-the-art models achieve over 94% mAP in active speaker detection, solely based on **face** and **audio data**. However, this approach may not be suited for uncooperative poses, non-guaranteed face access, or unreliable image/audio quality. How well do these models perform in such scenarios? And can body information aid in this task? To answer these questions (and more), we propose WASD, a Wilder Active Speaker Detection dataset.

Table 5.1: Feature comparison of ASD datasets. AVA-ActiveSpeaker is represented as AVA. If datasets contain information regarding a feature, its absence is presented with ×, while its presence with ✓. WASD has a high number of hours, with increased number of faces and reduced face tracks (culminating in higher average video duration), FPS variability (Var), and increased talking percentage. The most discriminative factors are demographic representation (DR), surveillance conditions (SC), and body data annotations.

Dataset	Total Hours	Faces (M)	Face Tracks (m)	Video Duration (s)	FPS Var	Talking %	DR	SC	Body Data
Columbia [570]	1.5	0.2	-	-	×	-	×	×	✓
Talkies [72]	4.2	0.8	23.5	1.5	-	-	-	×	×
EasyCom [571]	6.0	-	-	-	×	-	×	×	×
ASW [60]	30.9	-	11.5	∼10	-	57.9	-	×	×
AVA [59]	37.9	3.7	38.5	≤10	✓	24.2	-	×	×
WASD	30.0	7.4	9.8	∼28	✓	84.6	✓	✓	✓

Speaker Detection (WASD) Dataset.

WASD aims to preserve the challenging characteristics of AVA-ActiveSpeaker while increasing the difficulty of ASD by targeting the two key components state-of-the-art models use: face and audio. We select videos from YouTube and group them into 5 categories, based on a set of features targeted at face and audio impairment. The categories range from optimal conditions (face availability and good audio quality), to surveillance settings (non-guaranteed face access, subject cooperation, or sound quality). The increasing scale of ASD challenges can be useful for: 1) assess the ability of current models to deal with wild conditions and specific aspect impairment (audio, face, or a combination of both); 2) evaluate the limitations of AVA-ActiveSpeaker to prepare models for wild conditions; and 3) show the limitations of face and audio dependency for wild ASD, easing the identification of model improvements towards this goal. By selecting YouTube videos from real interactions, WASD also contains expressions, sudden interruptions, and interactions that movies hardly contain. These additional challenges, enhanced by the variability of demographics in WASD, contribute to a challenging ASD dataset where state-of-the-art models can not easily perform. Furthermore, WASD provides body data annotations and a body-based approach to motivate the development of models using body information

to complement face and audio data in (wild) ASD. Finally, the importance of WASD can be extended to other tasks, aside ASD, namely: speech recognition (what is the person saying), speech diarization (segment the audio per speaker), specific action recognition (dancing, talking, raising hand, social interactions), among others. To summarize, the main contributions are:

- We propose WASD, a ASD dataset divided into 5 categories with incremental ASD challenges, targeting audio quality and face availability, ranging from optimal conditions to surveillance settings. WASD is also innovative by providing ASD body annotations for various challenging conditions to motivate the development of body-based approaches;
- We show the limitations of AVA-ActiveSpeaker to prepare state-of-the-art ASD approaches for wilder conditions, and show that current ASD models can not reliably perform in such settings, even if trained in WASD data, in particular for audio impairment, facial occlusion, and surveillance settings;
- To prove the importance of body information for wild ASD, we propose a baseline that complements body with face and audio data, surpassing all state-of-the-art models, in particular for uncooperative and challenging sets (surveillance conditions) of WASD, and outperforms face-based implementations in Columbia.

5.2 Related Work

Active Speaker Detection. Works on ASD have evolved from facial visual cues [572, 573, 574] to audio as primary source [575, 576], to multi-modal data combination [59, 69, 1, 72, 70]. Since the introduction of AVA-ActiveSpeaker [59], combining audio with facial features is the *de facto* way to predict active speakers. Large 3D architectures [61], hybrid 2D-3D models [62], and large-scale pretraining [63, 64] for audio-visual combination are amongst some of the following works. Despite the viability of these approaches, feature embedding improvement [65] or attention approaches [66, 67, 68] were necessary to improve ASD. Creating two-step models, where the first focuses on short-term analysis (audio with face combination) and the second on multi-speaker analysis, is the approach from various recent works [69, 70, 71, 72]. ASC [69] focused on long-term multi-speaker analysis via temporal refinement, ASDNet [70] used a similar approach for inter-speaker relations, with improved visual backbones, and UniCon [71] relied on audio-visual relational contexts with various backbones. Improving speaker relation representation via GCN [73] is also a viable approach to assess context information [72, 74]. Diverging from two-step training, end-to-end models have also emerged for ASD [1, 75, 74, 76, 77]. TalkNet [1] focused on improving long-term temporal context with audio-visual synchronization, TS-Talknet [77] considered pre-enrolled speaker embedding to complement this synchronization, EASEE [75] included GCN to complement spatial and temporal speaker relations, and Light-ASD [76] proposed a lightweight model by splitting 2D and 3D convolutions for audio-visual feature extraction, and applied BGRU for cross-modal modeling.

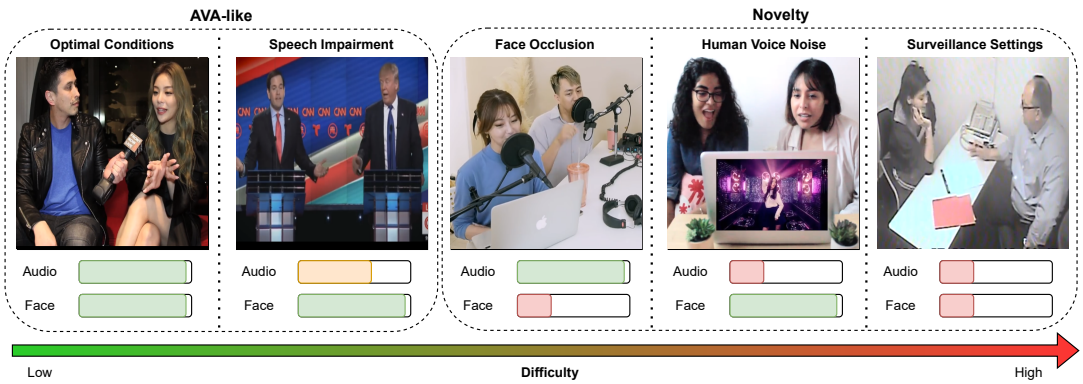


Figure 5.2: Considered categories of WASD, with relative audio and face quality represented. Categories range from low (Optimal Conditions) to high (Surveillance Settings) ASD difficulty by varying audio and face quality. Easier categories contain similar characteristics to AVA-ActiveSpeaker (AVA-like), while harder ones are the novelty of WASD.

Datasets. There is a variety of available datasets suited for ASD, such as frontal speaker data, designed for speech recognition [577, 578], voice activity detection [579], and diarization [565] datasets. However, these are limited in subject diversity and talking scenarios, diminishing their relevance. With increased talking variability, datasets derived from movies and TV shows have also been reported [580, 581, 582, 583], limited by the low number of annotated hours. Other related setups are lip reading datasets [584, 585, 586, 587, 588, 589], whose purpose diverges from ASD since their goal is to infer the words pronounced from a given speaker. Recently there is a greater focus on specific ASD datasets [570, 72, 571, 59, 60], whose task is to determine the talking speaker from a set of admissible candidates. Columbia [570] contains 87 minutes of a panel discussion, with up to 3 visible speakers. Talkies [72] focuses on low duration videos, totalling 4 hours, with an average of 2.3 speakers and off-screen speaking. Easycom [571] is designed for multiple tasks related with augmented reality, composed of various sessions of speakers sat at a table, with background noise. AVA-ActiveSpeaker [59] is the state-of-the-art dataset, with over 150 Hollywood videos, totalling almost 38 hours, with demographic diversity and dubbed dialogues. ASW [60] was proposed with over 30 hours, from 212 videos randomly selected from the VoxConverse [590], containing various sets of interviews. The proposed dataset, WASD, brings challenging sets, *in-the-wild* videos, demographic diversity, and body data annotations. The main characteristics of our dataset relative to others are presented in Table 5.1.

5.3 Dataset

We propose WASD, a dataset that aims to show the limitations of current state-of-the-art models by compiling a set of videos from real interactions with varying accessibility of the two key components for ASD: *audio* and *face*. By dividing our dataset into 5 categories with varying degrees of audio and face quality, we can assess how models adapt to these scenarios and which factors are more relevant for ASD. We create a balanced demographics dataset (regarding language, race, and gender), with several challenging factors,

Table 5.2: Category feature matrix. Feature description: FA, Face Availability; SO, Speech Overlap; DS, Delayed Speech; FO, Facial Occlusion; HVB, Human Voice as Background Noise; SS, Surveillance Settings. The absence of a certain feature is presented with ×, while its presence with ✓. Features containing ? refer to non-guarantee of its presence or absence. Green cells refer to features favorable for ASD, while red ones are unfavorable.

Category	FA	SO	DS	FO	HVB	SS
Optimal Conditions	✓	×	×	×	×	×
Speech Impairment	✓	✓	✓	×	×	×
Face Occlusion	✓	×	×	✓	×	×
Human Voice Noise	✓	×	×	×	✓	×
Surveillance Settings	?	?	?	?	?	✓

complemented with body annotations data. We discuss the process of dataset creation in the following sections.

5.3.1 Video and Category Selection

We select videos from YouTube and group them into 5 categories based on a set of features, whose values were attributed by human assessment. The main features used for category division are shown in Table 5.2, with the complete list in the supplementary materials. In sum, videos are grouped as follows:

- **Optimal Conditions**: People talking in an alternate manner, with minor interruptions, cooperative poses, and face availability;
- **Speech Impairment**: Frontal pose subjects either talking via video conference call (*Delayed Speech*) or in a heated discussion, with potential talking overlap (*Speech Overlap*), but ensuring face availability;
- **Face Occlusion**: People talking with at least one of the subjects having partial facial occlusion, while keeping good speech quality (no delayed speech and minor communication overlap);
- **Human Voice Noise**: Communication between speakers where another human voice is playing in the background, with face availability and subject cooperation ensured;
- **Surveillance Settings**: Speaker communication in scenarios of video surveillance, with varying audio and image quality, without any guarantee of face access, speech quality, or subject cooperation.

Some important aspects to consider from Table 5.2: 1) all categories, aside Surveillance Settings, guarantee face availability, which corresponds to cooperative scenarios and close-up faces; 2) we consider speech delay and overlap as variations of slight speech impairment, thus their grouping in the same category; and 3) Surveillance Settings does not have any guarantee regarding the analyzed features, corresponding to wild conditions. These considerations support the range of ASD difficulty between Optimal Conditions (easier)

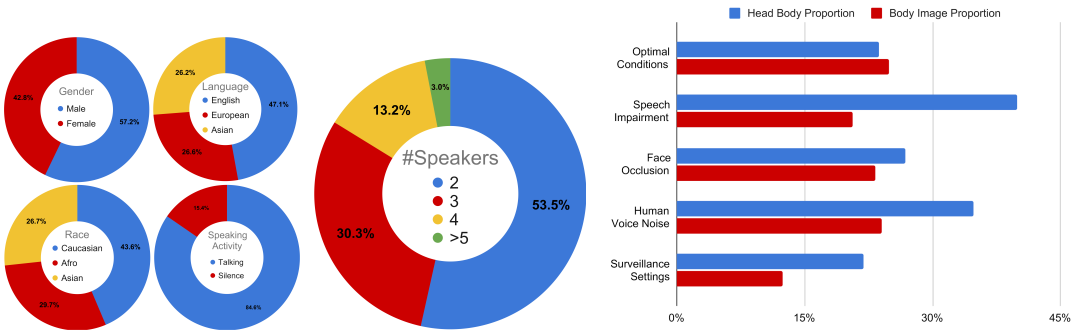


Figure 5.3: Gender, language, race, speaking activity, and number of speakers distribution of WASD. Afro refers to African and Afro-American people. On the right, distribution of head-body and body-image proportions of WASD categories. WASD is a balanced demographics dataset, with *talking* being the predominant speaking activity, mainly composed of few people conversations, where audio impaired categories (Speech Impairment and Human Voice Noise) have speakers closer to the camera, and Surveillance Settings has speakers further from it.

and Surveillance Settings (harder), since the impairment of audio and face is incremental and controlled throughout the categories. Figure 5.2 displays representative images of each category and the relative variation of audio and face quality. For a detailed list of all the considered features per video in WASD, we refer to the supplementary materials and the metadata Comma-Separated Values (CSV)) containing this information, available on GitHub.

WASD Groups. Aside category division, we also form two groups of videos for our experiments: **Easy** and **Hard**. The easy group contains the categories that more closely resemble AVA-ActiveSpeaker (*Optimal Conditions* and *Speech Impairment*) while the hard group has categories where one or both factors (face and audio) are specifically degraded (remaining 3 categories of WASD). The inclusion of Speech Impairment in the easy group relates to how speech overlap is admissible in AVA-ActiveSpeaker (as recurrent from normal conversations) and speech delay as a result of dubbed movies (existent in AVA-ActiveSpeaker).

5.3.2 Main Characteristics

One focus of the proposed dataset is ensuring that each category is balanced regarding language, race, and gender distribution to mitigate any potential bias in future experiments. The languages are grouped into English, European, and Asian, while races are grouped into Caucasian, Afro, and Asian. The considered languages and races, their grouping, and other related considerations are discussed in the supplementary material. The distribution of demographics, number of speakers, and head-body proportions of WASD is presented in Figure 5.3. WASD only considers two admissible labels, with talking being the dominant speaking activity (contrary to AVA-ActiveSpeaker), and is mainly composed of few people conversations. Surveillance Settings is the one with lesser camera proximity to speakers while Speech Impairment and Human Voice Noise have speakers closer to the camera.

Following the AVA-ActiveSpeaker approach, the maximum length considered for each

video is 15 minutes. Contrary to AVA-ActiveSpeaker, where each subvideo duration ranges up to 10 seconds, we segment each subvideo up to 30, with varying video FPS, mainly ranging from 24 to 30. Regarding the number of videos, WASD is composed of 164 videos (*vs.* 153 of AVA-ActiveSpeaker), totalling 30 hours of video annotations, divided into train and test with a similar proportion to AVA-ActiveSpeaker (80/20), with each category having the same amount of hours, (*i.e.*, 6 hours) and demographics balance.

5.3.3 WASD Annotations

Body bounding boxes drawing and tracking are obtained using YOLOv5 [540] and DeepSort [591], serving as input to Alphapose [521, 522, 523], which outputs pose information for each subject per frame. Then, we obtain face bounding boxes [592] from pose data, using eyes, ears, and nose keypoints as reference for bounding box drawing. The size of face bounding boxes is based on body bounding box height, which is adjusted manually per video to ensure adequate face capture. All face and body annotations are manually revised by a human and adjusted/fully annotated when necessary via Computer Vision Annotation Tool (CVAT) [593]. For speaking annotations, we design a custom Graphical User Interface (GUI) program in Python for manual annotation, outputting a file with the format used by AVA-ActiveSpeaker. Further details regarding annotations can be seen in the supplementary material.

5.4 Experiments

5.4.1 Datasets, Models, and Evaluation Metric

Datasets. The AVA-ActiveSpeaker dataset [59] is an audio-visual active speaker dataset from Hollywood movies. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723 and 8,015 video utterances, respectively, ranging from 1 to 10 seconds. The main challenges of this dataset are related to language diversity, FPS variation, the existence of faces with low pixel numbers, blurry images, noisy audio, and dubbed dialogues. Similar to other works, we report the obtained results on the AVA-ActiveSpeaker validation subset. We also use the proposed dataset, WASD, which is described in Section 5.3. Unless explicitly stated, all models trained in WASD use the whole training split (with 5 categories). We also consider Columbia [570] for cross-domain experiments with model modifications to include body information. Columbia consists of an 87-minute panel discussion video, with five speakers (Bell, Boll, Lieb, Long, and Sick) taking turns speaking, with 2-3 speakers visible at any given time.

Models. The considered models are the ones with state-of-the-art results and publicly available implementations: ASC [69], MAAS [72], TalkNet [1], ASDNet [70], TS-TalkNet [77], and Light-ASD [76]. ASC, MAAS, and ASDNet are trained in a two-step process, while TalkNet, TS-TalkNet, and Light-ASD are trained end-to-end. MAAS did not provide its Multi-modal Graph Network setup so we present the results from the available

Table 5.3: Comparison of AVA-ActiveSpeaker trained state-of-the-art models on AVA-ActiveSpeaker and categories of WASD, using the mAP metric. We train and evaluate each model following the authors' implementation. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. AVA refers to AVA-ActiveSpeaker, Light refers to Light-ASD, and TS-Talk to TS-TalkNet.

Model	AVA	WASD					
		OC	SI	FO	HVN	SS	Avg
ASC [69]	83.6	86.4	84.8	69.9	66.4	51.1	74.6
MAAS [72]	82.0	83.3	81.3	68.6	65.6	46.0	70.7
ASDNet [70]	91.1	91.1	90.4	78.2	74.9	48.1	79.2
TalkNet [1]	91.8	91.6	93.0	86.4	77.2	64.6	85.0
TS-Talk [77]	92.7	91.1	93.7	88.6	79.2	64.0	85.7
Light [76]	93.4	93.1	93.8	88.7	80.1	65.2	86.2

Table 5.4: Comparison of state-of-the-art models on the different categories of WASD, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings. Light refers to Light-ASD, and TS-Talk to TS-TalkNet.

Model	Easy		Hard			Avg
	OC	SI	FO	HVN	SS	
ASC [69]	91.2	92.3	87.1	66.8	72.2	85.7
MAAS [72]	90.7	92.6	87.0	67.0	76.5	86.4
ASDNet [70]	96.5	97.4	92.1	77.4	77.8	92.0
TalkNet [1]	95.8	97.5	93.1	81.4	77.5	92.3
TS-Talk [77]	96.8	97.9	94.4	84.0	79.3	93.1
Light [76]	97.8	98.3	95.4	84.7	77.9	93.7

implementation.

Evaluation Metric. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [59] that computes mAP, while for Columbia we use F1 score.

5.4.2 Limitations of AVA-ActiveSpeaker Training

We start by training models in AVA-ActiveSpeaker and evaluate their performance on AVA-ActiveSpeaker and WASD, in Table 5.3. The inverse cross-domain (WASD to AVA-ActiveSpeaker) is included in the supplementary materials.

Similar to AVA-ActiveSpeaker. Regardless of the model, their performance on Easy categories (Optimal Conditions and Speech Impairment) is similar to the one displayed in AVA-ActiveSpeaker, suggesting the presence of similar characteristics between this group and AVA-ActiveSpeaker. This highlights the importance of face and audio quality for current ASD models, and shows that with high quality data and reliable face access, simultaneous talk or slight speech delay do not significantly hinder model performance. Furthermore, the similar performance of models in AVA-ActiveSpeaker and Easy categories support the quality of WASD annotations.

Face and Audio Importance. However, the cross-domain performance is significantly worse in Hard categories. In Face Occlusion, Human Voice Noise, and Surveillance Settings, there is a decrease in performance relative to other categories, suggesting that impairment of face access or audio quality significantly impact models, with a cumulative

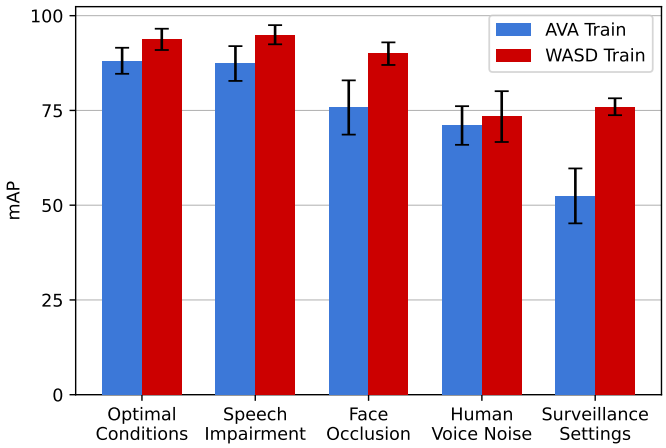


Figure 5.4: Average performance (mAP) variation of the four models on WASD categories, when trained on AVA-ActiveSpeaker and WASD. AVA-ActiveSpeaker is represented as AVA.

degrade when both are present (Surveillance Settings). Furthermore, facial occlusion is not as impactful as audio impairment (Human Voice Noise) in ASD, meaning that even when a model can not assess the talking person via face, it can still deduct it via audio analysis. The inverse is not as easily solved, since the existence of audio impairment with human voices (Human Voice Noise) leads to poorer performance relative to the Face Occlusion.

Model Performance Variance. One particular aspect regarding the performance of models for the Easy group is TalkNet and TS-TalkNet performing slightly better in Speech Impairment than in Optimal Conditions. This is mainly due to their long-term audio assessment approach, making the models robust against local audio variations and resilient to settings with slight audio impairment (Speech Impairment), even when trained in AVA-ActiveSpeaker, partially due to the existence of Speech Delay in AVA data (dubbed movies). Given this context, for TalkNet(s), Speech Impairment and Optimal Conditions tend to be "similar" environments, suggesting that the performance difference might be more related to the background of the videos and not their ASD challenges (*i.e.*, for TalkNet(s), Optimal Conditions and Speech Impairment only represent different videos and the latter is not necessarily harder). Additionally, Speech Impairment category has faces closer to the camera relative to Optimal Conditions (Figure 5.3, head-body proportion bar), which further contributes to the improvement of TalkNet(s) performance (*i.e.*, Speech Impairment is less challenging for TalkNet(s) than other models, and the face being closer to the camera leads to a slightly easier setting than Optimal Conditions).

Best Performers. Despite a performance degrade with increasing category difficulty, TalkNet, TS-TalkNet, and Light-ASD tend to perform better than the remaining models. This could be linked to their end-to-end approach for ASD, contrary to the other models, improving its generalization and performance in cross-domain. Furthermore, TalkNet and TS-TalkNet focus on long-term temporal context, benefiting from longer videos which is the case of WASD, while the BGRU for cross-modal modeling and lightweight model of Light-ASD contribute to more robustness for unseen conditions.

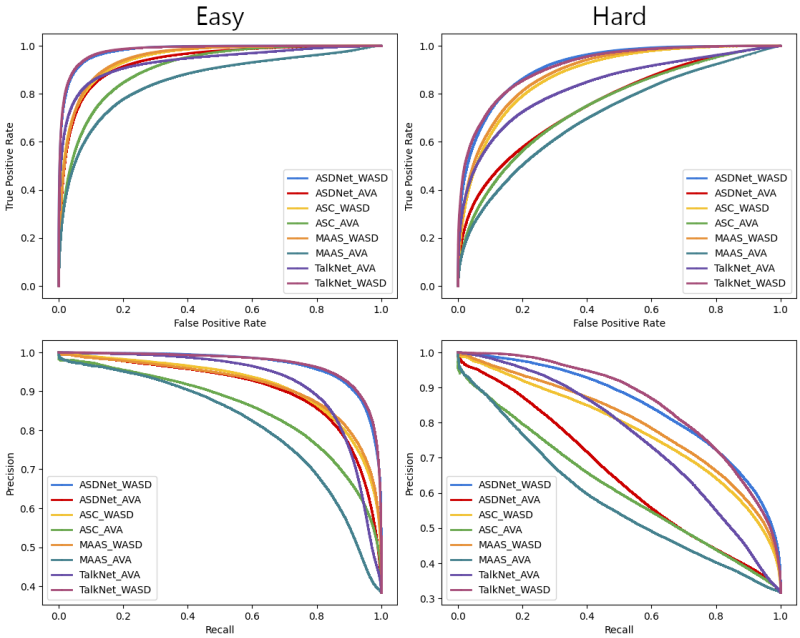


Figure 5.5: ROC and PR curves for models trained in AVA-ActiveSpeaker and WASD, and evaluated in Easy (left) and Hard (right) groups of WASD. AVA-ActiveSpeaker is represented as AVA.

5.4.3 Models Robustness in WASD

To evaluate the robustness of ASD models on challenging data, we train them in WASD and compare their performance with AVA-ActiveSpeaker training in Table 5.4 and Figure 5.4.

Performance Increase. Relative to AVA-ActiveSpeaker training, models trained in WASD tend to slightly improve their performance in Easy setups (Optimal Conditions and Speech Impairment), with higher increase in Face Occlusion and Surveillance Settings scenarios. The increase in Face Occlusion to closer values of those in Easy setups shows that, if trained accordingly, current models can perform ASD in such scenarios. This relates to how models can map different speaker relations in a scene, allowing the inference of one speaker relative to others, even if the face is occluded. Regarding Surveillance Settings, it shows that AVA-ActiveSpeaker does not contain data similar to these settings, but models can perform better in such scenarios if given the proper training. Similar to Face Occlusion, relating different speakers in a scene may give models the tools to perform in such scenarios, even when face access is not reliable.

Model Performance Differences. Regarding the performance of Human Voice Noise and Surveillance Settings, some models tend to perform better in the latter, even if said category is expected to be more challenging. We can look at the results considering two big groups: 1) ASDNet, TalkNet, Light-ASD, and TS-TalkNet (where Surveillance Settings is worse or equal to Human Voice Noise); and 2) MAAS and ASC (where Surveillance Settings is clearly better than Human Voice Noise). Older models (MAAS and ASC) are more heavily based on combining audio and face without attention-based approaches, and Human Voice Noise challenges an assumption of these models (given their training data and approaches): *audio of a given scene/video is very likely to come from ASD candidates*

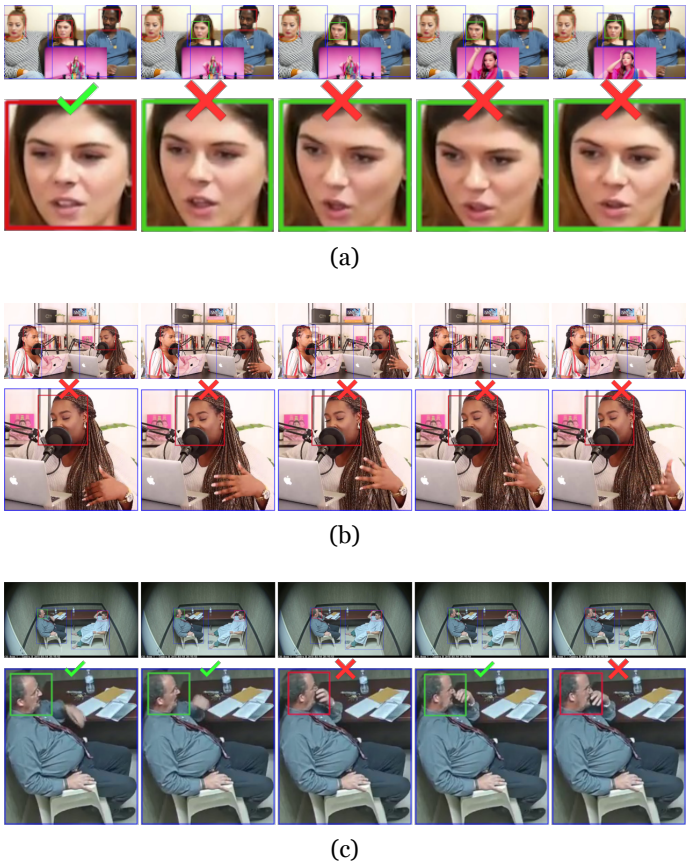


Figure 5.6: Incorrect model inference in different scenarios. Source of misconception: a) awe expression, with sudden and subtle mouth movement, while having human voice in the background; b) partial facial occlusion from scene object; and c) slight mouth occlusion from hand movement.

(*i.e.*, people in the scene). This is not guaranteed in this category since the majority of its videos have talking/singing audio from people not in the scene, which is distinctive from all other categories, such as Surveillance Settings (*i.e.*, in Surveillance Settings all talking audio comes from people in the scene), which might justify the underperformance of older models and the resilience of the most recent ones.

Model Limitations. When trained in WASD, models can not improve their performance in the presence of disruptive/distracting human voice background (Human Voice Noise), which shows the limitations of current approaches. The guaranteed face access may induce a false sense of security to classify a person as talking when they do micro expressions in the presence of (background) human voice. Furthermore, the disparity between the results with human voice background or surveillance settings and the other scenarios (77% *vs* >92%) shows the limitations of current models to perform in wilder ASD contexts, particularly in impaired audio conditions.

Performance in WASD Groups. To complement model performance assessment, we compute the ROC and PR curves of models in different experimental settings, in Figure 5.5. The results show that: 1) in the Easy group, ASDNet and TalkNet trained in AVA-ActiveSpeaker are competitive with other models trained in WASD, showing the robust-

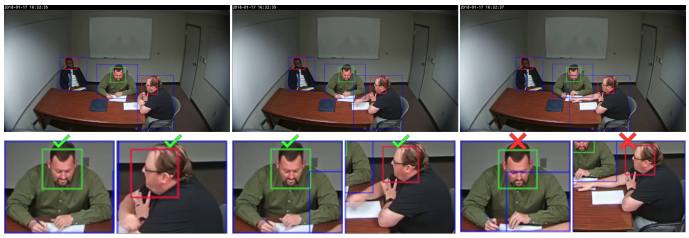


Figure 5.7: Incorrect model inference by mixing active speakers. Hand and arm movement (from right speaker, better viewed with zoom in) suggest a change in conversation between the two speakers, whose analysis would aid understanding speaker swap mid conversation.

ness of the best performing models and the similarity between AVA-ActiveSpeaker and Easy group of WASD; 2) for the Hard group, all models trained in WASD have superior performance relative to AVA-ActiveSpeaker training, suggesting the difference of data between this group and AVA-ActiveSpeaker; and 3) TalkNet trained in AVA-ActiveSpeaker displays a different tendency relative to other models, expressed in both Easy and Hard group, with higher predominance in PR curves. TalkNet has a cautious and precise approach in determining the active speaker (high precision), while not keeping a similar performance in identifying all the active speakers as other models (lower precision with higher recall). This is linked to the lower talking percentage of AVA-ActiveSpeaker and the end-to-end approach of TalkNet with emphasis on long term context: identifying only active speakers with high confidence is a good strategy in AVA-ActiveSpeaker but not as reliable in WASD.

5.4.4 Qualitative Analysis

We analyze different scenarios where WASD is distinctive from AVA-ActiveSpeaker and body data analysis is more relevant for ASD, namely in Human Voice Noise, Face Occlusion, and Surveillance Settings, represented in Figures 5.6a, 5.6b, 5.6c, and 5.7, respectively. Head boxes are colored with models predictions, trained in WASD: green, person is talking; red, not talking. Figures are accompanied with zoom ins containing wrong and correct signs, displaying the correctness of ASD prediction. By not using body information, state-of-the-art models can not reliably deal with scenarios where someone expresses slight lip movement (*e.g.*, awe expression) when another person (not in scene) is talking (Figure 5.6a), or with facial occlusion (Figure 5.6b), even in the context of speaker proximity and cooperation. In surveillance settings (Figures 5.6c and 5.7) the benefits of body data evaluation is even more pronounced. Accessing hand movement with slight face occlusion helps understanding that the same person is talking (Figure 5.6c), as well as inferring when one person is requesting other to stop talking (Figure 5.7).

5.5 Proposed Baseline

Given the superiority of TalkNet in WASD and its end-to-end approach, we select it as our baseline, modifying its architecture to include body information while maintaining

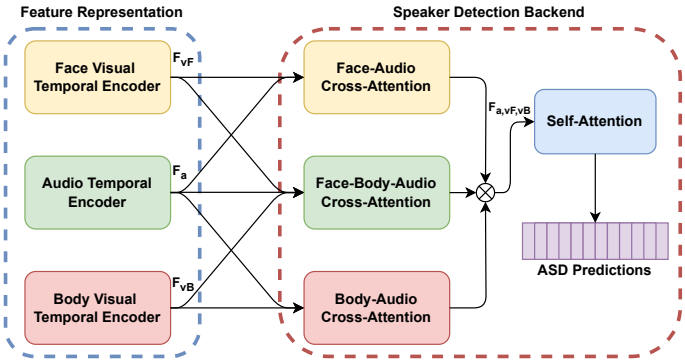


Figure 5.8: Adaptation of TalkNet [1] architecture to include body data and serve as our baseline: inclusion of a visual temporal encoder for body data and cross-attention for audio and body.

Table 5.5: Comparison of TalkNet, using face and body data, and proposed Baseline on WASD categories, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings.

Model	Easy		Hard		
	OC	SI	FO	HVN	SS
TalkNet$_{Face}$	95.8	97.5	93.1	81.4	77.5
TalkNet$_{Body}$	91.1	95.5	88.4	73.1	75.0
Baseline	96.9	98.1	95.4	83.8	81.5

its key characteristics, as shown in Figure 5.8. In experiments with only one visual data input (face or body), we use the original TalkNet.

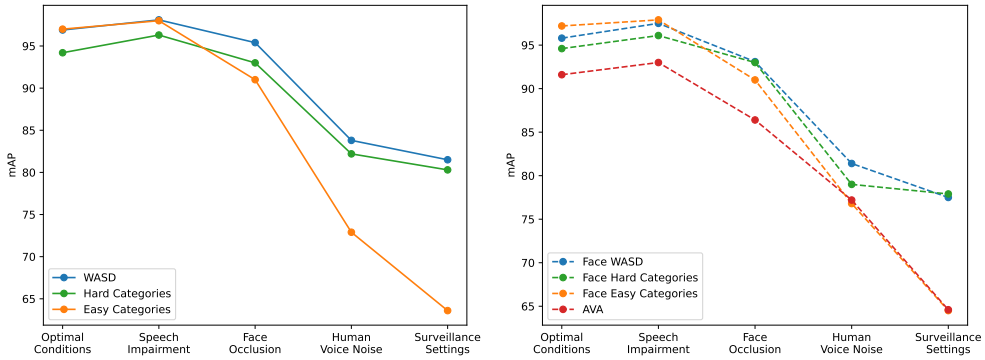
5.5.1 TalkNet Architecture Modifications

To ensure adequate body data encoding, we add a new visual encoder for body data, with the same architecture as the one used for face (Video-Temporal Convolutional (V-TCN) [1] with Depth-wise Separable Convolutional Layers (DS-Conv1D) [594]). For multi-modal cross-attention, we combine audio and body features following the approach used for face and audio [595]. Feature combination was adapted to include body and self-attention block input was updated to consider these changes. Given that the original training loss function considers the prediction of feature combination and each feature individually with weighted factors, we include a body loss with the same weight as face.

5.5.2 Body Importance Assessment

We assess the performance of the proposed Baseline and TalkNet using face and body data, in Table 5.5.

Body Data Alone. Although body data contains facial cues, model attention to them diminishes when using body as sole input, leading to worse performance overall. The categories less affected by this change are Speech Impairment and Surveillance Settings, given the similarity of face and body data (head body proportion in Figure 5.3), and reduced face importance due to its unreliable access, respectively. This shows that current ASD approaches can not effectively process body data as input and are more prepared to



(a) (b)

Figure 5.9: WASD group training performance (mAP) influence on: a) Proposed Baseline; and b) TalkNet with face as input. AVA refers to TalkNet trained in AVA-ActiveSpeaker, and WASD refers to training using all training data.

analyze closed-up facial/mouth cues for ASD prediction.

Face Body Importance Discrepancy. Regarding the difference between using face or body alone, Human Voice Noise is the category where this difference is maxed. Human Voice Noise has videos with sporadic/involuntary body movements that might suggest that a person is talking, while not actually being (*e.g.*, dancing, pointing at something, raising hand in reaction to a given event). In these cases, considering only body information as visual input in Human Voice Noise data which contains movements that are rarely/never seen in other categories, might be detrimental towards ASD and lead to false positives based on sudden/abrupt body movements (*i.e.*, when looking at other categories, raising hands or similar actions are heavily related to speaking, which is not guaranteed in Human Voice Noise).

Face Body Complement. The combination of body with face data translates into better results for all categories, with the biggest improvement on Hard categories. This shows that face-body complement is a viable strategy for current models and body importance is emphasized in harder setups, particularly when face can not be reliably accessed.

5.5.3 Training Influence in Model Performance

We explore the effect of training in WASD groups (Easy and Hard, Section 5.3.1) on the proposed Baseline and TalkNet with face as visual input, in Figure 5.9.

Group Training on Baseline. For our Baseline, training in Easy has similar performance to training in WASD for cooperative setups, but significantly underperforms in unseen challenging sets. In contrast, Hard training follows the trend of WASD training (with only slightly worse performance), showing the importance of training in challenging data for robustness in within and cross-domain settings.

Training Influence on Face-based Models. Similar to our Baseline, face-based models overall performance is better when training in Hard than in Easy. However, while in Baseline there is a gap between Hard and WASD training, in face-based models the performance is more similar, showing the importance of having additional training data

Table 5.6: Comparison of F1-Score (%) on the Columbia dataset.

Model	Speaker					
	Bell	**Boll**	**Lieb**	**Long**	**Sick**	**Avg**
TalkNet [1]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [596]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [76]	82.7	**75.7**	**87.0**	74.5	**85.4**	**81.1**
Baseline	**83.0**	64.5	79.6	**86.2**	78.9	78.4

for more robust approaches (body-face models). Relative to TalkNet training in AVA-ActiveSpeaker, its performance on WASD follows the same trend as Easy training, reinforcing that AVA-ActiveSpeaker and Easy data are similar, and that neither adequately prepare models to deal with challenging sets.

5.5.4 Body Importance in Columbia

We also assess the performance of the proposed baseline in Columbia, following the procedure of Light-ASD [76] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table 5.6. Since AVA-ActiveSpeaker does not have body data annotations, we obtain body bounding box annotations from AVA Actions Dataset [140] and complement them with speaking labels of AVA-ActiveSpeaker to train the proposed baseline. The results show that body inclusion massively improves the proposed baseline performance relative to its initial architecture (TalkNet), showing that body information leads to increased robustness to perform ASD and can be useful in other settings other than the challenging conditions of WASD.

5.6 Conclusion

We propose WASD, a innovative challenging ASD dataset with degraded audio quality, facial occlusions, surveillance conditions, and body data annotations. We show that current datasets do not prepare models for wild ASD, and state-of-the-art models underperform in such conditions, particularly in audio impairment and surveillance settings. We also propose a baseline that complements body data with audio and face to prove the importance of body for wild ASD and support the development of subsequent approaches, given the unreliability of audio quality and subject cooperation in wilder settings. Given the challenges of the surveillance data relative to existing datasets, future directions should be centered on increasing the amount of challenging data available to prepare models to perform in wilder conditions.

Chapter 6

BIAS: A Body-based Interpretable Active Speaker Approach

State-of-the-art Active Speaker Detection (ASD) approaches heavily rely on audio and facial features to perform, which is not a sustainable approach in *wild* scenarios. Although these methods achieve good results in the *standard* AVA-ActiveSpeaker set, a recent wilder ASD dataset (WASD) showed the limitations of such models and raised the need for new approaches. As such, we propose BIAS, a model that, for the first time, combines audio, face, and body information, to accurately predict active speakers in varying/challenging conditions. Additionally, we design BIAS to provide interpretability by proposing a novel use for Squeeze-and-Excitation blocks, namely in attention heatmaps creation and feature importance assessment. For a full interpretability setup, we annotate an ASD-related actions dataset (ASD-Text) to finetune a ViT-GPT2 for text scene description to complement BIAS interpretability. The results show that BIAS is state-of-the-art in challenging conditions where body-based features are of utmost importance (Columbia, open-settings, and WASD), and yields competitive results in AVA-ActiveSpeaker, where face is more influential than body for ASD. BIAS interpretability also shows the features/aspects more relevant towards ASD prediction in varying settings, making it a strong baseline for further developments in interpretable ASD models, and is available at `https://github.com/Tiago-Roxo/BIAS`. The work presented in this chapter has been published in [4].

6.1 Introduction

Current Active Speaker Detection (ASD) models are known to perform reliably using only audio and face-based information. This is mainly due to the fact that state-of-the-art ASD datasets have good audio and face quality, yielding from controlled setups (movies, from AVA-ActiveSpeaker [59]) and cooperative settings (interviews, from AWS [60]). Recently, WASD [3] has been announced as a more challenging set, with degraded audio and face data quality, corresponding to less constrained data acquisition scenarios.
In this context, the existing models are not suitable for wilder settings, where audio quality might be poor and faces occluded (Figure 6.1). As such, we propose BIAS, an approach for ASD, that uses, for the first time, body data to complement face and audio-based features, achieving state-of-the-art results in challenging sets (WASD and open-settings of Columbia), and competitive results in more cooperative conditions (AVA-ActiveSpeaker), where the body relevance is reduced since the face is the predominant feature in this data. Furthermore, we propose a novel use of Squeeze-and-Excitation (SE) blocks [2] to provide reasoning for model decision and analyze the importance of different features. This



Figure 6.1: Illustration of the **BIAS insight**: in surveillance settings, where facial and audio-based features might not be always available, body data should be crucial to accurately detect the active speakers. In such challenging conditions, providing reliable explanations for the reasoning behind the provided responses is also an important feature. This paper describes BIAS, which singularly uses facial, audio, and body-based features, also providing visual interpretability and feature importance assessment for its responses.

way, we are able to obtain visual interpretability and assess feature influence in varying ASD settings, via SE vector manipulation. Finally, to improve the interpretability of ASD-related scenarios, we complement the attention of BIAS (visual) with descriptions of existing actions, via caption generation (textual). Given the absence of training data of subtle actions related to ASD tasks (*e.g.*, raise hand, cross arms, open mouth to talk), we use WASD data to annotate an ASD-related actions dataset (ASD-Text) and finetune a GPT model on it which, in conjunction with the visual interpretability of BIAS, can be used for a full interpretability setup. To summarize, the main contributions are:

- We propose BIAS, an ASD model based on a long temporal context approach, using audio, face, and body-based information. With respect to the state-of-the-art, BIAS is innovative by being the first method to use body-based information for ASD, yielding competitive results in AVA-ActiveSpeaker and is state-of-the-art in WASD and Columbia (open settings version);

- We propose a novel use of SE blocks for attention heatmap creation (visual interpretability) and perceive feature importance in the decisions, which can be easily included in common and customized architecture, for image and video settings;

- To improve the interpretability of ASD-related scenarios, we annotate an ASD-related

actions dataset (ASD-Text) for text scene captions, providing a training setup for fine scene description in the ASD context.

6.2 Related Work

ASD Context. Given the task to determine the talking speaker from a set of admissible candidates, various state-of-the-art ASD datasets have been recently proposed [570, 72, 571, 59, 60]. Columbia [570] contains 87 minutes of a panel discussion and Talkies [72] focuses on low duration videos, totalling 4 hours, with off-screen speaking. Easycom [571] is designed for multiple augmented reality tasks, composed of various speaking sessions with background noise. AVA-ActiveSpeaker [59] contains Hollywood videos totalling almost 38 hours, with demographic diversity and FPS variation, and has application in other topics such as audio anomaly detection [597]. ASW [60] has over 30 hours, with videos randomly selected from the VoxConverse [590], containing various sets of interviews. Recently, WASD [3] has been announced, with 30 hours of data grouped based on the audio and face quality, with balanced demographic diversity and body annotations data.

ASD Models. Based on the available data for ASD, current state-of-the-art heavily rely on face and audio data, with audiovisual combination using 3D architectures [61], hybrid 2D-3D models [62], large-scale pretraining [63, 64], feature embedding improvement [65], and attention mechanisms [66, 67, 68]. Various ASD works [69, 70, 71, 72] based their model on a two-step process, where the first focuses on short-term analysis (audio with face combination) and the second on multi-speaker analysis. ASC [69] focuses on multi speaker analysis via temporal refinement, ASDNet [70] uses a similar approach for inter-speaker relations with improved visual backbones, and UniCon [71] relies on audio-visual relational contexts with various backbones. The improvement of ASD performance by assessing contextual information via speaker relation using GCN [73] has also been explored [72, 74, 75]. Diverging from two-step training, end-to-end models have also emerged for ASD [1, 75, 74, 76]. TalkNet [1] focused on improving long-term temporal context with audio-visual synchronization, EASEE [75] included GCN to complement spatial and temporal speaker relations, and Light-ASD [76] proposed a lightweight model by splitting 2D and 3D convolutions for audio-visual feature extraction, and applied BGRU for cross-modal modeling.

Body Information for Attribute Recognition. Although recent works on ASD do not use body information, this data contains information that could contribute to improve model performance, particularly in wilder conditions (*e.g.*, surveillance settings), where face is not reliably accessed. PAR datasets [506, 507, 508] are examples of these scenarios, containing person cropped images from surveillance settings, used to identify attributes (*e.g.*, clothing, accessory usage, gender, age) under challenging covariates such as occlusion, pose, image resolution, and luminosity. Works in this area focused on different strategies ranging from different architecture combination [548, 470, 483], attention-based approaches [481, 484], and attribute relation importance [475, 476, 478].

Model Interpretability. For visual interpretability, we can group methods into two

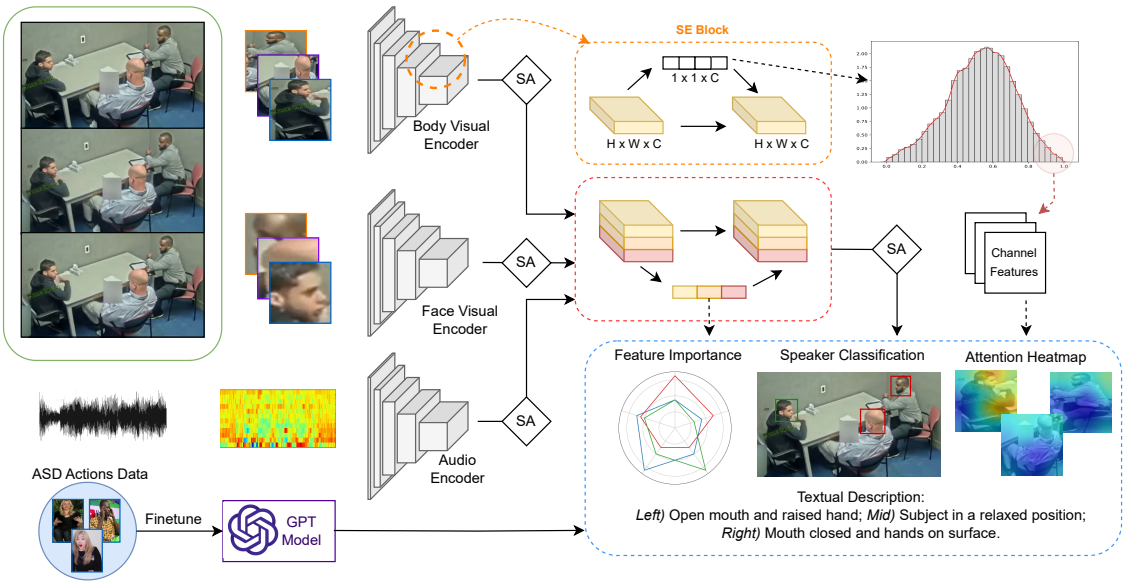


Figure 6.2: Overview of the BIAS architecture and pipeline, with GPT model integration: body and face-based data is fed into the respective visual encoders, while audio is processed into MFCC before encoding. SE blocks are used in visual encoders and feature combination for attention heatmaps and feature relative importance, respectively. SA refers to self-attention blocks. Heatmaps are created by combining channel features of the respective top 10% SE vector values. BIAS prediction is based on feature combination, accompanied with visual interpretability and feature importance assessment, complemented by text descriptions from a GPT model finetuned in ASD-related actions data (ASD-Text).

main categories: *gradient based* [598, 599, 600, 601, 602] (gradients of each layer, computed through backpropagation) and *attribution propagation* [603, 604, 605, 542, 606] (recursive decomposition of layers contributions, all the way to model's input). Saliency based methods [607, 608], Excitation Backprop [609], and Perturbation methods [610, 611] are also visual interpretable approaches in computer vision, with Transformer-based interpretability [612] being recently explored. Although most works explore model interpretability in object classification datasets, its use in face [613, 614], body [615], and PAR [616, 480] data is not unprecedented. Contrary to current approaches, we propose a SE block-based visual interpretation, obtainable in inference time, without requiring additional computational cost for attention heatmaps creation.

6.3 BIAS Approach

We propose BIAS, an interpretable model for ASD, based on a long temporal context approach, using audio, face, and body data. We process data with modified encoders and customized feature vector combination, using SE blocks to provide visual interpretability and feature importance assessment, respectively. BIAS is distinctive from other state-of-the-art models by using, for the first time, body data in ASD tasks, translating into state-of-the-art performance in challenging sets and competitive results in cooperative conditions. The overall architecture is displayed in Figure 6.2, with details of each part in the following subsections. Note that, although body data also contain facial cues, these are not as easily perceived as body movements (given the face/body proportion), which

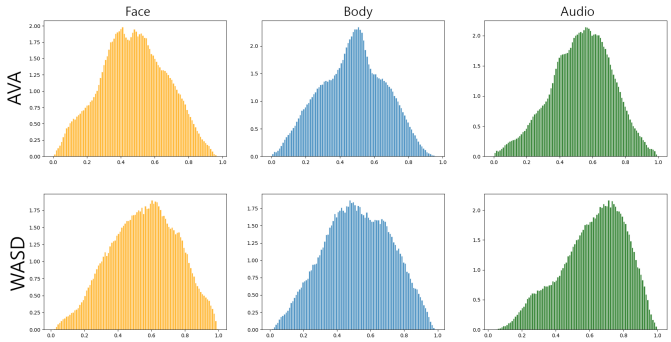


Figure 6.3: SE vector values from feature (audio, body, and face) combination, for BIAS trained in AVA-ActiveSpeaker and WASD. Both datasets follow a normal distribution. AVA refers to AVA-ActiveSpeaker.

creates a stronger focus on hands and arms movements.

6.3.1 Visual and Audio Encoders

Backbone. We obtain a sequence of frame-based embedding using a customized backbone, based on the ResNet18 architecture, composed of a 3D convolutional layer followed by ResNet18 [538] layers and a SE block, before average pooling. The inclusion of a SE block is to provide a way to retrieve informative features, via assessment of SE vector values, useful for visual interpretability, which we discuss in Section 6.3.2.

Temporal Enconding. For long-term visual spatio-temporal representation, we use a visual temporal network, composed of 5 depth-wise separable convolutional layers (DS Conv1D [594]) with residual connections, followed by a Conv1D layer for feature dimension reduction (to 128-feature embedding).

Audio Encoder. We obtain audio temporal encoding using a ResNet34 with SE blocks in its layers (SE-ResNet34). Audio frames are represented as 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC), with the audio encoder outputting them as a 128-dimensional audio embedding for subsequent visual and audio conjunction.

6.3.2 Squeeze-and-Excitation for Interpretability

The novel use of SE block in BIAS is included in two parts of its architecture: end of visual backbones (visual interpretability) and feature combination (perceived feature importance).

Visual Interpretability. In our model, we use SE in encoders to obtain the features (channels) perceived as highly influential towards ASD (*i.e.*, high value in SE vector). However, the perception of *high* is relative to the video/categories used, so we can not use a default/hard-coded value to define what is high or low importance. As such, given that the values of the SE vector follow a Normal distribution (Figure 6.3), we use its formula to obtain the top 10% values of the SE vector (and, consequently, the top 10% channels) to use for ASD evaluation. The reason for considering the top 10% in our experiments is to balance model performance and visual interpretability (higher % leads to worse performance, and lower % leads to less clear visual interpretability). Finally, we conjugate

Table 6.1: ASD-Text dataset action labels, grouped by body parts.

Body Part	#	Action Label
Hand	1	Hand raised in the air
	2	Hand touching the face
	3	Hand raised with object
	4	Hand movement (not raised)
Mouth	5	Mouth occlusion with object
	6	Mouth occlusion with hand
	7	Mouth move from speech
	8	Mouth move from expression
	9	Mouth not moving
Arms	10	Crossed arms
	11	Arms behind back
Body	12	Body in relaxed position
	13	Body facing forward
	14	Wild body movement

the selected channels by bicubic interpolation on top of the original image to provide a heatmap of the most important regions for ASD.

Perceived Importance and Feature Influence. Another use of SE in BIAS is to combine audio, face, and body information (embeddings of 128 dimensions). For feature importance assessment, we create three different graphs that display the importance of each feature (audio, face, and body) for the categories of WASD (Table 6.2). For each feature, we normalize the SE values for visualization purposes to better display the feature importance on WASD categories. The normalization is done based on the values obtained by the model for a video and it is not included in the BIAS training process, meaning that BIAS does not rely on additional/external information and can be used for other domains without requiring adaptations. Regarding feature influence in model performance, when the experiments do not contain a certain feature it means that we set the SE vector values to 0 for the channels of the considered feature (*e.g.*, when we consider *only face*, we set the values to 0 for audio and body channels).

6.3.3 Self-attention and Loss Function

Self-attention. To improve feature importance assessment, we use an attention layer based on query, key, value strategy [66], after visual encoders and feature conjugation. Initially used in the Transformer architecture, this approach has been extensively used for attention purposes, with a role in BIAS of improving audio-visual correlation. This translates into more accurate distinction between speaking and non-speaking frames, given each speaker.

Loss Function. The main source for ASD is the conjunction of 3 data sources: audio, face, and body. To improve independent feature relevance for ASD, and to not motivate ASD prediction solely based on their conjunction, we include weighted cross entropy losses using individual data on ASD training prediction. For inference, only the conjunction of 3 data sources is considered.

Table 6.2: Category feature matrix. Feature description: FA, Face Availability; SO, Speech Overlap; DS, Delayed Speech; FO, Facial Occlusion; HV, Human Voice as Background Noise; SS, Surveillance Settings. The absence of a certain feature is presented with ×, while its presence with ✓. Features containing ? refer to non-guarantee of its presence or absence. Green cells refer to features favorable for ASD, while red ones are unfavorable. Retrieved from [3].

Category	FA	SO	DS	FO	HV	SS
Optimal Conditions (OC)	✓	×	×	×	×	×
Speech Impairment (SI)	✓	✓	✓	×	×	×
Face Occlusion (FO)	✓	×	×	✓	×	×
Human Voice Noise (HVN)	✓	×	×	×	✓	×
Surveillance Settings (SS)	?	?	?	?	?	✓

6.3.4 ASD-Text Dataset

For text scene description, we finetune a ViT-GPT2 [617, 618] model on WASD data, with ViT trained in ImageNet [619] and GPT2 in WebText [618]. Before annotations, we define a set of 14 admissible ASD-related actions, in Table 6.1. Then, for each action we create 3 admissible captions for each gender to ease data annotation, and construct a GUI to annotate. For each subject image, the annotators select the gender and all existing actions on it, using the GUI to create annotations with predefined admissible captions from the selected actions and gender. The total annotated data is composed of 47 246 captions, and 11 733 images from WASD subvideos (up to 30s), randomly selected from all videos of the WASD train set. Similar to the COCO captions dataset [620], we divide our annotations into 90/10 train/test [621]. The caption prediction of ViT-GPT2 model in Section 6.5.3 is done on images from the test set of WASD.

6.3.5 Implementation Details

BIAS is trained with an Adam optimizer, with a initial learning rate of 10^{-4}, decreasing 5% for each epoch. All visual data is reshaped into 112 x 112, audio data is represented by 13-dimensional MFCC, and both visual and audio features have an encoding dimension of 128. Self-attention uses a transformer layer with 8 attention heads. For visual augmentation, we perform random flip, rotate and crop, while for audio augmentation, we use negative audio sampling [1]. In sum, given a video data during training, a audio track of a new one is randomly selected from the same batch as noise, maintaining the same speaking label of the original soundtrack. Since AVA-ActiveSpeaker does not have body data annotations, we obtain body bounding box annotations from AVA Actions Dataset [140] and complement them with speaking labels of AVA-ActiveSpeaker. ViT-GP2 model is finetuned for 3 epochs, using AdamW optimizer, with a learning rate of $5{\times}10^{-5}$, without weight decay or warmup steps.

6.4 Experiments

6.4.1 Datasets, Models, and Evaluation Metrics

Datasets. The AVA-ActiveSpeaker dataset [59] is an audio-visual active speaker dataset from Hollywood movies. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723 and 8,015 video utterances, respectively, ranging from 1 to 10 seconds. The main challenges of this dataset are related to language diversity, FPS variation, the existence of faces with low pixel numbers, blurry images, noisy audio, and dubbed dialogues. Similar to other works, we report the obtained results on the AVA-ActiveSpeaker validation subset.

The WASD dataset [3] compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*. With 30 hours of labelled data, WASD is divided into 5 categories with varying degrees of audio and face quality, grouped into categories: Optimal Conditions (OC), Speech Impairment (SI), Face Occlusion (FO), Human Voice Noise (HVN), and Surveillance Settings (SS). Table 6.2 presents the main characteristics of WASD categories. WASD contains 164 videos, with varying FPS, averaging 28 second duration, with balanced demographics, and similar train/test division as AVA-ActiveSpeaker (80/20). We report the results on each category and Easy-Hard grouping, following WASD experiments (Easy: OC and SI, Hard: FO, HVN, and SS).

We also consider Columbia [570] following the methodology of Light-ASD [76] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning. Columbia consists of an 87-minute panel discussion video, with five speakers (Bell, Boll, Lieb, Long, and Sick) taking turns speaking, with 2-3 speakers visible at any given time.

Models. The considered models are the ones with state-of-the-art results and publicly available implementations: ASC [69], MAAS [72], TalkNet [1], ASDNet [70], and Light-ASD [76]. ASC, MAAS, and ASDNet are trained in a two-step process, while TalkNet and Light-ASD are trained end-to-end. MAAS did not provide its Multi-modal Graph Network setup so we present the results from the available implementation.

Evaluation Metrics. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [59] that computes mAP, while for Columbia we use F1 score. Following the Microsoft COCO Image Captioning Challenge approach, caption generation is evaluated by ROUGE-L [622], METEOR [623], and BLEU-1 to 4 [624].

6.4.2 BIAS Performance in WASD

To assess the importance of body information for ASD we compare BIAS with the reported results of state-of-the-art models in WASD [3], divided by categories, in Table 6.3.

Similar Performance in Easy. The inclusion of body information in BIAS culminates in state-of-the-art results across all categories, obtaining slightly better results in Easy setups. In these scenarios, the reliability of face access and sound quality, with minor degradation (OC and SI), is enough to warrant a good performance from state-of-the-art models. As such, the complement of body information from BIAS does not translate into

Table 6.3: BIAS and state-of-the-art models performance on the different categories of WASD, grouped by Easy and Hard, using the mAP metric.

Model	Easy		Hard		
	OC	SI	FO	HVN	SS
ASC [69]	91.2	92.3	87.1	66.8	72.2
MAAS [72]	90.7	92.6	87.0	67.0	76.5
ASDNet [70]	96.5	97.4	92.1	77.4	77.8
TalkNet [1]	95.8	97.5	93.1	81.4	77.5
Light-ASD [76]	97.8	98.3	95.4	84.7	77.9
BIAS	**97.8**	**98.4**	**95.9**	**85.6**	**82.5**

Table 6.4: Comparison of BIAS and state-of-the-art models on the AVA-ActiveSpeaker, grouped by the visual encoder used. Models with $*$ customized the reported backbones.

Model	Visual Encoder	Par(M)	Body Data	mAP
ASC [69]	RN18 2D	23.3	×	87.1
MAAS [72]	RN18 2D	21.7	×	88.8
TalkNet [1]	RN18* 2D-3D	15.0	×	92.3
BIAS	RN18* 2D-3D	31.6	✓	**92.4**
ASDNet [70]	RNx101 3D	49.7	×	93.5
EASEE-50 [79]	RN50 3D	74.7	×	94.1
Light-ASD [76]	Conv 2D-1D	1.0	×	94.1

a significant improvement over state-of-the-art models, given the cooperative settings of the Easy group for ASD.

Improvement in Hard. The major difference of BIAS relative to other models is in Hard categories, with degraded audio and face image quality. In particular, audio degraded categories (HVN and SS) are the ones where BIAS obtains the biggest improvement over state-of-the-art, which is linked to body information access from BIAS, diminishing the dependence of audio cues for ASD. Regarding scenarios without reliable access to mouth movement (FO), conjunction of body cues with audio information is a more reliable approach for ASD, translated by the increased performance.

6.4.3 BIAS Performance in Other Datasets

Although we show that inclusion of body information is important for ASD in WASD, in particular for categories with more degraded audio and face data, we also assess its importance in other setups using AVA-ActiveSpeaker and Columbia in Tables 6.4 and 6.5, respectively.

BIAS in AVA-ActiveSpeaker. We compare BIAS with the other state-of-the-art models, by grouping them into sets, based on the backbone used. The results show that body information inclusion contributes to state-of-the-art performance in ResNet18 grouping, but it is slightly worse than superior or custom backbones, which suggests that body data can be used but it is not as relevant in scenarios where face access and audio quality are more reliable [3]. Figure 6.4 shows the face/body area proportion in AVA-ActiveSpeaker and WASD, grouped by categories, corroborating the reduced importance of body in AVA-

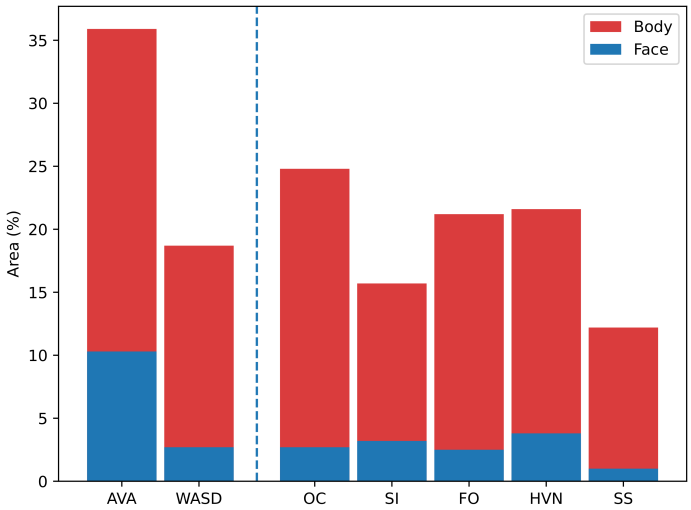


Figure 6.4: Comparison of face and body area, relative to image dimension, in percentage. AVA-ActiveSpeaker contains data with subjects closer to the camera, expressed by higher face and body percentage, relative to WASD and any of its categories. Surveillance Settings (SS) is the category with further distance of subjects from camera.

Table 6.5: Comparison of F1-Score (%) on the Columbia dataset.

Model	Speaker					
	Bell	**Boll**	**Lieb**	**Long**	**Sick**	**Avg**
TalkNet [1]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [596]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [76]	82.7	**75.7**	87.0	74.5	85.4	81.1
BIAS	**89.3**	75.4	**92.1**	**88.8**	**88.6**	**86.8**

ActiveSpeaker data. In this context, the proximity to camera and reliable face access makes body information less contributing to improve ASD, since most of the information usable already exists in the face data. As such, BIAS approach to include body does not translate to a state-of-the-art performance in such settings.

Robustness of BIAS in Columbia. We also assess the performance of BIAS in Columbia, following the methodology of Light-ASD [76] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table 5.6. In this more challenging setting, BIAS approach to combine body with face and audio information leads to a state-of-the-art performance and highlights its robustness to perform ASD in cross-domain settings. Given the results, BIAS is a resilient state-of-the-art ASD model, applicable in scenarios with varying data quality.

6.4.4 Ablation Studies

SE Feature Combination. To complement ASD feature importance for BIAS, we also assess the effect of SE for feature combination and using different features towards BIAS performance in AVA-ActiveSpeaker and WASD, in Table 6.6. The results show that: 1) When using only one visual feature, face is more relevant than body for both datasets; 2) For two feature combination, audio with face is the approach with better results, particularly for AVA; and 3) The aggregated feature combination is better for both datasets, with

Table 6.6: Ablation studies on the effect of SE for feature combination and the contribution of different features towards BIAS performance (mAP) in AVA-ActiveSpeaker (AVA) and WASD.

Settings				AVA	WASD
SE	Audio	Face	Body		
✓	×	×	✓	62.7	84.6
✓	×	✓	×	79.9	90.2
✓	✓	×	✓	75.3	86.2
✓	×	✓	✓	81.2	92.1
✓	✓	✓	×	91.9	92.2
×	✓	✓	✓	91.3	91.8
✓	✓	✓	✓	92.4	94.1

Table 6.7: BIAS performance using different backbones in AVA-ActiveSpeaker (AVA), while maintaining the remaining of the architecture.

Visual Backbone	Par(M)	AVA
ResNet18	31.6	92.4
ResNet50	55.1	92.2
ResNet101	91.3	92.4

Table 6.8: BIAS performance (mAP) in AVA-ActiveSpeaker (AVA) and WASD, with varying body inputs and as the only model input for ASD.

Model Input	AVA	WASD
Body w/ Face Region	62.7	84.6
Body w/o Face Region	44.8	76.0

increased performance in WASD due to the importance of body information in its most challenging categories; and 4) Regarding the effect of SE, its inclusion translates into improved results for both datasets, with a bigger improvement on WASD, which is linked to its more challenging data where adequate feature selection is of utmost importance (*i.e.*, careful selection of face *vs.* body features is more relevant when face may not be reliably accessed).

Backbone Variance. We explore the influence of having bigger backbones for visual feature extraction in Table 6.7, which shows varying backbones do not influence BIAS performance. With ASDNet and EASEE, we see that their higher extraction power is also accompanied with additional computationally heavy components to assess actors' relation (*e.g.*, GCN), and customized lightweight models [76] also achieve state-of-the-art performance which confirms that simply having higher extraction power does not necessarily translate into better performance. This also occurs in BIAS since the combination of face and body features is not directly influenced by varying the backbone given that the combination is done after feature extraction.

Body *vs.* Face Influence. Given the novelty of BIAS to include body for ASD, we explore the importance of body *vs.* face influence, in Table 6.8, where we ignore facial cues of body data by including black boxes in the face region. The results show that, although facial cues are important for ASD, body-related actions also contribute towards BIAS performance. However, for AVA-ActiveSpeaker a significant portion of body information comes from the face region, mainly due to the proximity of subjects to the camera in its

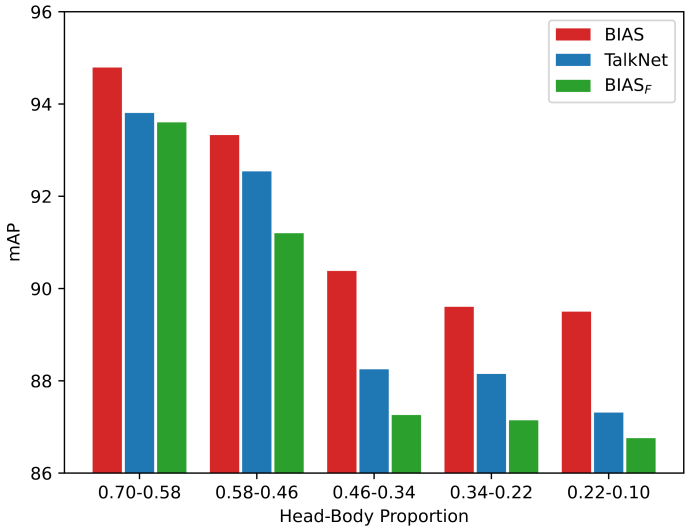


Figure 6.5: Performance of BIAS, TalkNet, and BIAS$_F$ relative to HBP in WASD, across 5 equidistant intervals based on minimum (0.1) and maximum (0.7) HBP. BIAS$_F$ refers to BIAS with only face as visual input.

data (Figure 6.4), which further supports the inability of BIAS to achieve a state-of-the-art performance in such settings. Our approach in this experiment was to deliberately omit information, which can be seen as an adversarial attack, thus decreasing the model performance [625, 626]. As such, we also assess the importance of face *vs.* body with varying HBP, in Figure 6.5, for WASD. The results show that when body is significantly predominant relative to face (*i.e.*, low HBP, meaning that face is small), the model that uses body information (BIAS) is superior to others that only use face (BIAS$_F$ and TalkNet), and this discrepancy is bigger the lower the HBP is. This shows that the performance discrepancy is independent of the model and is mainly due to the difference of data inputs (*i.e.*, it is not facial cues that justify the differences but rather the information present in body data). As such, even if body information also contains facial cues, body is the most relevant feature when face is small and/or not easily accessed.

6.5 BIAS Interpretability

6.5.1 Squeeze-and-Excitation Visual Interpretability

To assess the importance of body for ASD, we create attention heatmaps using SE channels, following the methodology described in Section 6.3.2. These heatmaps highlight the region where BIAS is focusing on when predicting the active speaker of a given scene. We consider different scenarios to assess the relative importance of face and body via BIAS attention heatmaps, in Figure 6.6, where each scenario contains the initial setting (bigger image) and the next three frames for body and face movements (and their respective heatmaps).

Body Importance for ASD. The scenarios in the intersection of *ASD Related Body Focus* with *Face Available* show that BIAS gives higher attention to body when the scenario contains challenging features, even if face is available or partially accessible. In particular,



Figure 6.6: Context, body and face crops, and respective BIAS attention heatmaps of various scenarios. We group the examples into 4 overlapping sets: 1) face available; 2) occluded or non available face; 3) focus on ASD-body movements, and 4) ignoring body information.

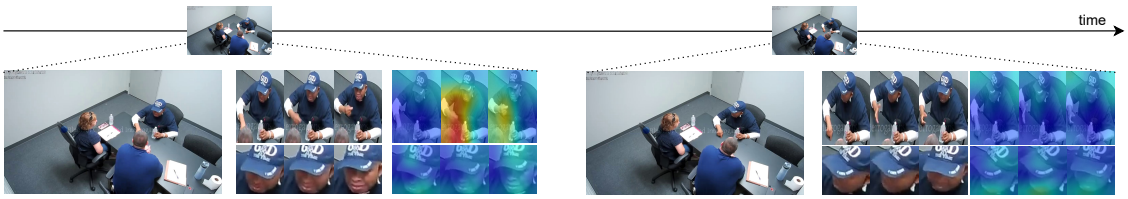
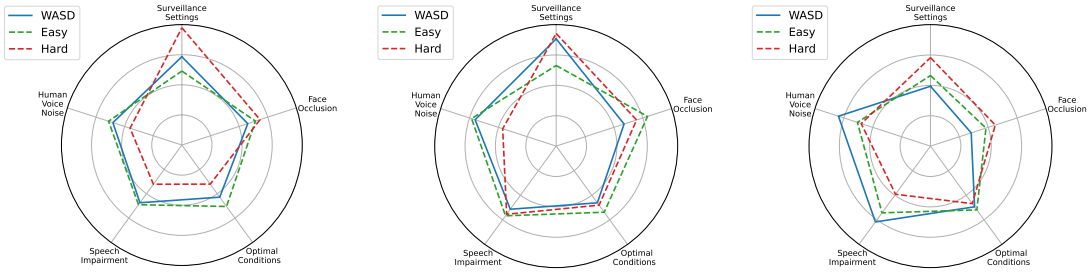


Figure 6.7: Variation of body movement attention in different time lapses of the same context. Although sudden hand movements related to speech are considered by BIAS, subtle ones are not perceived as important, within the same challenging set.

with audio quality subpar (video playing) or limited face availability (hat), body movement importance increases for ASD. When considering the intersection of *ASD Related Body Focus* with *Occluded or Non Available Face*, we are limiting the assess to facial cues which further increases the relative importance of body, translated by BIAS attention to hand movements (in these examples). This shows that body information complement is mainly required in more challenging scenarios, where audio and face are not easily retrieved.

Face over Body. The intersection of *Ignores Body Information* with *Face Available*, shows examples where BIAS opts to give less relevance to body information (heatmaps of body with little/no coloring). In these cooperative conditions, face is the most relevant/reliable data, leading BIAS to rely exclusively on it and ignore body cues (hand movement in both cases).

No Body Focus. The intersection of *Ignores Body Information* with *Occluded or Non Available Face* shows situations where BIAS can also ignore body information in occluded or non-visible faces. In most cases, BIAS attention to body is mainly linked to pronounced (hand) movement. However, in the scenarios of the intersection, subjects only perform



(a) Audio (b) Body (c) Face

Figure 6.8: Relevance of audio, body, and face-based features for ASD on the five different categories of WASD. Hard-BIAS gives higher audio importance in audio impaired categories, perceives body importance similar to WASD-BIAS for the most body-reliant scenario (Surveillance settings) and gives, overall, less face importance to face-dependent contexts, relative to WASD and Easy-BIAS.

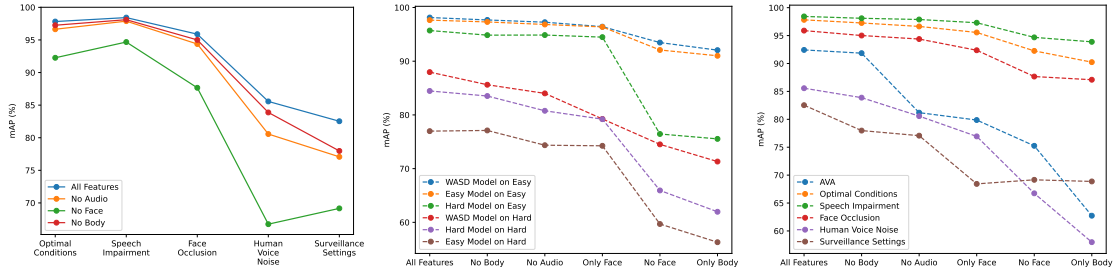


Figure 6.9: Influence of the different features on BIAS performance (mAP): *left*) Effect of audio, face and body absence in WASD categories; *middle*) Feature variation influence relative to different training and testing sets (Easy, Hard, and WASD); and *right*) Feature effect on AVA-ActiveSpeaker (AVA) and WASD categories.

subtle movements when speaking and BIAS response to it is (wrongly) ignoring body information. This shows that BIAS body attention is not entirely linked to speaking activities but more on pronounced/abrupt movements, as exemplified by Figure 6.7. With different time frames of the same challenging set, body focus varies in similar scenes (hand movement while talking), with the difference being on the second timeframe having more subtle movement than the first.

6.5.2 Feature Importance Assessment

Although visual explanation provides some reasoning behind BIAS prediction, we also assess the relative importance of features (audio, face, and body data). We compare the effect of training sets on the perceived feature importance, using WASD and Easy/Hard group training, in Figure 6.8: WASD-BIAS, Hard-BIAS, and Easy-BIAS refer to BIAS training in WASD, Hard, and Easy data, respectively.

Relative Feature Importance. The main conclusions are: 1) For WASD-BIAS and Easy-BIAS, all the categories have similar audio importance (Figure 6.8a). The most divergent variation is Hard-BIAS, which gives less importance to audio in categories with faces available and higher importance in scenarios with occluded faces; 2) Both WASD-BIAS and Hard-BIAS give higher importance to body in surveillance scenarios relative to

Table 6.9: Effect of group training, in mAP, on the categories of WASD. Easy group training contains data of Optimal Conditions and Speech Impairment, while Hard group contains data of the remaining WASD categories.

Category	Train Set		
	Easy	**Hard**	**WASD**
Optimal Conditions	97.1	94.7	97.8
Speech Impairment	98.2	96.7	98.4
Face Occlusion	90.4	93.5	95.9
Human Voice Noise	77.3	83.4	85.6
Surveillance Settings	63.2	79.4	82.5

other categories (Figure 6.8b), given the various challenges in this setting. Furthermore, with Hard-BIAS, body is less important for the category with more close-up faces and less visible body. This trend is not seen in the other models given their different training conditions (*i.e.*, they have access to other categories with close-up faces and less visible body), leading to a well rounded body importance; and 3) Regarding face importance (Figure 6.8c), Hard-BIAS is the more evenly distributed model, meaning that face is overall important for ASD regardless of the category. Regarding WASD-BIAS (and Easy-BIAS to some extent), face is more relevant for scenarios with close-up faces and impaired audio, leading to increased face importance relative to the other categories.

Performance Influence. We explore the effect of group training and feature influence in model performance, in Table 6.9 and Figure 6.9, respectively. Overall, face is the most influential feature for BIAS in WASD, and the removal of body information has higher impact in surveillance settings, which highlights the body importance in these conditions (Figure 6.9a). When considering the impact of different features (audio, face, and body) in different training sets (Figure 6.9b), we observe that subsequent removal of available features translates into decreased performance, with training with all data (WASD) translating in more resilient results than training only on Easy or Hard data alone. This and the results in Table 6.9 highlight the importance of robust training data for improved resilience, particularly to avoid situations where training may lead to a false sense of face reliability (drop of performance from Hard-BIAS on Easy data when not using face). We also extend the analysis of feature influence in AVA and WASD training (Figure 6.9c). With subsequent removal of features, BIAS underperfoms the most on categories with audio impairment (HVN) and unreliable face access and audio quality (SS). Furthermore, surveillance settings is where body information is of utmost importance, based on the performance stability with using *Only Face*, *No Face* and *Only Body* (*i.e.* good face/body importance balance). Finally, BIAS training in AVA is not reliant on body, and has greater focus on audio and face features, heavily degrading in performance without these features. This shows that AVA data does not promote adequate feature combination, translating in models less resilient to varying data quality (*i.e.*, when audio and/or face quality are affected).

Table 6.10: ViT-GPT2 performance increase with finetune on ASD-Text dataset. Metric description: RL, ROUGE-L; M, METEOR; B1-4, BLEU1-4.

Model	RL	M	B1	B2	B3	B4
ViT-GPT2$_{Base}$	0.28	0.18	0.14	0.03	0.01	0
ViT-GPT2$_{ASD}$	0.61	0.58	0.61	0.50	0.37	0.31

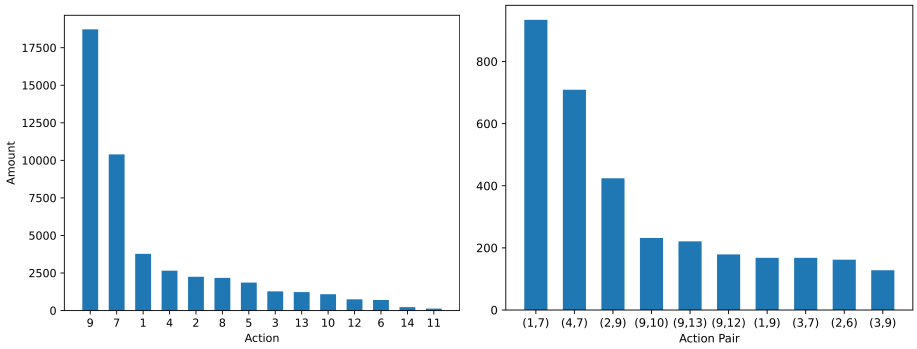


Figure 6.10: Distribution of ASD-Text dataset actions (left) and pair of actions more commonly associated (right). Action labels refer to their numbers in Table 6.1.

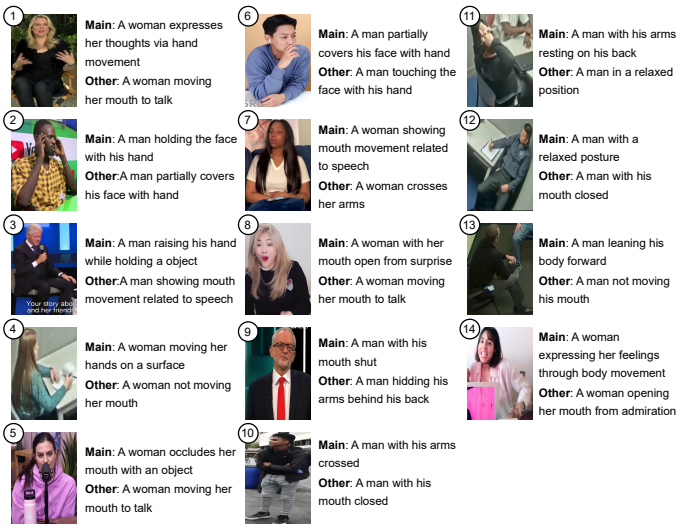


Figure 6.11: ASD-Text dataset finetuned ViT-GPT2 model predictions in representative images of the 14 considered ASD-related actions of Table 6.1. Each figure number relates to the respective ASD-action number. ViT-GPT2 predictions contain the main caption and a related one (other).

6.5.3 ASD-Text Dataset

To improve the interpretability of ASD-related scenarios, we complement the visual interpretability of attention heatmaps with text scene description, via caption generation of a ViT-GPT2 model. Given the absence of relevant data to train models for captions on ASD-related actions, we start by showing the relevance of the proposed ASD-Text dataset (Section 6.3.4), annotation distribution, and its importance towards a full interpretability setup.

ASD-Text Dataset Importance. We assess the performance of a pretrained (Base) and finetuned (ASD) ViT-GPT2 on ASD-Text dataset, on Table 6.10, where the results support the need for ASD-Text dataset given the improvements of a finetuned ViT-GPT2

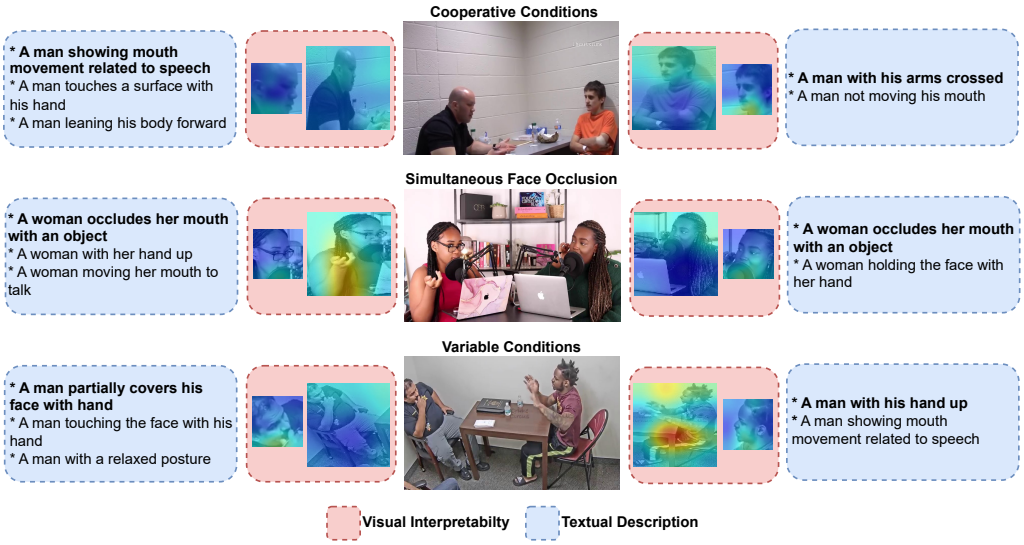


Figure 6.12: Combination of SE visual interpretability with ViT-GPT2 model subject description in various scenarios with varying ASD challenges. Bold descriptions are the main prediction of ViT-GPT2.

in predicting ASD-related captions. Based on the increase of all caption related metrics to reference levels [627], finetuned ViT-GPT2 can reliably describe ASD scenes in various challenging scenarios (WASD data). Furthermore, we also display ViT-GPT2 predictions on representative examples of the considered training actions (Table 6.1), in Figure 6.11. We present two captions per image based on the top 5 predictions of ViT-GPT2, with *main* being its first predicted caption and *other* being the first caption not related to the action of the main.

Annotation Distribution. Figure 6.10 shows the overall distribution of ASD-Text dataset actions. The majority of subjects have mouths closed or actively moving, relating to the speaking labels considered (talking *vs.* not talking). Regarding the actions more commonly paired, we have subjects talking with hand movement, either raised or on a surface, illustrating body movement associated with talking. More linked to non-talking subjects we have hands touching faces, without occluding mouth, as the most common pair association.

Full Interpretability Setup. We can also use ASD-Text dataset to create a complete interpretability setup that combines the visual information of BIAS (attention heatmap) with the text description of ViT-GPT2, as shown in Figure 6.12. In the three different scenarios, with varying challenges, visual information and text scene description represent the key characteristics for the decision behind accurate ASD.

6.6 Conclusion

In this paper we propose BIAS, a multi-modal approach for Active Speaker Detection (ASD) that singularly considers audio, face, and body-based information, which is state-of-the-art in challenging settings and has competitive results in more cooperative conditions. Furthermore, we propose a novel application of Squeeze-and-Excitation blocks

to assess ASD feature importance in different settings and provide visual interpretability, complementing them with text descriptions from a ViT-GPT2 model (finetuned in ASD-Text dataset) for a full interpretability setup. Our work highlights the importance of body inclusion for ASD in unconstrained/challenging conditions and serves as baseline for models to perform in wilder scenarios such as surveillance settings.

Chapter 7

How to Squeeze An Explanation Out of Your Model

Deep learning models are widely used nowadays for their reliability in performing various tasks. However, they do not typically provide the reasoning behind their decision, which is a significant drawback, particularly for more sensitive areas such as biometrics, security and healthcare. The most commonly used approaches to provide interpretability create visual attention heatmaps of regions of interest on an image based on models gradient backpropagation. Although this is a viable approach, current methods are targeted toward image settings and default/standard deep learning models, meaning that they require significant adaptations to work on video/multi-modal settings and custom architectures. This paper proposes an approach for interpretability that is model-agnostic, based on a novel use of the Squeeze and Excitation (SE) block that creates visual attention heatmaps. By including an SE block prior to the classification layer of any model, we are able to retrieve the most influential features via SE vector manipulation, one of the key components of the SE block. Our results show that this new SE-based interpretability can be applied to various models in image and video/multi-modal settings, namely biometrics of facial features with CelebA and behavioral biometrics using Active Speaker Detection datasets. Furthermore, our proposal does not compromise model performance toward the original task, and has competitive results with current interpretability approaches in state-of-the-art object datasets, highlighting its robustness to perform in varying data aside from the biometric context. The work presented in this chapter has been published in [628].

7.1 Introduction

Application scenarios for deep learning are diverse and steadily increasing everyday, ranging from robotic [629], person re-identification [630], security [631, 632], anomaly detection [597], biometrics [633], and even healthcare [634]. One of the key reasons behind this widespread relates to the ability of deep learning to perform various tasks, given the appropriate training data. The main drawback of using these approaches is mainly linked to their black-box settings, *i.e.* the lack of interpretability/explanations behind model decisions, which can be critical in more sensitive areas such as healthcare and security-related tasks. In particular, for biometric applications, explainability related to facial attributes is crucial for face verification systems, while assessment of behavioral biometrics, such as speaking mannerisms and non-verbal cues (body language), is also highly influential.
There has been an increased interest in interpretability/explainable approaches for deep learning in recent years. The most predominant technique relates to visual attention maps via backpropagation of model gradients [598] where the areas

of more "interest" for the model are highlighted, providing an explicit idea of where the model is mostly sourcing outputs. Other approaches not linked to visual interpretability are less common, although there are various works using game theoretic approaches [635] to explain the output of machine learning models, which are useful to assess models limitations and the most influential features [480]. Although this topic has gained interest over the years, most works mainly focus on gradient-based approaches in image settings and for standard deep learning models. This means that other settings (such as video/multi-modal) and custom models (variations of available implementations) are less explored, and currently available interpretability approaches do not work in such conditions or require heavy adaptations.

Given these limitations, we propose a novel use of the SE block [2] to create a model-agnostic interpretability approach (Figure 7.1) that works regardless of the model, dataset and setting (image/video) used. SE block was initially proposed to account for the interdependencies between different feature channels to improve model performance. We adapted it to assess the most influential features to create visual heatmaps, similar to current interpretability approaches, via SE vector manipulation. Our proposal simply requires the inclusion of an SE block prior to the classification layer of deep learning models to be able to assess models attention. We explore the viability of SE-interpretability in state-of-the-art object datasets and in the biometric context, namely facial attributes and speaking behavioral features using CelebA [636] and ASD data, respectively. The results show that the inclusion of SE blocks is a viable approach for visual interpretability, competitive with existing state-of-the-art approaches, and can be applied to different standard and custom models, using various datasets and settings, without degrading their performance on the original task.

The remaining of the paper is structured as follows. In Section 2, we discuss the related work. In Section 3, we describe SE-related concepts and present our novel use of SE for interpretability. In Section 4 we report the experiments and discuss the results obtained. Finally, Section 5 concludes the paper.

7.2 Related Work

Model Interpretability. For visual interpretability, we can group methods into two main categories: (i) *gradient-based* [599, 600, 601, 602], where gradients of each layer are computed through backpropagation; and (ii) *attribution propagation* [604, 605, 542, 606], where there is a recursive decomposition of the contribution of the layers all the way to the input of the model. Other examples of visual interpretable approaches are *Saliency-based methods* [607, 608], which multiply the original image by the obtained saliency map to draw the focused region, *Excitation Backprop* [609], which calculate a marginal winning probability by summing the values across channels, and *Perturbation methods* [610, 611] that perturb the input and observe the changes made to the outputs. The increasing interest in Vision Transformers led to the *Transformer-based interpretability*, inspired by standard techniques, such as gradient-based [637, 638, 639, 640],

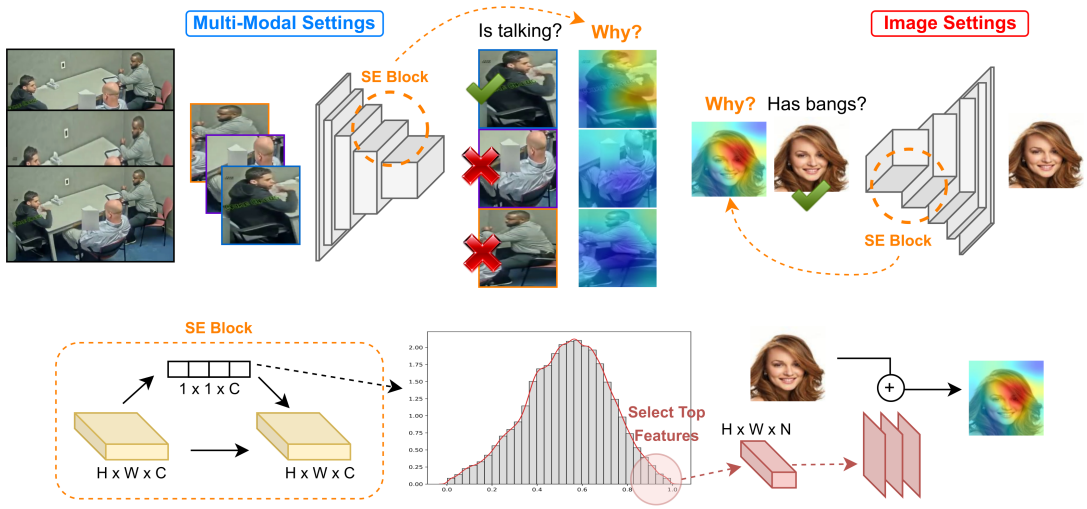


Figure 7.1: Overview of the inclusion of SE blocks for model-agnostic interpretability. Two different settings are depicted (multi-modal, on the left, and image, on the right) where SE block is similarly included in feature extractors to output attention heatmaps of models predictions. Heatmaps are created using channel features of the respective top 10% SE vector values, via channel interpolation and combination with the original image. Audio encoding of the multi-modal setting (ASD) is not displayed for simplicity.

attribution [612, 641], and redundancy reduction [642]. Although most works explore visual interpretability in object classification datasets, its use in biometric-related topics such as face [613, 614], body [615], and PAR [616, 548] data is not unprecedented.

Video and Multi-Modal Interpretability. The inclusion of GradCAM [598] and its variants is not commonly explored in video or multi-modal settings, aside from a few examples [643, 644]. Using GradCAM in video settings requires the video to be converted into a sequence of frames, ignoring the spatiotemporal correlations [645]. Furthermore, using the standard interpretability approaches in custom architectures, such as ASD context (multi-modal settings), is not extensively explored or smoothly done using the available implementations. Other approaches [646, 647] involve using a separate model/process to explain the decision of the original model. Our SE-based visual interpretation does not require additional computational cost or external methods to create attention heatmaps, is applicable for image and video/multi-modal settings, and can be included in standard and custom architectures.

7.3 Squeeze-and-Excitation for Interpretability

7.3.1 SE Definition and Concepts

The original objective of SE blocks [2] was to take into account the interdependencies between different feature channels to improve models performance, via three key stages: 1) Squeeze; 2) Excitation; and 3) Scale and Combine. Figure 7.2 presents the way SE blocks operate.

1) Squeeze Phase: The goal of this phase is for the SE block to get a global understanding of each channel by squeezing them into a single numeric value, via global average pooling over the spatial dimensions (height H and width W) of each channel. Formally, we can

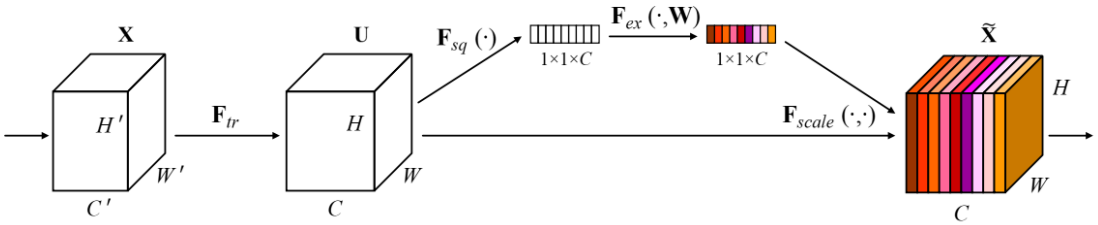


Figure 7.2: Overview of a Squeeze-and-Excitation block, where input U is squeezed, for each channel (u_c, in (1)), via average pooling through spatial dimensions $H \times W$ to output a vector of weight importance (s, in (2)), used to highlight the most important feature via channel-wise multiplication of weight importance vector and respective channel of the inputted feature map. Retrieved from the original paper [2].

define $z \in \mathbb{R}^C$, which is generated by shrinking input $U \in \mathbb{R}^{H \times W \times C}$ through its spatial dimensions $H \times W$, such that the *c-th* element of z is calculated by:

$$z_c = \frac{1}{H \times W} \sum_{H}^{i=1} \sum_{W}^{j=1} u_c(i,j). \quad (7.1)$$

2) Excitation Phase: This phase aims to fully capture channel-wise dependencies of z, using two FC, ReLU and Sigmoid layers, outputting vector s, with the same shape of z, representing the learned importance weights for each channel. Formally, $\mathbf{s}$ is defined by:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})), \quad (7.2)$$

where σ refers to the sigmoid activation, δ is the ReLU fuction and $\mathbf{W}$ refer to the FC layers.

3) Scale and Combine: Finally, the weights of s are incorporated into the input channels to highlight the most important features, via rescaling U with s using channel-wise multiplication between each channel of s and the respective channel of the feature map $\mathbf{u}$, outputting $X \in \mathbb{R}^{H \times W \times C}$:

$$\mathbf{x}_c = s_c \mathbf{u}_c. \quad (7.3)$$

7.3.2 SE Visual Interpretability

We use SE to obtain the features (channels) perceived as highly influential (*i.e.*, high value in SE vector, $\mathbf{s}$ in (2)) for models to perform in different settings. However, the perception of *high* is relative to the task considered, so we can not use a default/hard-coded value to define what is high or low importance. As such, given that the values of the SE vector in our experiments follow a Normal distribution (Figure 7.3), we use its formula to obtain the top 10% values of the SE vector (and, consequently, the top 10% channels) for visual interpretability purposes. We obtain the normal distributions by grouping the SE values of all models for each dataset, by normalization with mean to 0. Finally, we conjugate the selected channels using bicubic interpolation on top of the original image to provide a heatmap of the most important regions for model decision.

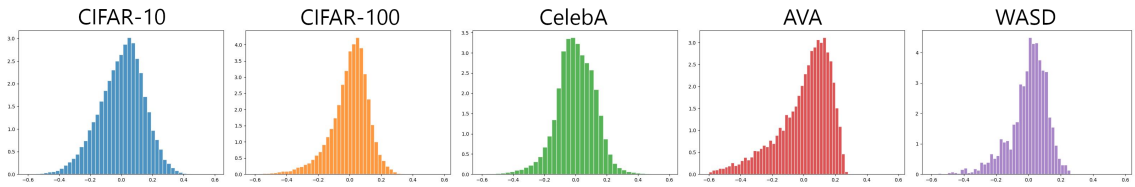


Figure 7.3: SE value distribution of all considered models for the datasets CIFAR-10, CIFAR-100, CelebA, AVA, and WASD (from left to right). SE values are normalized with mean to 0.

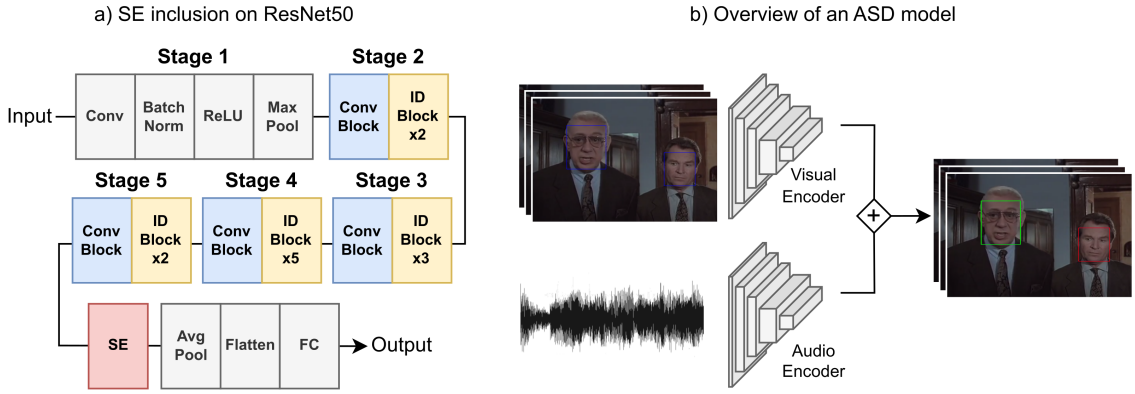


Figure 7.4: Inclusion of SE block in ResNet50 (left), and overview of how standard ASD models perform (right).

Motivation. The novelty of our approach is the inclusion of a single SE block at the end of different models, in particular before average pooling (left side of Figure 7.4) to provide visual interpretability similar to those obtained from GradCAM-like approaches. The key motivation is that inherently SE blocks already aim to assess the most influential channels to perform a given task, so we make use of the SE vector ($\mathbf{s}$ in (2)) to select the features of the most influential channels ($\tilde{\mathbf{X}}$, from Figure 2) to create model attention heatmaps. Given the simplicity of our proposal, SE blocks can provide visual interpretability to any model (standard or custom) in various settings (image or video-based).

SE Interpretability Application. SE interpretability is applicable to image-based data, as most interpretability-based approaches are able to perform. Additionally, it is also applicable in video/multi-modal settings, as we explore in our experiments using ASD as a proof-of-concept to assess behavioral biometrics. ASD models are commonly custom made (not standard ResNets) and use facial and audio data to predict who is talking, given a video input, which makes this area a prime candidate to assess SE interpretability robustness (right side of Figure 7.4).

7.3.3 Implementation Details

Image models are trained using an SGD optimizer with a learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.0005, during 50 epochs starting from pretrained weights on ImageNet. All inputs are resized to 128×128, with the exception of Inceptionv3, which uses 299×299 inputs due to its architecture constraints. Multi-modal models are trained with an Adam optimizer, with an initial learning rate of 10^{-4}, decreasing 5% for each epoch. All visual data is reshaped into 112×112, audio data is represented by

13-dimensional MFCC, and both visual and audio features have an encoding dimension of 128. For visual augmentation, we perform random flip, rotate and crop. For audio augmentation, we use negative audio sampling [1].

7.4 Experiments

7.4.1 Datasets, Methods, Models, and Evaluation Metrics

Datasets. For image-based settings, we consider 3 datasets: CIFAR-10, CIFAR-100, and CelebA. CIFAR-10 dataset [648] is a subset of the Tiny Images dataset [649] and consists of 60 000 32×32 color images, labeled into 10 mutually exclusive classes, each containing 6 000 images, with 5 000 training and 1 000 testing images per class. CIFAR-100 dataset [648] contains the same data as CIFAR-10, except it has 100 classes with 600 images each, divided into 500 training images and 100 testing images per class. CelebA dataset [636] contains 202 599 face images with size 178×218 from 10 177 celebrities, each annotated with 40 binary labels indicating facial attributes. In our experiments, we use CelebA for binary classification (only train models to predict one attribute).
Regarding multi-modal settings, we consider 2 ASD datasets: AVA-ActiveSpeaker [59], and WASD [3]. The AVA-ActiveSpeaker dataset [59] is an audio-visual active speaker dataset from Hollywood movies, where typically only train and validation sets are used for experiments: 120 for training and 33 for validation, ranging from 1 to 10 seconds. WASD [3] compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*, containing 30 hours of labeled data, 164 videos, varying FPS, averaging 28-second duration, and balanced demographics.
Interpretability Methods. We compare the qualitative performance of our approach with GradCAM [598], GradCAM++ [650], EigenGradCAM [651], and FullGradCAM [600] from publicly available implementations [652].
Models. For image-based settings, the considered models are MobileNetv2, ResNet18, ResNet50 and InceptionV3, while for multi-modal conditions we select WASD Baseline [3] and Light-ASD [653], given their end-to-end approach. Relative to Light-ASD, which only takes face and audio as inputs, WASD Baseline also considers body data as input.
Evaluation Metrics. For CIFAR-10, CIFAR-100, and CelebA we use accuracy, and for AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [59] that computes mAP. For quantitative interpretability evaluation we use the Deletion and Insertion metric [654].

7.4.2 SE Inclusion on Model Performance

We start by assessing the effect of incorporating SE blocks into different architectures for image settings, in Table 7.1. The results show that using SE blocks can be applied in object and biometrics data, while not affecting the performance of models, regardless of their architectures and datasets, and without significantly increasing the number of model pa-

Table 7.1: SE inclusion performance (Acc) influence on CIFAR-10, CIFAR-100, and CelebA. Results on CelebA refer to gender attribute prediction. Par(M) refers to the number of parameters in millions.

Model	Par(M)	SE	CIFAR-10	CIFAR-100	CelebA
MobileNetv2	2.13	×	94.59	77.28	97.20
		✓	94.29	76.63	97.44
ResNet18	10.66	×	94.49	76.75	97.45
		✓	94.58	76.98	97.64
ResNet50	22.44	×	96.25	80.73	97.04
		✓	95.84	81.31	97.44
Inceptionv3	23.97	×	97.07	84.25	97.38
		✓	96.92	84.04	97.51

Table 7.2: SE inclusion performance (mAP) influence on AVA-ActiveSpeaker and WASD. AVA refers to AVA-ActiveSpeaker and Par(M) refers to the number of parameters in millions.

Model	Visual Encoder	Par(M)	SE	AVA	WASD
Light-ASD [653]	Conv 2D-1D	1.0	×	93.6	93.7
			✓	93.7	93.9
Baseline [3]	RN18 2D-3D	31.6	×	92.3	94.0
			✓	92.4	94.1

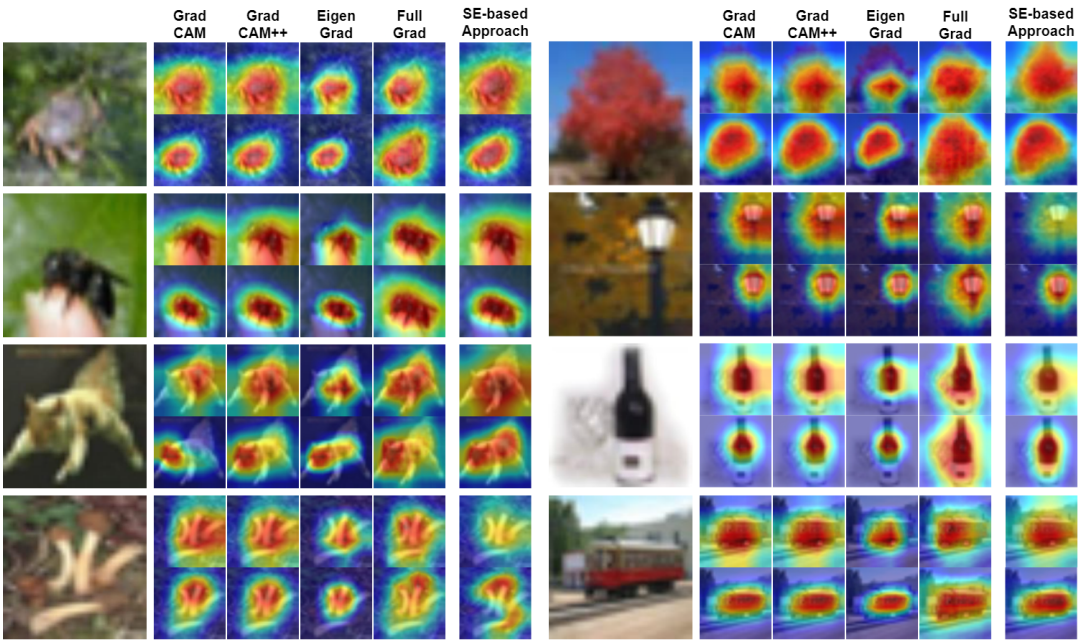


Figure 7.5: SE Interpretability (last column per original image) relative to other standard visual interpretability approaches: GradCAM, GradCAM++, EigenGrad, and FullGrad. For each original image the top row relates to interpretability of ResNet50 and the bottom to InceptionV3.

rameters. We also extend this evaluation to the multi-modal context, in particular ASD settings to assess behavioral biometrics, where audio and visual inputs are used for model predictions, in Table 7.2. The results confirm that SE blocks do not limit model performance, even in more challenging and variable conditions such as ASD data, showing its viability in behavioral biometrics related to speaking.

Table 7.3: ResNet18 performance (Acc) on different attributes of CelebA.

Attribute	Arched Eyebrows	Bald	Bangs	Bushy Eyebrows	Double Chin	Eyeglasses	Beard	Earrings	Hat
Acc	83.25	98.65	95.75	92.39	96.32	99.63	95.49	89.88	98.99

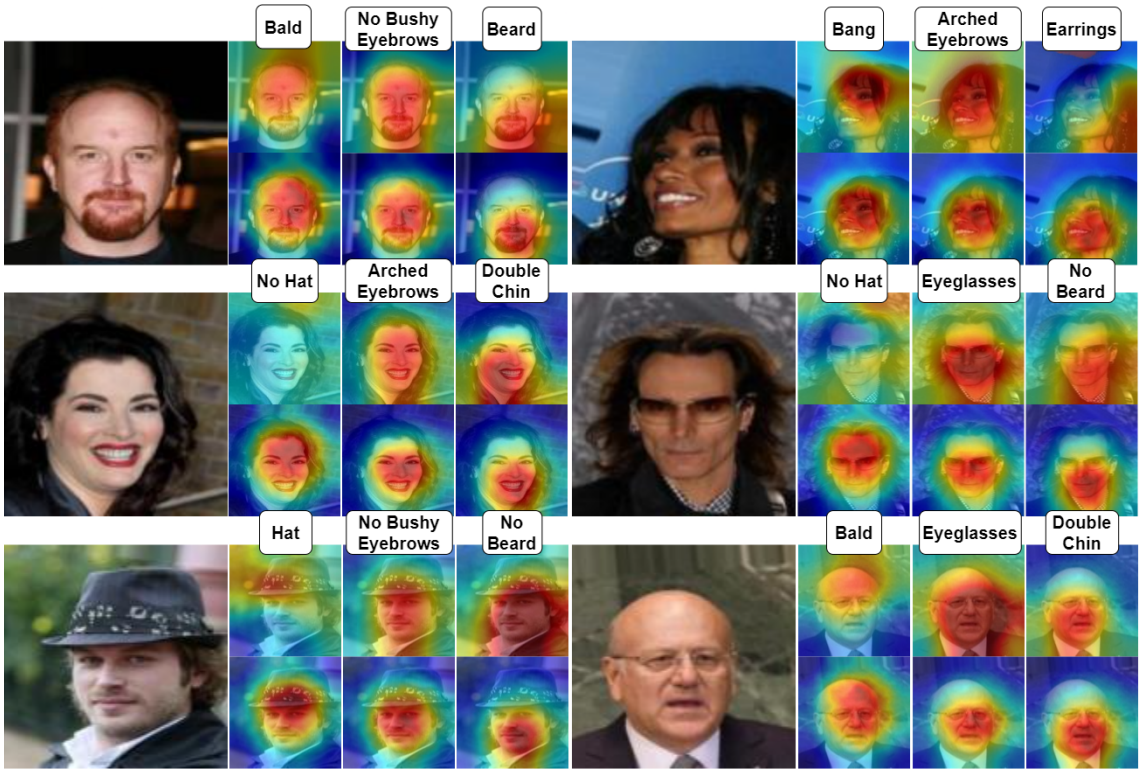


Figure 7.6: SE Interpretability of ResNet18 for different CelebA attributes. Each column after the original image refers to attributes more commonly located to specific parts of the face: top, middle, and bottom, respectively. For each original image the top row relates to interpretability of our approach (SE-based) and the bottom to FullGradCam.

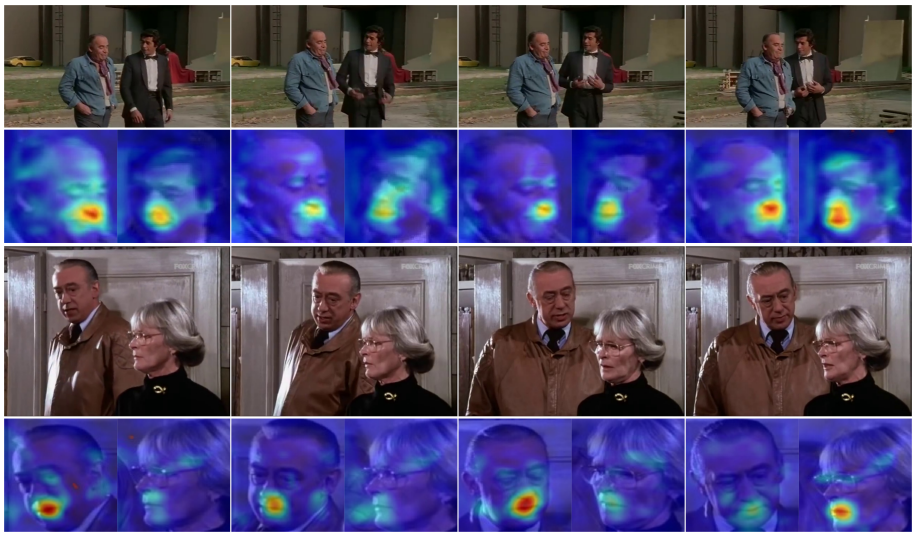


Figure 7.7: Attention of Light-ASD for facial cues, via SE blocks, in AVA-ActiveSpeaker data.

7.4.3 Qualitative Performance of SE Interpretability

Given that SE blocks do not influence model performance, regardless of architectures or training/testing settings, we assess the interpretability of models via SE blocks for CIFAR-100 in Figure 7.5. We use an object dataset to compare the viability of SE-interpretability relative to standard visual interpretability approaches (GradCAM variants), given the varying features available to classify in this data. The results show that our approach is able to output reliable and interpretable results for various settings: animals, plants, objects, with closed-up images, and varying positions of objects. Additionally, SE approach is applicable regardless of the architecture (ResNet *vs.* Inception), although Inceptionv3 tends to have more focused attention, which is due to the input size requirements of this model (299x299 *vs.* 128x128 of other models). We also assessed the influence of model attention in the biometric context, based on different attribute training using CelebA data, in Figure 7.6 using our SE-based interpretability and FullGradCAM. The results show that our SE-based approach can be used to highlight the varying model attention based on the different attribute training: 1) hat presence, baldness or bangs relate to focus on top locations of the face; 2) eyeglasses and eyebrows lead to overall face attention; and 3) beard, accessories, and chin features relate to bottom locations of the face. Relative to our approach, FullGradCAM also highlights similar facial regions given different attributes, but with greater focus to the face (and less to the background). For reference, the performance of ResNet18 for each of the attributes assessed is shown in Table 7.3.

7.4.4 Multi-Modal Feature Importance via SE

The relevance of SE blocks for interpretability also extends to other settings such as video/multi-modal, where we use ASD data in our experiments to assess the behavioral biometric context (facial attributes/cues analysis). Current ASD models take as input video facial data and associated audio to predict who are the active speakers in a scene. We start by exploring Light-ASD attention in different scenarios in Figure 7.7. The results show that the model focuses mainly on facial cues when people are talking, given that these are the most distinctive features, with greater attention to the person talking relative to other people in the scene, as seen in the second scenario.

We also complement our experiments with WASD Baseline, which takes body data as an additional input, distinctive from other state-of-the-art ASD approaches, allowing for behavioral biometric assessment via body cue analysis (non-verbal speaking cues). We consider different scenarios of WASD to assess the relative importance of face and body attention heatmaps, in Figure 7.8, where each scenario contains the initial setting (bigger image) and the next three frames for body and face movements (and their respective heatmaps). The results show greater attention to the body when the scenario contains challenging features, even if the face is available or partially accessible. In particular, with audio quality subpar (video playing, first row) or limited face availability (hat, second row), body movement importance increases for ASD. Furthermore, when limiting the access to facial cues (last two scenarios), the relative importance of the body further



Figure 7.8: Body attention via SE blocks of WASD Baseline in challenging sets of WASD.

Table 7.4: Comparative evaluation in terms of Deletion (lower is better) and Insertion (higher is better) scores of a ResNet50 on CIFAR-10 and CIFAR-100 datasets.

Approach	**CIFAR-10**		**CIFAR-100**	
	Deletion	Insertion	Deletion	Insertion
GradCAM	0.31	0.93	0.17	0.77
GradCAM++	0.30	0.93	0.15	0.78
EigenGradCAM	0.29	0.93	0.14	0.78
FullGradCAM	0.29	0.94	0.13	0.79
SE Interpretability	0.34	0.91	0.14	0.76

increases, translated by visual attention to hand movements (in both examples).

In these experiments, heatmaps for model attention tend to be consistent across consecutive frames and do not require big deviations (not talking and starts talking) to highlight model attention, although heatmap consistency across sequential frames may vary depending on the ASD context (*i.e.* more cooperative ASD data leads to greater heatmap consistency). Given the results, our SE approach is a viable strategy to assess model attention in multi-modal settings, in particular for behavioral biometrics, regardless of the model architecture and dataset.

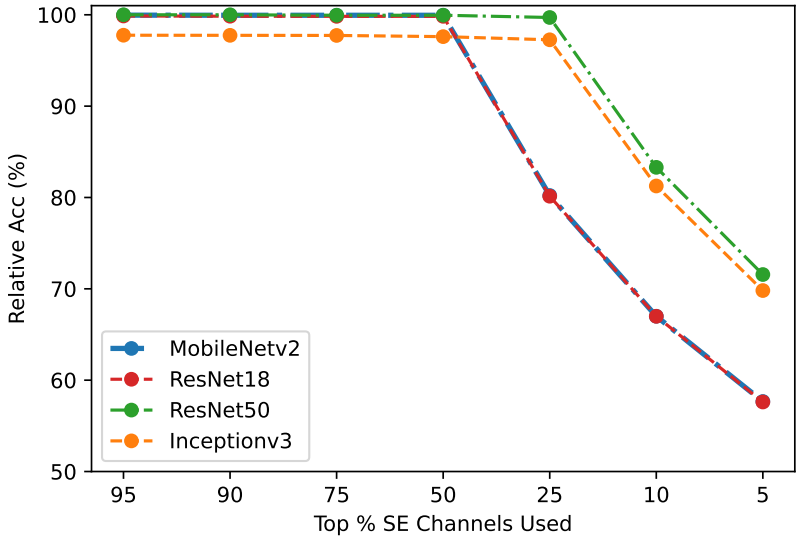


Figure 7.9: Accuracy with top % of SE channels used in CIFAR-100, relative to using all available feature channels.

7.4.5 Quantitative Performance of SE Interpretability

Deletion and Insertion. To objectively compare our approach with standard interpretability methods, we use the Deletion and Insertion metric [654] for ResNet50 in Table 7.4. *Deletion* assesses the decrease in the probability of the predicted class as the most relevant pixels are removed (given the heatmaps), while *Insertion* takes a complementary approach, *i.e.* it measures the increase in probability as more and more pixels are introduced, with higher Area Under Curve indicative of a better explanation. The results show that our SE-based approach is competitive with standard interpretability approaches for CIFAR-10 and 100, with the advantage of being applicable to any model in various settings (image and video/multi-modal), making it a suitable option for model interpretability.

SE Channel Importance Variance. We also explore the influence in performance when using feature channels corresponding to the top % SE vector values, in Figure 7.9. We select CIFAR-100 for this experiment given that it is the most challenging data in our experiments, based on the reported results in Table 7.1 and their inherent challenges, as discussed in Section 7.4.3. Although only the top percentage of SE channels are relevant to assess model attention, the decrease of feature channels used leads to performance decrease, with a greater emphasis using the top 25% channels or less, for all considered models. Noticeably, ResNet50 and Inceptionv3 are more resilient to using fewer feature channels, given their performance with the top 25% channels relative to MobileNetv2 and ResNet18, which is linked to model robustness (increased number of parameters).

7.5 Conclusion

In this paper we propose a novel application of SE blocks to assess model attention, with competitive qualitative results relative to common variants of GradCAM. The effective-

ness of the proposed approach was proven in the biometrics context, in particular by assessing facial attributes and speaking features, via extensive testing with well known datasets both for images and videos, namely in the context of ASD. The interpretability provided by the SE blocks derives from its inherent intention to highlight the most discriminative features via SE vector, and can be applied to standard and custom architectures, in image or video/multi-modal settings, without compromising model performance. Future work includes exploring the effect of including SE blocks at different stages of a model to assess its attention to high (global) and low-level (local) features, and extend the experiments to assess whether the most salient zones are coherent with the predicted class using segmented face regions.

Chapter 8

ASDnB: Merging Face with Body Cues For Robust Active Speaker Detection

State-of-the-art Active Speaker Detection (ASD) approaches mainly use audio and facial features as input. However, the main hypothesis in this paper is that body dynamics is also highly correlated to "speaking" (and "listening") actions and should be particularly useful in *wild* conditions (*e.g.*, surveillance settings), where face cannot be reliably accessed. We propose ASDnB, a model that singularly integrates face with body information by merging the inputs at different steps of feature extraction. Our approach splits 3D convolution into 2D and 1D to reduce computation cost without loss of performance, and is trained with adaptive weight feature importance for improved complement of face with body data. Our experiments show that ASDnB achieves state-of-the-art results in the benchmark dataset (AVA-ActiveSpeaker), in the challenging data of WASD, and in cross-domain settings using Columbia. This way, ASDnB can perform in multiple settings, which is positively regarded as a strong baseline for robust ASD models (code available at `https://github.com/Tiago-Roxo/ASDnB`.). The work presented in this chapter is under review [655].

8.1 Introduction

Active Speaker Detection (ASD) aims to identify, from a set of potential candidates, active speakers on a given visual scene [59], with application in several topics such as speaker diarization [565, 566, 567], human-robot interaction [656, 657], automatic video editing [658, 659], and speaker tracking [568, 569]. State-of-the-art ASD models typically perform at the video frame level using face data and sound information. Using only facial cues as visual input is a viable strategy due to the correlation of mouth movement and speaking activity, but this approach is only reliable in cooperative and controlled settings. This widely used strategy is motivated by the benchmark ASD dataset AVA-ActiveSpeaker, composed of movies with good audio and face quality.

Another approach that has not been widely explored for ASD is the use of body information. When people speak or listen they typically use other forms of non-verbal behaviors such as head nodding, hand and body movements, which are not considered by current ASD models. This information increases in importance when face can not be reliably accessed (*e.g.*, face occlusion) or in wilder conditions such as surveillance-like settings [4], as shown in Figure 8.1. As such, complementing body information with facial cues could improve model robustness to perform in cooperative conditions as well as more challenging settings.

This paper proposes a model that brings Active Speaker Detection and Body (ASDnB)

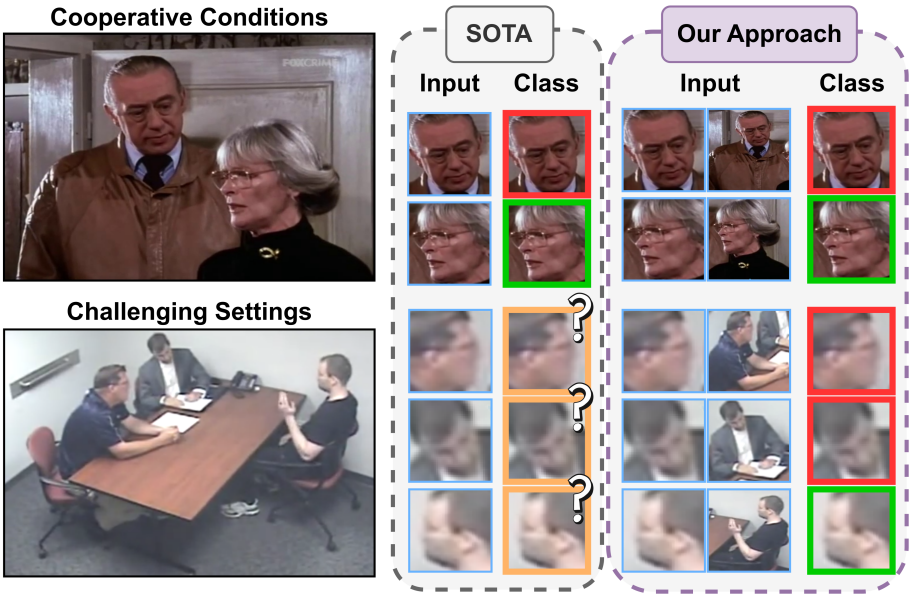


Figure 8.1: State-of-the-art ASD models solely rely on facial cues as visual input to perform. This approach is only reliable with cooperative (subjects) conditions, increasing uncertainty in model prediction in more challenging settings. Our approach aims to complement face with body cues to create more robust models that are able to perform in cooperative and unconstrained scenarios.

together. In particular, ASDnB has the novelty of efficiently combining body with face data using a single visual encoder, merging them at different steps of the extraction, and outputting combined visual features for robust ASD. We modified the visual encoder by splitting the 3D convolution into 2D and 1D to reduce computation cost without loss of performance, and we train ASDnB using adaptive weight feature importance, which results in improved visual encoder extraction and feature complement. Finally, we include temporal modeling in the classifier of ASDnB using BGRU layers to maintain the temporal notion for speaker label prediction. Our experiments show that ASDnB achieves state-of-the-art results in the benchmark ASD dataset AVA-ActiveSpeaker, as well as in challenging data with degraded audio and face data quality (WASD [3]) and cross-domain settings (Columbia [570]), making it a baseline for robust ASD models. To summarize, the main contributions are:

- We announce the first effective combination of body and face data for visual input in ASD, which is a novel approach to create robust models to perform in more challenging settings;
- We propose a intra visual encoder combination of dual inputs (face and body) and training with adaptive weight feature importance to effectively combine relevant features for robust ASD;
- Ablation studies, experimental evaluation, and performance analysis demonstrate ASDnB state-of-the-art results in the benchmark dataset, AVA-ActiveSpeaker, in the challenging data of WASD, and in cross-domain settings using Columbia.

8.2 Related Work

ASD Context. Active Speaker Detection is the task to determine the talking speaker from a set of admissible candidates. The benchmark dataset of this area is AVA-ActiveSpeaker [59], which is based on Hollywood videos totalling almost 38 hours, with demographic diversity and FPS variation, with applications in other areas [597]. Several other datasets [72, 571, 60] were announced since, guaranteeing face access and good audio quality, similar to AVA-ActiveSpeaker, which is not an accurate representation of wild conditions [59]. For ASD in more challenging data, WASD [3] has been recently proposed containing different categories, with varying audio and face quality, ranging from cooperative conditions to surveillance settings. Based on the available data for ASD, current state-of-the-art models heavily rely on face and audio data, combining them using 3D architectures [61], hybrid 2D-3D models [62], and attention mechanisms [66, 67, 68]. Earlier works are based on a two-step process, where the first focuses on audio with face combination and the second on multi-speaker analysis [69, 70, 71, 72], while recently end-to-end models have emerged [1, 75, 74, 76, 4]. Contrary to existing works, where face is the only visual input for ASD, ASDnB is the first to effectively combine face with body data for robust ASD.

Model Enhancement. Strategies to improve ASD models are typically based on improved feature extraction and combination. Works focus on temporal speech refinement [69], inter-speaker and audio-visual relations [70, 71], using GCN [73] to improve speaker relation representation [72, 74, 75], long-term temporal context with audio-visual synchronization [1], using a reference speech to improve ASD [77], and changing encoder architectures [76]. Despite the different strategies, combining audio with visual features is usually done post encoding using cross-attention approaches [1, 77] and complemented by temporal modeling. ASDnB is the first ASD model to combine different visual features (face and body), intra encoding, making it more robust to perform in challenging data where face can not be easily accessed.

Using Body Information. Current ASD state-of-the-art models rely on facial cues for visual input given the subject cooperation (face access guaranteed) of the mainstream datasets. However, this is not a viable approach in wilder conditions (such as surveillance settings), where face is not reliably accessed. As such, one potential strategy to improve ASD model robustness is using body information, as explored in other areas. PAR datasets [506, 507, 508] are examples of these scenarios, containing person cropped images from surveillance settings, used to identify various attributes under challenging covariates. Works in this area focus on different strategies ranging from different architecture combination [548, 470, 483], attention-based approaches [481, 484], assessing model limitations [480], and attribute relation importance [475, 476, 478]. In ASDnB, we propose a modification of standard ASD visual encoders, where face and body features are combined at different steps of the extraction, intra encoder, outputting combined visual features.

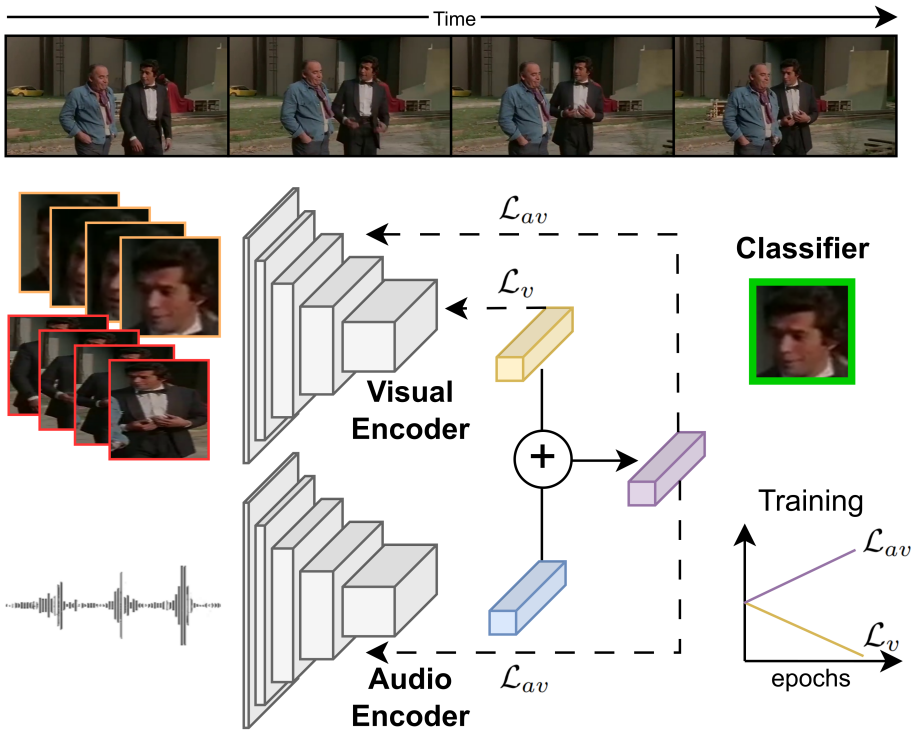


Figure 8.2: Overview of the ASDnB architecture. Body and face data are fed into the visual encoder, allowing intra encoder feature fusion and complement, while audio is processed through its respective encoder. Audio and visual features are combined to predict subject speaking label, using an adaptive weighted loss for combined and visual features ($\mathcal{L}_{av}$ and $\mathcal{L}_v$, respectively).

8.3 ASDnB

We propose ASDnB, an model that, for the first time, effectively combines face with body data to perform ASD in cooperative and challenging conditions. We combine dual visual inputs using a single encoder by complementing face and body features at different steps of extraction, and train with adaptive weight to improve feature extraction and combination. The overall architecture of ASDnB is displayed in Figure 8.2, with details of each model component in the following subsections.

8.3.1 Visual Encoder

Selecting Visual Encoding Approach. Several strategies for visual encoding in ASD are based on 3D convolutional neural networks, given their effectiveness in extracting spatiotemporal information of face sequences [70, 75]. These approaches are typically computationally expensive with increased number of parameters, which made other works explore the use of a 3D convolution prior to inputting to a 2D ResNet (typically 18), followed by visual temporal convolution networks [1, 77]. The key takeaway from the state-of-the-art is that reducing the visual inputs to a 2D context leads to good ASD performance, which can be further extended by splitting the 3D convolution into 2D and 1D to extract the spatial and temporal information, respectively [76]. We select this approach for our model since it combines the key strategies of previous ASD works while significantly reducing the number of parameters and computational cost, and without loss of performance.

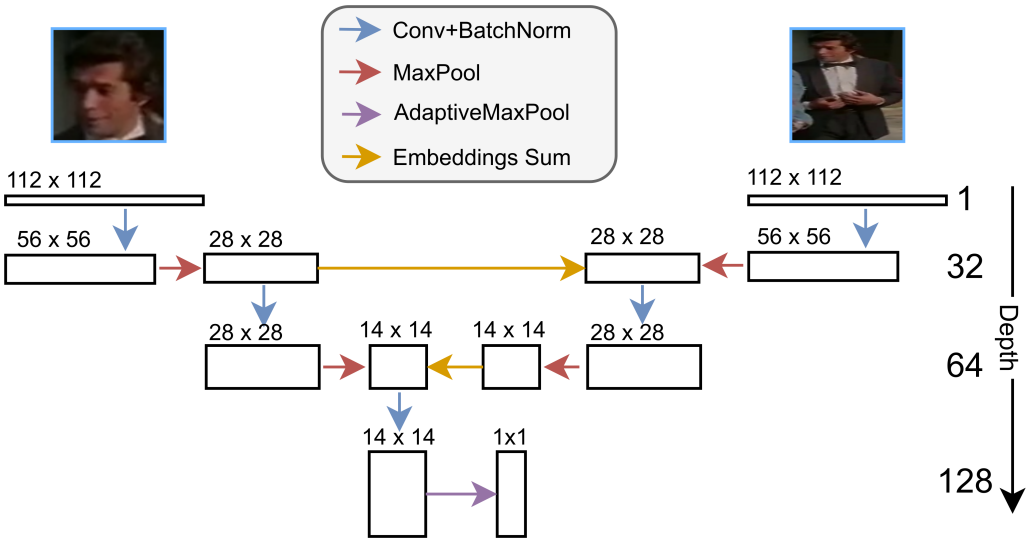


Figure 8.3: Overview of the flow of face and body combination in ASDnB visual encoder. The first convolution for both inputs downsamples via stride.

Combining Face and Body. Contrary to previous works, which only take face as input to visual encoders, we also consider body given its importance to complement facial cues for ASD in more challenging settings (*e.g.*, surveillance conditions). One possible approach is adding two visual encoders to extract face and body features, followed by combining them with audio prior to model classification. However, this does not force the model to consider face and body information in conjunction on feature extraction, resulting in ignoring information complement from the two sources, leading to subpar performances, namely in cooperative settings where face data is reliable [4]. As such, our motivation is to combine face and body data intra visual encoder to output combined visual feature, using an approach inspired by UNet [660] where face and body features are combined at different steps of the extraction, as shown in Figure 8.3. We combine face features into body at early extraction steps to make body data as a complement to facial cues since these are the main features for ASD, while recombining body data to face feature at latter stage to reinforce data conjunction. Finally, to ensure spatiotemporal feature extraction abundance at different receptive fields, each convolutional block contains two paths, one with kernel size of 3 and another of 5, given its superiority to other kernel combinations [76].

8.3.2 Audio Encoder

For audio encoding, we adapt the audio signal to serve as input to a 2D encoder by generating MFCC, with a sampling rate of 16 kHz, analysis window of 0.025 ms, step between successive windows of 0.010, and a audio representation of 13 cepstrums. As audio encoder we use ResNet34 with SE blocks in its layers (SE-ResNet34), outputting a 128-dimensional audio embedding for subsequent visual and audio conjunction.

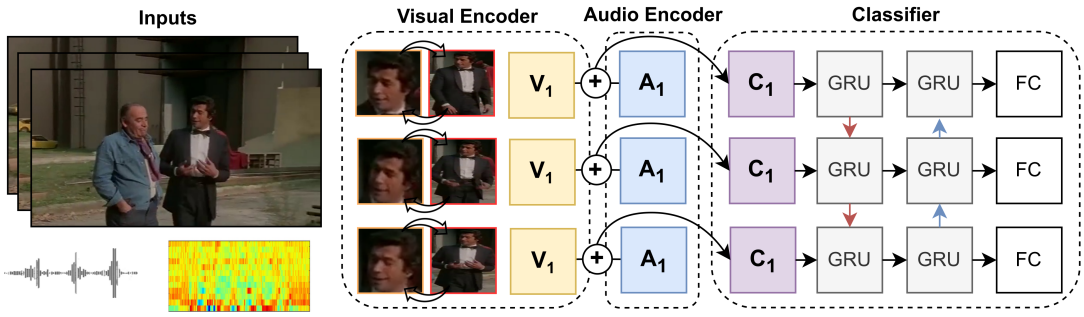


Figure 8.4: Bidirectional GRUs of visual and audio combination in ASDnB classifier, before inputting to FC layers for speaker classification.

8.3.3 Temporal Modeling in Classifier

To improve ASD model performance we apply temporal modeling to the combined multi-modal features from audio and visual encoders. The key motivation is to provide a temporal notion to the model to predict if a subject is talking given that speaking is a continuous action in time, *i.e.* if one subject is talking in a given frame it is more likely to be talking in sequential frames, with a similar logic applying to a non-talking subject. Our approach for ASDnB is shown in Figure 8.4, where the combined multi-modal features are obtained by summing visual and audio features, followed by a BGRU, before passing to a FC layer to predict if the candidate is talking.

8.3.4 Loss Function

Selecting Loss Strategy. Existing ASD approaches are typically based on audio and face, with losses tending to rely on the conjunction of both data to assess if a subject is talking. However, recent works [1, 76] explore the inclusion of visual and audio features as sole inputs for losses to complement the cross-modal interaction, leading to improved feature extraction and model performance. In our loss, we also consider visual features importance into it, given that cooperative ASD settings (like AVA-ActiveSpeaker) tend to benefit from assessing facial cues to predict subject talking (*i.e.*, mouth movement heavily relates to talking).

Feature Importance and Combination. Unlike previous works, we include the notion of gradually increment the importance of combined features (audio with visual) with a relative decrease of visual feature importance throughout training. The key motivation is that, although visual features are important to improve feature combination, its relevance is more crucial in earlier stages of training, with the end goal of ASD models being to assess if a subject is talking via both audio and visual, and not solely based on facial cues. This is particularly important for ASD in more challenging data (*e.g.*, surveillance settings), where the importance of visual features as sole input is not a reliable approach. Furthermore, since our visual encoder has to combine face and body features to output visual embeddings, we want the training to focus early on improving the visual encoder and later at combining audio with visual features.

Formally, we define ASD as a frame-level classification, where the predicted label se-

quence is compared with the ground truth via Cross-Entropy:

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T}(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \quad (8.1)$$

where T refers to the number of video frames, p_i and y_i are the predicted and ground truth label for the i^{th} frame, respectively. Finally, the complete loss function is expressed by:

$$\mathcal{L}_{ASDnB} = \alpha\, \mathcal{L}_{av} + (1 - \alpha)\, \mathcal{L}_{v}, \quad (8.2)$$

where $\mathcal{L}_{av}$ and $\mathcal{L}_{v}$ refer to the losses of the combined features (audio with visual) and of the visual feature classification, respectively. α refers to the weight coefficients for combined features, with α starting at 0.5 and incrementing to 1 throughout training, as follows:

$$\alpha = \alpha_0 + \delta(\varepsilon - 1), \quad (8.3)$$

where α_0 is set to 0.5 as the initial coefficient importance, δ is set to $\frac{1}{60}$ as the coefficient decay degree, and ε refers to the training epoch.

8.3.5 Obtaining Body Data

One of our key contributions relates to combining body with facial cues to retrieve visual features relevant for ASD in varying conditions. However, most ASD datasets do not provide this type of data since current approaches rely solely on face information as visual input. Regarding AVA-ActiveSpeaker, we obtain body bounding box annotations from AVA Actions Dataset [140] (groundwork dataset) and complement them with speaking labels of AVA-ActiveSpeaker, by matching entity id of the original annotations. For Columbia, we use the S3FD face detector [661] based on previous works [76, 1], resizing its predictions to retrieve the body regions by using twice the width and ending the bottom of the bounding box at three times the predicted height. This approach is only viable for this setting given that subjects in Columbia are cropped to the upper body region (sitting). Regarding WASD [3], the original dataset already contains body data annotations.

8.3.6 Implementation Details

ASDnB is trained for 30 epochs with an Adam optimizer, with a initial learning rate of 10^{-4}, decreasing 5% for each epoch. All visual data is reshaped into 112 x 112, audio data is represented by 13-dimensional MFCC, and both visual and audio features have an encoding dimension of 128. For visual augmentation, we perform random flip, rotate and crop, while for audio augmentation, we use negative audio sampling [1]. In sum, given a video data during training, a audio track of a new one is randomly selected from the same batch as noise, maintaining the same speaking label of the original soundtrack.

8.4 Experiments

8.4.1 Datasets and Evaluation Metrics

AVA-ActiveSpeaker. The AVA-ActiveSpeaker dataset [59] is an audio-visual active speaker dataset from Hollywood movies, ranging from 1 to 10 seconds, with 5.3 million face crops, where typically only train and validation sets are used for experiments [1, 76, 70, 77, 4].
WASD. The WASD dataset [3] compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*. With 30 hours of labelled data, WASD is divided into 5 categories ranging from optimal conditions to surveillance settings. We report the results on WASD and on each category, following WASD experiments.
Columbia. We also consider Columbia [570] following the methodology of Light-ASD [76] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning. Columbia consists of an 87-minute panel discussion video, with five speakers (Bell, Boll, Lieb, Long, and Sick) taking turns speaking, with 2-3 speakers visible at any given time.
Evaluation Metrics. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [59] that computes mAP , while for Columbia we use F1-Score.

8.4.2 ASDnB Performance in AVA-ActiveSpeaker

We compare ASDnB performance with state-of-the-art models in AVA-ActiveSpeaker, in Table 8.1. ASDnB outperforms other models while being **lightweight and trained end-to-end**. The increased number of parameters from other ASD approaches derive from heavier extraction power of visual inputs and model components (*e.g.*, GCN) to consider author relation which ASDnB simplifies by splitting 3D convolutions into 2D and 1D (lesser computation cost without loss of performance) and using temporal modeling in the classifier. However, the major contribution of ASDnB relative to existing approaches is the **first efficient combination of face with body data**. The inclusion of body information for ASD is extremely important, in particular for challenging data where face can not be reliably accessed, which is a novel strategy that state-of-the-art models do not yet consider. Only BIAS has previously considered complementing body with face data, but their approach was to treat body has another visual input rather than a complement to face leading to subpar results in AVA-ActiveSpeaker, where face is a reliable visual input for ASD. We are able to efficiently include body for ASD by combining face and body inputs intra encoder, outputting a single combined visual feature that complements facial cues with relevant body movements. Finally, our **pretraining strategy** is mainly to prepare ASDnB for adequate body information extraction by using WASD [3], a ASD dataset containing challenging data such as surveillance settings where face is not reliably accessed. Although AVA-ActiveSpeaker is the benchmark dataset, it is not a good representation of *in-the-wild* [59], with mainly cooperative conditions. As such, for body to have further importance in ASD we prepare the model to perform in harder settings prior to assess ASDnB in more cooperative conditions: AVA-ActiveSpeaker.

Table 8.1: Comparison of ASDnB with state-of-the-art models in AVA-ActiveSpeaker.

Model	Audio Encoder	Visual Encoder	Par(M)	Pre Train	End-to End	Body Data	mAP
ASC [69]	RN18 2D	RN18 2D	23.3	✓	×	×	87.1
MAAS [72]	RN18 2D	RN18 2D	21.7	✓	×	×	88.8
UniCon [71]	RN18 2D	RN18 2D	23.8	✓	×	×	92.2
TalkNet [1]	SE-RN34	RN18+V-TCN	15.0	×	✓	×	92.3
BIAS [4]	SE-RN34	RN18+V-TCN	31.6	×	✓	✓	92.4
ASD-Trans [662]	RN18 2D	RN18+V-TCN	15.0	×	✓	×	93.0
ASDNet [70]	SincDsNet	RNx101 3D	49.7	✓	×	×	93.5
TS-TalkNet [77]	SE-RN34	RN18+V-TCN	36.8	×	✓	×	93.9
EASEE-50 [79]	RN50	RN50 3D	74.7	✓	✓	×	94.1
Light-ASD [76]	Conv 1D	Conv 2D-1D	1.0	×	✓	×	94.1
SPELL [74]	RN18 2D	RN18+TSM	22.5	✓	×	×	94.2
ASDnB	SE-RN34	Conv 2D-1D	2.2	✓	✓	✓	**94.6**

8.4.3 ASDnB Performance in Other Datasets

Challenging data of WASD. WASD is divided into categories with incremental challenges to audio and face data, with the most challenging data having face occlusion (FO), audio impairment with background voices (HVN) and surveillance settings (SS), where face access and audio quality is not guaranteed. Table 8.2 shows that ASDnB is superior to all models that only consider face as visual input, in particular for the Hard categories where face and audio quality is impacted, which rarely occurs in AVA-ActiveSpeaker. The biggest performance discrepancy is in surveillance settings (without reliable face input), where only BIAS performance is similar to ASDnB, given its strategy to also consider body data, highlighting the importance of including body information to create robust ASD models.

Table 8.2: Comparison of ASDnB with state-of-the-art models on the different categories of WASD, using the mAP metric. *OC* refers to Optimal Conditions, *SI* to Speech Impairment, *FO* to Face Occlusion, *HVN* to Human Voice Noise, and *SS* to Surveillance Settings.

Model	Easy		Hard			Avg
	OC	SI	FO	HVN	SS	
ASC [69]	91.2	92.3	87.1	66.8	72.2	85.7
MAAS [72]	90.7	92.6	87.0	67.0	76.5	86.4
ASDNet [70]	96.5	97.4	92.1	77.4	77.8	92.0
TalkNet [1]	95.8	97.5	93.1	81.4	77.5	92.3
TS-Talknet [77]	96.8	97.9	94.4	84.0	79.3	93.1
Light-ASD [76]	97.8	98.3	95.4	84.7	77.9	93.7
BIAS [4]	97.8	98.4	95.9	85.6	82.5	94.5
ASDnB	**98.7**	**98.9**	**97.2**	**89.5**	**82.7**	**95.6**

Robustness of ASDnB in Columbia. We also assess the performance of ASDnB in Columbia, following the methodology of Light-ASD [76] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table 8.3. Although Columbia data contains cooperative subjects, the cross-domain evaluation raises challenges for the models. In this context, ASDnB approach to combine body with face and audio information leads to a state-of-the-art per-

Table 8.3: Comparison of F1-Score (%) on the Columbia dataset.

Model	Speaker					
	Bell	**Boll**	**Lieb**	**Long**	**Sick**	**Avg**
TalkNet [1]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [596]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [76]	82.7	75.7	87.0	74.5	85.4	81.1
BIAS [4]	89.3	75.4	92.1	88.8	88.6	86.8
ASDnB	**91.6**	**81.2**	**93.1**	**91.7**	**90.6**	**89.6**

formance, and highlights the relevance of complementing face with body data for model robustness to perform in varying conditions such as cross-domain settings.

8.4.4 Ablation Studies

Feature Influence. Given the ASDnB novelty of body inclusion for ASD, we explore the influence of different features, and pretraining in WASD, in Table 8.4. The main conclusions are: 1) the variant with only face as visual input and pretraining does not achieves state-of-the-art performances, meaning that body is a necessary feature; 2) pretraining benefits more the ASDnB variant with body than with only face (2.6% vs 0.5%), highlighting that pretraining in the challenges of WASD raises more importance to body information relative to facial cues; 3) the combination of body with face information, without pretraining, is the ASDnB variant with best performing results meaning that face with body complement is necessary for ASD but its relevance increases in more challenging data given that the best results require the WASD pretraining, which contains challenges not seen in AVA-ActiveSpeaker.

Table 8.4: Ablation studies on the effect of WASD pretraining, face and body influence towards ASDnB performance (mAP) in AVA-ActiveSpeaker.

Face	Body	Pretrain	mAP
×	✓	×	83.9
		✓	86.5
✓	×	×	93.7
		✓	94.2
✓	✓	×	94.1
		✓	94.6

Feature Extraction. We explored variations of audio and visual encoders for ASDnB and summarized the results in Table 8.5. Regarding visual encoders, the approach of splitting 3D convolution into 2D and 1D to extract the spatial and temporal information, respectively, significantly outperforms the standard approach of using a ResNet with temporal convolutional network, while also having lower number of parameters. For ASD visual inputs, facial and body movements are the most relevant aspects, meaning that simpler models capture these notions better, without dispersion to other visual features. For audio encoder, the most robust approach of SE-ResNet34 leads to improved results when combined with the lightweight visual encoder. This is mainly due to the pretraining on WASD, with varying audio quality, that requires more robust extraction of audio fea-

tures such as distinguish between relevant audio and background voices to combine with adequate visual information.

Table 8.5: Variation of audio and visual encoders regarding the number of parameters and model performance in AVA-ActiveSpeaker.

Visual Encoder	Audio Encoder	Par(M)	mAP
RN18+V-TCN	Conv 1D	16.5	92.5
RN18+V-TCN	SE-ResNet34	17.6	92.7
Conv 2D-1D	Conv 1D	1.1	94.1
Conv 2D-1D	SE-ResNet34	2.2	94.6

Loss Function. We compare our loss with approaches from other works in Table 8.6. Standard losses only consider the combined audio and visual features as relevant for ASD , while recent works [76, 1], focus on complementing this combined loss with weight importance of individual features. The latter strategies tend to perform better since they motivate ASD models to improve visual feature extraction such that visual cues are a reliable source to predict ASD . Light-ASD further improves this aspect by including a temperature coefficient to control feature importance throughout training epochs. Our approach is based on similar concepts but with two key changes: 1) our starting weight importance for visual and combined features increases the relative importance of visual features to motivate a better conjunction of face with body information in the earlier training stages; and 2) we vary visual and combined features importance through training, such that in later stages visual features lose relevance and combining audio with visual features is the strategy for ASD . Our loss translates into better results relative to existing approaches, highlighting the influence of adaptive weight importance for more reliable ASD .

Table 8.6: Loss effect on ASDnB performance (mAP) in AVA-ActiveSpeaker. I_v refers to the importance of visual features, while I_{av} refers to the importance of combined audio and visual inputs towards ASD prediction. All approaches have Cross-Entropy has the underlying training loss. τ refers to the temperature coefficient.

Approach	I_v	I_{av}	Extra	mAP
Standard	0	1	$\times$	93.1
TalkNet	0.4	1	$\times$	94.0
Light-ASD	0.5	1	τ	94.2
Our	[0.5-0]	[0.5-1]	$\times$	94.6

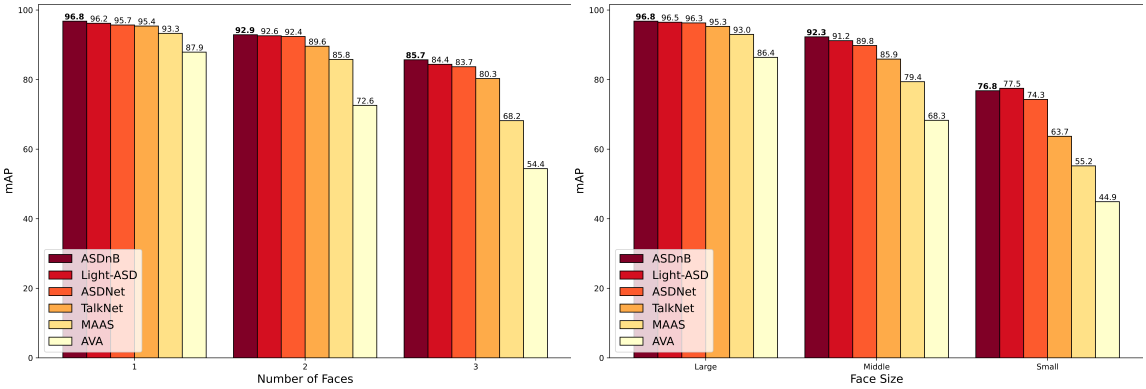
Temporal Modeling for ASD. We assess different temporal modeling approaches for ASDnB classifier, in Table 8.7. Given the ASD context, speaker prediction benefits from including a temporal relation between frames as shown by the results of not having temporal modeling. Increasing this relation by bidirectional (*vs.* forward) temporal modeling translates into better results, with Gated Recurrent Units (GRU) outperforming LSTM. LSTM tends to be more reliable for long-term information while the simplified version of GRU makes neighboring frames more informative, which is a better approach for ASD.

Table 8.7: Performance of temporal modeling methods in ASDnB classifier.

Method	Par(M)	mAP
None	2.02	89.8
Forward LSTM	2.15	93.7
Forward GRU	2.12	93.8
Bidirectional LSTM	2.28	94.4
Bidirectional GRU	2.22	94.6

8.4.5 ASDnB Performance Analysis

Face Size and Number of People. We assess the robustness of ASDnB to deal with variations of AVA-ActiveSpeaker data in Figure 8.5, similar to other works [76, 1, 72, 70]. The methodology considers a face as *Small* with width under 64, *Middle* with width between 64 and 128, and *Large* with width over 128, while for the number of people in the scene the data is divided into three mutually exclusive groups (1, 2, and 3) based on the number of faces detected in a frame, totalling 90% of AVA-ActiveSpeaker data. For all variations ASDnB performance is superior to existing state-of-the-art models, with only a slight underperformance in the smaller face size relative to Light-ASD (76.8% *vs*. 77.5%). With smaller faces and no relevant body information ASDnB is not as robust, meaning that there is room for improvement in these settings, namely in background people of a scene. Aside this scenario, ASDnB is an all-around model for ASD , outperforming existing approaches in varying conditions.



(a) Models performance by the number of faces on each frame

(b) Models performance by faces size

Figure 8.5: Comparison of ASDnB performance relative to ASD state-of-the-art models for (a) number of faces per frame and (b) various face sizes in AVA-ActiveSpeaker.

Relative Body Importance. To further explore the importance of body for ASD, we compare the performance of ASDnB and Light-ASD with varying head-body proportion (HBP), in AVA-ActiveSpeaker, in Figure 8.6. We use all the available testing data in the first pair of bars, and use less data moving left to right on the x-axis, corresponding to lower HBP values. For instance, in the pair of bars at 20%, we are using the data with the lowest 20% HBP meaning that these are settings where the face is significantly smaller than body. The results show that, with a decrease of HBP, ASDnB performance is pro-

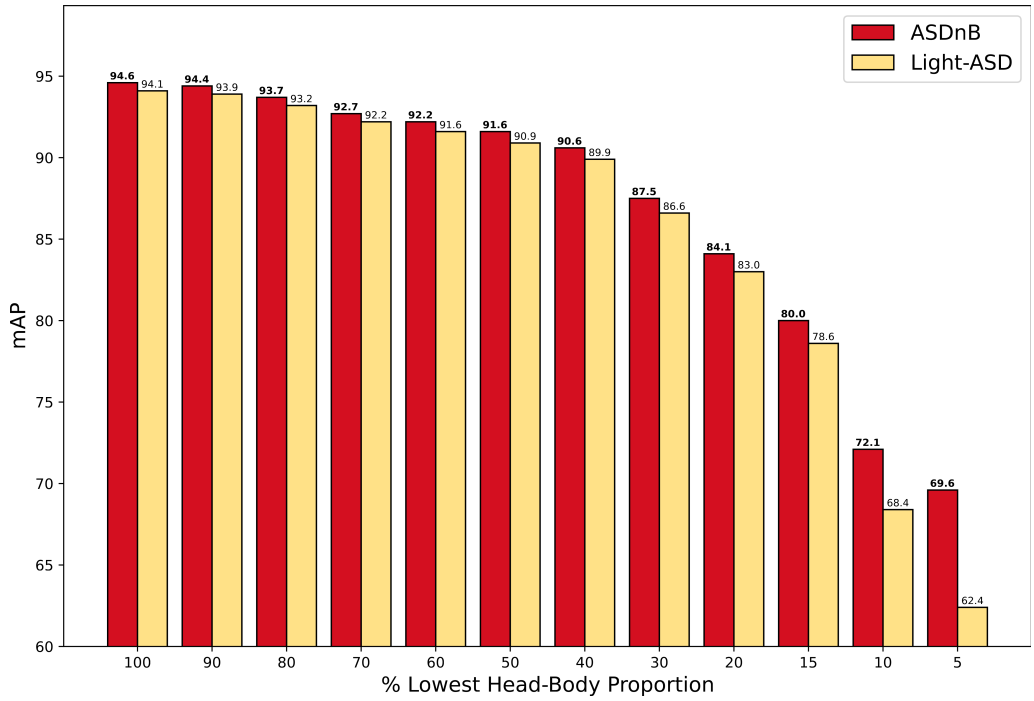


Figure 8.6: Relative body importance of ASDnB and Light-ASD with decremental head-body area proportion in AVA-ActiveSpeaker.

gressively better relative to Light-ASD, highlighting that the absence of reliable face access raises importance in body information. This is particularly predominant in wild conditions such as surveillance settings, meaning that ASDnB strategy is a viable approach to increase ASD robustness to perform in such conditions. Although ASDnB performs worse in smaller faces than Light-ASD (Figure 8.5), this aspect is mitigated with relevant body information meaning that the underperformance of ASDnB in such conditions is mainly due to background people without visible body. With relevant body information ASDnB is able to outperform Light-ASD, which reinforces the notion that body data is a relevant feature to use in challenging ASD data (WASD [3]), or when the subject is not close to the camera (instances of AVA-ActiveSpeaker).

Qualitative Analysis. We complement our experiments with a qualitative analysis of ASDnB and Light-ASD in challenging scenarios of AVA-ActiveSpeaker and WASD, in Figure 8.7. The considered scenarios contain subjects far from the camera and in suboptimal cooperative settings (top-down view) which makes it harder to predict who is speaking using only facial cues. In both examples, Light-ASD misclassified the switch of speakers while ASDnB was more resilient by analysing the hand movement that preceded subject speaking. The results support the importance of body analysis for ASD in wild conditions, where face can not be reliably accessed, making ASDnB a viable baseline for robust ASD models.

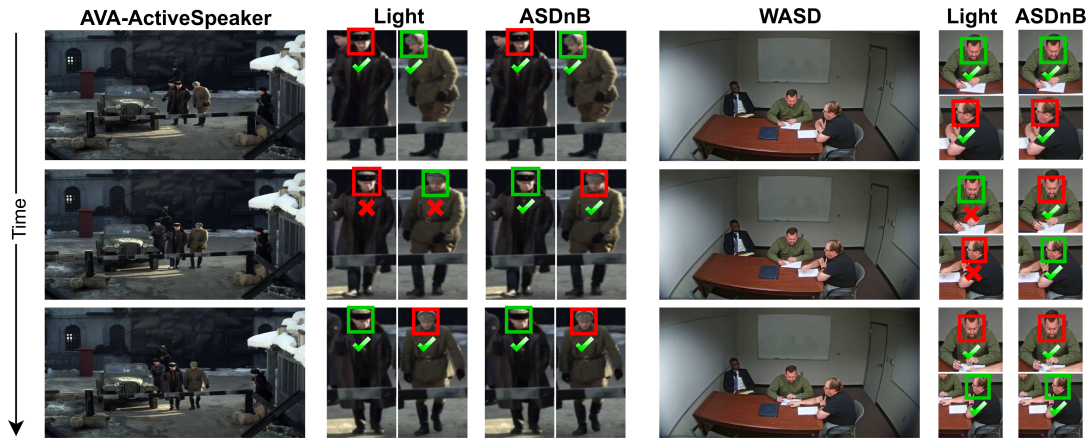


Figure 8.7: ASDnB and Light-ASD (Light) qualitative performance assessment in challenging scenarios of AVA-ActiveSpeaker and WASD. Red bounding boxes denote model prediction of subject not talking, green to speaking, and predictions with red cross denote missclassification relative to the ground-truth. In both examples with subjects far from the camera, Light-ASD misclassified the switch of speakers while ASDnB was more resilient by analysing the hand movement that preceded subject speaking.

8.5 Conclusion

This paper describes ASDnB, a lightweight multi-modal model that, for the first time, efficiently combines face with body information for Active Speaker Detection. The key contribution of our proposal relates to combining face and body features at different feature extraction steps, inspired by the UNet approach, yielding state-of-the-art performance both on cooperative conditions (benchmark dataset AVA-ActiveSpeaker) and on more challenging settings (WASD and cross-domain of Columbia). The obtained results show that complementing body information with facial cues is of utmost importance for ASD robustness, and is particularly important for *wild* conditions (*i.e.*, surveillance settings), where state-of-the-art models do not reliably perform.

Chapter 9

Conclusion and Future Directions

9.1 Conclusion

This thesis proposes a new perspective on how visual-based deep learning approaches can be applied to human behavior analysis. Given the increasing number of works published using deep learning, very few focus their motivation on applying it to human behavior given its inherent challenges. Our proposal aims to define a set of features proven as influential in human behavior (via Psychology literature review) that can be analyzed via visual-based models and without requiring subject cooperation. This way, it is possible to analyze behavioral features to create subject profiling which can be used to detect anomalous behaviors, service customization, among others. Note that the analysis of these features do not correspond to categorization of human behaviors (as depressive, aggressive, *e.g*) but serve as a set of data applicable to different profiling purposes, meaning that it does not replace behavioral experts but rather can be used to facilitate/complement their work.

By defining this new perspective, we also highlight the importance of improving certain areas to work in wilder conditions (surveillance settings), given that some works are mainly focused on cooperative settings (good image quality with subject cooperation), while also demonstrating areas worth exploring given their lesser number of works (*e.g*., human interaction, *i.e*. ASD) relative to more mainstream topics. With this in mind, we showed the limitations of current approaches for biometric assessment in wilder conditions, using gender as a proof of concept, and designed an approach more resilient to these conditions by strategically complementing face and body information for gender inference in surveillance-like settings.

The initial works focused on displaying the limitations of a mainstream topic (gender/ biometric inference) in more unconstrained conditions, while the latter aim to further expand the body of knowledge by targeting a topic not as extensively explored: human interaction. In particular, we focus on ASD, whose state-of-the-art models mainly perform using face and audio data in good quality data (AVA-ActiveSpeaker). With the goal of highlighting their limitations in more unconstrained conditions, we created a new wilder ASD dataset (WASD) given the lack of challenging data in this area. This dataset showed that current models underperform in scenarios with degrading audio and image quality, highlighting the importance of using body to increase model robustness, particularly in surveillance settings. Following these findings, we proposed a new model (BIAS) that, for the first time, complements body with audio and face information for robust ASD in varying data (challenging and cooperative), while also providing visual interpretability. We further extend our contributions for the ASD field, by proposing an improved model (ASDnB) that

combines body and face information intra encoding, translating into a combination of relevant body information with facial cues, resulting in state-of-the-art performances in both cooperative and challenging conditions.

9.2 Future Directions

Given our new perspective of human behavior analysis via assessment of particular features to create subject profiling, several different topics can also be explored:

- **Assess performance of other features in unconstrained conditions**: Similar to our works on gender/biometric and ASD assessment in wilder conditions, the resilience of other topics in these conditions can also be explored (*e.g.* clothing). The findings of models' limitations can then serve as basis to the proposal of more robust approaches;

- **More data in unconstrained conditions**: We found a lack of surveillance-like data when trying to assess the limitations of ASD in these conditions, which highlight a area where future works can focus on: create datasets of uncontrained conditions. Several feature areas are already well supported with this type of data but others can be improved (*e.g.* physical human interaction). This is particularly important to motivate the development of models more robust and not ones that only reliably perform in perfect conditions.

- **Further ramifications of the features proposed**: Although the features considered have Psychology literature to support their choice, the list of admissible features is not extensive, meaning that other areas or sub-areas can also arise to complement the current proposal (*e.g.* gait assessment). Furthermore, future papers/projects can also increase the relevance/motivation of their work by associating to the topic of human behavior analysis using this thesis as reference;

- **Improvement on ASD in surveillance conditions**: The complement of body to improve ASD performance in these settings is a step in the right direction but it can be further improved to reliably perform in the "real world". One potential improvement could be the assessment of people/crowd interaction with a candidate to active speaker in a scene. When a person is talking, others typically present nonverbal behaviors to indicate engagement in ones conversation/interaction and these type of interactions are not duly considered by state-of-the-art models, which shows room for improvement in future works;

- **Embedded system to user profiling**: The basis of the new behavioral perspective is to create a user profile to better model their behavior. By analysing the set of features proposed using state-of-the-art models it is possible to obtain a data to group individuals based on their behavior (*e.g.* via clustering). This can then be applied to shop service customization (identify the type of customer one is, buyer

or viewer, and act on it) or even detect anomalous behavior (*e.g.* person running around in a library, fight in streets, crowd running).

Appendix A

Other publications resulting from this doctoral research program not included in the thesis.

On Exploring Audio Anomaly in Speech

Tiago Roxo
Inst. de Telecomunicações
Univ. da Beira Interior
Covilhã, Portugal
tiago.roxo@ubi.pt

Joana Cabral Costa
Inst. de Telecomunicações
Univ. da Beira Interior
Covilhã, Portugal
joana.cabral.costa@ubi.pt

Pedro R. M. Inácio
Inst. de Telecomunicações
Univ. da Beira Interior
Covilhã, Portugal
prmi@ubi.pt

Hugo Proença
Inst. de Telecomunicações
Univ. da Beira Interior
Covilhã, Portugal
hugomcp@di.ubi.pt

***Abstract*—Existing anomaly detection works mainly focus on abnormal activities in image and video settings, while assessing audio manipulation, namely the presence of anomalous audio in speech, has not yet been explored. To overcome this limitation, we propose a setup in the context of Active Speaker Detection (ASD) by defining a methodology to perceive audio anomaly, assessing the performance of anomaly models, and establishing setup variations. This way, we evaluate models performance in identifying the presence of anomalies (detection) and localizing the timeframe where they occur (localization). To complement anomaly detection, we propose Anomaly Score (AS), a metric to assess anomaly localization that balances precision and mis-localization. Given the sequential nature of audio, we explore the performance of a density-based approach for video anomaly (CPD) and recurrent models (LSTM and RNN) on detecting and localizing audio anomalies. The results show that: 1) anomaly inclusion in talking portions increases models resilience toward anomaly localization; 2) CPD is superior in anomaly detection, while recurrent models perform better in anomaly localization; 3) anomaly with distinctive audio benefits precise anomaly localization; and 4) using original ASD audio is overall the best approach, relative to other processing approaches. The setup and experiments of this work serve as a baseline for future works on speech anomaly detection.**

***Index Terms*—Active speaker detection, anomaly setup, audio anomaly.**

I. Introduction

Anomaly detection is typically linked to the assessment of abnormal events in a given setting, such as detecting violence in city surveillance videos [1], defect detection [2] or medical imaging [3], with most works designed to perform in image or video settings. However, with a raise on image and audio deep fakes, there is still no work reported on identifying audio manipulation in speeches, *i.e.*, a model that is able to perceive if a subject audio speech was altered to something else. In particular, the context we are focused on is perceiving changes mid speech, where the majority of audio belongs to a subject but a portion of it was changed to audio of someone else, which is a more challenging set than completely generated fake/anomalous audio.

In this paper we propose a speech anomaly detection setup, by defining a methodology for anomaly detection and localization in the context of Active Speaker Detection (ASD). As such, we use AVA-ActiveSpeaker, an ASD dataset, and retrieve the audio encoding of a standard ASD model (Figure 1) to feed anomaly models. By considering audio as sequential data, we explore the performance of a density-based approach for video anomaly [4] (CPD) and recurrent models (LSTM and RNN) on detecting and localizing audio anomalies, for different setups: anomaly size and probability variation, similarity between audio and anomaly used, audio preprocessing, or audio combination with visual data. We identify anomalies via audio partitioning using sliding window to retrieve audio intervals, used by anomaly models to predict anomaly presence. For this reason, we complement our setup with the proposal of Anomaly Score (AS), a new metric that relates the precision of anomaly interval identification with misclassifications. This way, we can perceive how our models perform in identifying the presence of anomalies (detection) and also where they occur (localization). The results show that throughout the various experimental settings, CPD is better than recurrent models at detecting anomalies, with the latter being better at localizing them, and anomaly inclusion in talking portion of the original ASD audio is the better approach for more resilient anomaly models, overall. In sum, the main contributions are:

- We propose a speech anomaly detection setup, in the context of Active Speaker Detection, by defining a methodology to perceive audio anomaly, assessing anomaly models performance, and establishing setup variations;
- We propose Anomaly Score, a new metric that relates the precision and misclassification of anomaly localization

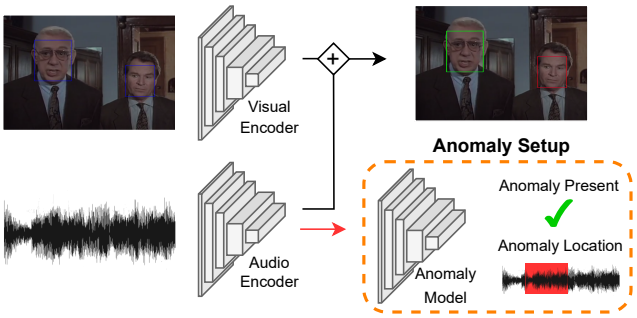


Fig. 1. Overview of the considered audio anomaly detection context. The features extracted from the audio encoder of a standard ASD model are given to an anomaly model to perceive the presence of an anomaly (detection) and at which frames an anomaly most likely occurred (localization).

This work was supported by the Portuguese FCT/MCTES through National Funds and co-funded by EU funds under Project UIDB/50008/2020 and by the FCT Doctoral Grant 2020.09847.BD and Grant 2021.04905.BD.

that complements standard classification metrics;
- We explore the performance of a density-based video anomaly model (CPD) and recurrent models (LSTM and RNN) on various experimental settings to assess its applicability in audio anomaly detection and localization.

II. Related Work

Video Anomaly Detection. Most works on anomaly detection are based on video/image, and typically rely on **reconstruction-based** approaches, *i.e.*, try to reconstruct the initial input, with increased error in reconstruction linked to anomaly presence. These methods use either auto-encoders [5] or Generative Adversarial Networks (GAN) [6] for reconstruction and improve anomaly detection via architecture changes (*e.g.*, memory-based autoencoders [7], [8]), training optimization [9], [10], distillation [11], [12], bag of videos [13], and motion inclusion [14], [15]. There are also **density-based** approaches for video anomaly detection that model the distribution of normal data (no anomaly), with the assumption that abnormal data differs (low likelihood) from normal one [16]. Several works consider that density can be expressed through some pre-defined distributions [17], [18], although other approaches explore density estimation via deep generative models [4], [19].

Audio Anomaly Detection. For audio anomaly, the works are mainly based on industrial [20] (machine and motor sounds) and healthcare [21] (phonocardiogram) sound anomaly. The goal is to infer if a given sound is different from the expected (normal) one, applicable to detect if a machine is malfunctioning or if a patient is suffering from some disease. Approaches can obtain features via hand-crafting [22], through auto-encoders [23], or signal processing methods [24], and assess audio anomaly via reconstruction and density-based approaches, similarly to video anomaly detection. Relative to the state-of-the-art, our setup is done in human speech, where the anomalous sound appears in portion of the speech, while keeping parts of the original audio in. The sound anomaly derives from speeches of another person, forcing the anomaly detection methods to perceive: 1) what is the original sound; and 2) what is the new *anomalous* sound that changed mid speech. This a more challenging setup than the state-of-the-art, and can be applied to perceive deep fake audio manipulation with malicious intent (defamation).

III. Speech Anomaly Setup

A. Anomaly in Active Speaker Detection

ASD aims to identify, from a set of potential candidates, active speakers on a given visual scene [25], done at the video frame level using facial cues and audio information. We retrieve features from the audio encoder of TalkNet [26], and feed them to anomaly models (Figure 1). Following their approach, audio frames are represented by a vector of Mel-Frequency Cepstral Coefficients (MFCCs) with a window step of 10 ms. In the context of our work, we consider as *original* the audio from AVA-ActiveSpeaker, and *audio with anomaly* as one that suffered manipulation, *i.e.*, portions of the original audio are replaced by audio of another person. Note that the anomalous audio (anomaly) is from a different subject of the dataset and is not artificially generated.

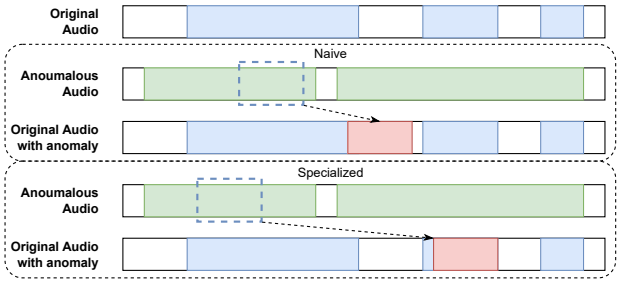


Fig. 2. Illustration of anomaly in speech manipulation for *naive* and *specialized* setups: while naive can alter audio in any frame, specialized forces anomaly inclusion in talking frames. Both original and anomalous audio belong to people of datasets, with the colored bars representing the frames containing talking audio.

B. Anomaly Models

The premise of anomaly detection considered is based on audio as a sequence of timeframes, where previous steps influence future ones, and the presence of anomalies is influenced by normality disruption, based on the timeframe sequence. As such, the type of anomaly models in video settings that more resemble this premise are **autoregressive models** which provide a general formulation for tasks involving sequential predictions, in which each output depends on previous observations. Our reference for these type of model is the one proposed by Abati *et al.* [4], expressed by:

$$p(z) = \prod_{i=1}^{n} p(z_i | z_{0,..,i-1}), \tag{1}$$

where z is the audio latent representation, and $p(z)$ relates conditional probability densities (CPD) of previous timeframes for a given event. In their work, they design autoregressive layers to guarantee that CPD is computed from inputs of previous occurrences, via Masked Fully Connections, and assess video audio anomaly by combining reconstruction and density-based predictions. In our experiments, we only use the density-based portion of this model, which we refer as **CPD** from now on.

To complement the assessment of anomaly detection via sequential data analysis, we select Long Short-Term Memory (**LSTM**) and Recurrent Neural Network (**RNN**), given that their underlying approach is based on this premise (recurrent connections). Furthermore, these two models complement each other regarding audio analysis, in particular the effect of longer memory of LSTM relative to RNN.

C. Anomaly Setups

Similar to video anomaly detection approaches, we train the anomaly detection model to reliably recognize standard settings (*i.e.*, without anomaly) so that the presence of an anomaly is regarded as a standout event. Given the ASD

context, during a scene subjects may talk uninterruptedly, talk interchangeably with other subjects, or have no talking moments. As such, to assess the effect of **different inference conditions** we start by considering two initial setups: 1) **naive**, where we allow the insertion of anomaly in any audio location; and 2) **specialized**, where audio anomaly is only inserted in talking frames, with guarantee that the initial talking frames belong to the original person prior to anomaly inclusion. Note that for both setups, the anomalous sound contains only talking audio via ground truth labels assessment. Figure 2 highlights the two setups considered and key differences between them.

For training, all models receive the original audio without manipulation, trained by means of Maximum Likelihood Estimation (MLE), *i.e.*, based on the log-likelihood term of [4] loss (mean log-likelihood, averaged along the batch axis).

In the experimental section of variations to anomaly setup, we assess the effect of using only talking audio and combining audio with visual data. To obtain **only talking audio**, we preprocess the dataset to combine only the talking frames of each subject, via ground truth labels assessment, before inputting to the anomaly model. For **audio visual combination**, we stack the feature vectors from audio and image encoders (Figure 1), before inputting to the anomaly model.

D. Audio Similarity

Another setup variation considered is the choice of an anomaly. Instead of randomly selecting one audio to include on another, we can opt to choose audios similar or distinctive to the original. For this, we use Dynamic Time Warping (DTW) [27] to **assess audio similarity**, a dynamic programming technique to find the optimal temporal matching between elements of two time series. It is typically used for temporal data analysis, does not rely on equal duration for similarity comparison, and is able to compare with accelerations and decelerations throughout audio, making it suitable to compare audio in ASD. We compute the DTW values of each audio pair of AVA-ActiveSpeaker (ASD Dataset) using the FastDTW variation [28], normalize them, and consider audios as *low similarity* if they belong at the bottom 5%, following the normal distribution percentile formula, and *high similarity* those at the top 5%.

E. Anomaly Detection and Localization

For audio anomaly in speech, we assess models performance based on: 1) detecting that an anomaly is present; and 2) localizing the frame interval where it occurred. As such, we select **Recall** as our metric for **anomaly detection**, given the importance of assessing how many anomalies can models detect from the pool of existing ones.

Contrary to anomalies in video-based approaches that assess anomaly values based on perceived value of *normality* (using reference values from training, *e.g.* [4]), we can only assess anomaly based on audio variation (*i.e.*, at what timeframe is the sound starting to sound different/from another person). As such, we follow a sliding window approach and divide the audio into intervals where each corresponds to

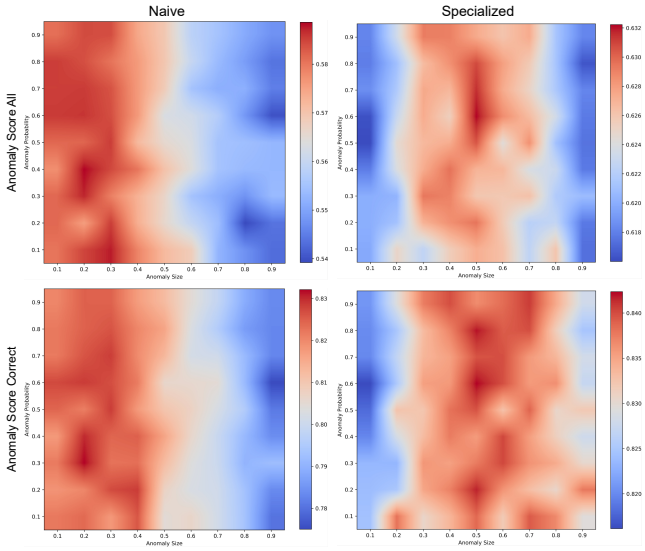


Fig. 3. AS_{All} and AS_C in naive (left) and specialized (right) experimental setups, with varying anomaly size (x axis) and probability (y axis).

10% of the original audio duration, with a stride of half the interval. We consider a interval as containing anomaly if the predicted anomaly value is in the top 10%, following the normal distribution percentile formula. Then, given the intervals for anomaly detection, models may accurately predict the interval where the anomaly frame belongs while also incorrectly predict other intervals where the anomaly did not occur. As such, **for anomaly localization** we propose a metric that balances correct predictions with incorrect ones, as:

$$AS_{[a,b]} = \begin{cases} 1, & \text{if} \quad \alpha \in [a,b] \\ 1 - \Delta/t, & \text{otherwise,} \end{cases} \tag{2}$$

where α refers to the anomaly frame, a and b are the initial and final frames of the interval, t is the number of intervals not containing α, and Δ is the number of intervals between $[a,b]$ and the closest one containing α. We assess AS for each interval considered as potential anomaly, with the final AS for each speech being the average for all intervals predicted with anomaly. Note that we consider two variations of AS in our experiments: 1) AS_{All}, which refers to the average of AS for all audios; and 2) AS_C, which only considers the audios where a model accurately predicted the anomaly interval. These variations highlight the overall anomaly localization (AS_{All}) and the precision of cases with accurate anomaly localization (AS_C).

IV. Experiments

A. Dataset, Models, and Evaluation Metrics

Dataset. The AVA-ActiveSpeaker (AVA) dataset [25] is an audio-visual active speaker dataset from Hollywood movies, totalling almost 38 hours. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723

TABLE I
CPD, LSTM, AND RNN ANOMALY DETECTION (RECALL) AND LOCALIZATION (AS) ASSESSMENT IN NAIVE AND SPECIALIZED SETUPS.

Anomaly	Model	Recall	AS_{All}	AS_C
Naive	CPD	93.68 ± 1.99	57.28 ± 1.30	81.09 ± 1.76
	LSTM	73.45 ± 1.07	65.71 ± 1.47	87.16 ± 0.93
	RNN	87.82 ± 0.68	72.45 ± 1.45	87.08 ± 0.71
Specialized	CPD	95.71 ± 0.51	62.63 ± 0.50	83.59 ± 0.64
	LSTM	73.13 ± 0.85	68.41 ± 0.75	87.45 ± 0.65
	RNN	88.66 ± 0.66	74.57 ± 0.51	86.91 ± 0.44

and 8,015 video utterances, respectively, ranging from 1 to 10 seconds

Models. For anomaly detection models, we use CPD [4] with 100 quantization bins, and recurrent models that consider inputs as sequence of data, such as LSTM and RNN, from Pytorch. The ASD model used to include anomaly detection models is TalkNet [26], a model focused on long-term analysis, trained end-to-end, following the authors methodology.

Evaluation Metrics. We use Recall as standard evaluation metric for anomaly detection, and for anomaly localization we use AS, described in Section III-E.

B. Anomaly Inclusion and Model Performance

We start by considering two setups of anomaly inclusion, as described in Section III-C: 1) *naive*, where anomaly can be inserted in any audio frame; and 2) *specialized*, where the anomaly must be inserted in a audio frame where the original subject was talking.

Anomaly Size and Probability. By varying anomaly size and probability of inclusion in both settings (Figure 3), we observe that with the naive approach models can only reliably localize anomalies with smaller sizes, decreasing their performance with increased anomaly size. This is mainly due to the randomness of audio inclusion in the naive setting, leading to potentially be included in frames without talking or replacing completely the original audio. This effect is enhanced with increased anomaly size, causing models to not be able to compare the original with changed audio, thus the decrease in both AS. Relative to the *specialized* setting, the areas with lower performance are the extreme sizes which is linked to models not being able to perceive audio changes when the anomaly is too small (not easily perceived) or too big (almost replacing the original audio). By including changes in talking portions, models are able to be more resilient to greater anomaly size variations, particularly in conditions where they can reliably locate the anomaly (AS_C). Finally, anomaly probability is not the distinctive factor in anomaly localization for both settings, given the relative homogeneity throughout probability variation.

Naive vs Specialized. To complement assessing the influence of settings in anomaly detection and localization, we compute the Recall and AS for the considered models, in Table I. When comparing naive with specialized settings, besides the lesser resilience of models to anomaly size variation in naive, they also increase the overall mislocalization of anomaly (AS_{All}). Relative to anomaly detection (Recall) and anomaly localization when detecting the right interval (AS_C), only CPD suffers a minor change, with the others being relatively stable. This shows that the naive approach does not greatly impacts anomaly detection (Recall) and mainly affects the anomaly localization (AS_{All}), while the cases where anomaly inclusion is easier to localize (AS_C) are not greatly impacted. Given the results, the remaining experiments use the specialized setting with anomaly sizes ranging from 0.3 to 0.7.

TABLE II
EFFECT OF AUDIO SIMILARITY VARIATION IN ANOMALY LOCALIZATION (AS_{All} AND AS_C), FOR ALL MODELS.

Metric	Similarity	CPD	LSTM	RNN
AS_{All}	High	62.17 ± 0.44	68.30 ± 0.67	74.36 ± 0.57
	Low	62.67 ± 0.42	68.57 ± 0.56	74.80 ± 0.45
AS_C	High	83.14 ± 0.32	87.11 ± 0.33	86.42 ± 0.27
	Low	84.37 ± 0.30	87.98 ± 0.17	87.47 ± 0.28

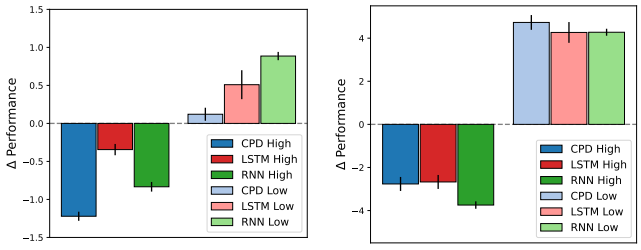


Fig. 4. Performance variation of audio similarity, relative to random anomaly inclusion, for overall (AS_{All}, left) and precise (AS_C, right) anomaly localization. *High* and *Low* refer to using anomaly with audios of high or low similarity relative to the original one.

Models Performance. Relative to models performance (Table I), CPD is the most reliable model to detect anomalies, while LSTM and RNN are better at localizing them. Although both LSTM and RNN share the same principle of recurrent networks and treating inputs as sequential data, the increased complexity of LSTM relative to RNN hinders its ability to extrapolate and perform in the testing setting, particularly in anomaly detection. By considering information in memory for a longer period of time relative to RNN, LSTM is not as resilient to subtle audio changes since it analyses audio more as a *whole* instead of smaller sequential changes such as RNN. Note that given the anomaly context (ASD), detecting and localizing anomalies is heavily based on smaller subsequent audio changes, which benefit RNN more than LSTM. For CPD, the model was initially created to detect anomalies in video settings, which is done in blocks of 16 frames [4]. In our setup, anomaly detection is similar to its original intent, but localization is different (*i.e.*, identify the frame where it occurred), thus its slight underperformance relative to other models.

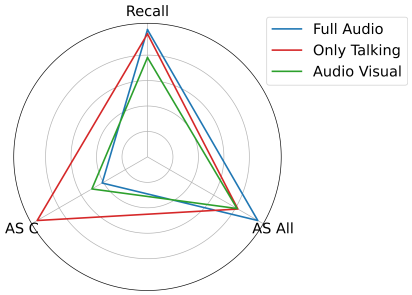


Fig. 5. Anomaly detection and localization comparison on CPD, using original audio (*Full Audio*), only talking portions (*Only Talking*) or complementing original audio with image data (*Audio Visual*).

C. Influence of Audio Similarity

To explore the effect of anomaly characteristics, instead of randomly selecting audio from other subjects of the dataset as anomaly, we assess the effect of using audio of subjects similar to the original audio (high similarity) and different from the original (low similarity) on anomaly localization. The values obtained are in Table II, with the performance variation (expressed by percentage error reduction/increase, relative to random anomaly sound) is shown in Figure 4. The results show that high similarity has a negative impact on anomaly localization, with relative increased performance using low similarity, in particular for precise anomaly localization (AS_C). As such, using distinctive audio as anomaly is a contributing factor for anomaly models. While overall anomaly localization (AS_{All}) also exhibits the same trend, the scale of the latter is smaller than for AS_C, meaning that audio similarity is not significantly impactful for overall localization but improves precise localization. This shows that by considering audio as sequential data, audio similarity has a slight impact in anomaly localization, with anomaly size being the more influential feature (Figure 3).

D. Variations to Anomaly Setup

Given the context of anomaly detection and localization (ASD), there are some variations to consider, namely: 1) assess the effect of additional data; and 2) process audio to guarantee only talking portions. We explore the effects of these variations on CPD with the specialized setup, in Figure 5.

Audio Visual Complement. By complementing audio with image frames and inputting them to CPD, anomaly detection and localization are lower than using only audio. Although visual data complements audio information, we can not guarantee that both data follows a distribution with similar pattern, thus forcing CPD to assess density distribution of combined data leads to underperformance in anomaly detection (Recall), with minor improvements in accurate anomaly detection (AS_C).

Audio *vs.* Only Talking. When comparing the effect of using all audio *vs.* only talking, the major difference is the more accurate anomaly localization when the correct anomaly frame interval is localized (AS_C). By using only talking audio, pauses and background noises (while not talking) are ignored, leading to less incorrect anomaly detection predictions (*i.e.*, increase of AS_C). This does not translate to overall improved anomaly detection predictions (AS_{All}), since the merging of only talking audio induces unnatural audio sequences, which hinders the perception of original and "changed" audio. Thus, although anomaly detection is the same, only talking requires preprocessing of original audio (less realistic scenario) and induces artifacts that impact the overall anomaly localization.

V. Summary and Discussion

One of the main contributions of this work is the definition of an audio anomaly setup in speech, whose importance and key points we highlight below:

- **How to perceive audio anomaly:** We consider analyzing audio as sequential data to perceive sequential disruption. This is mainly due to how humans perceive sound and are able to detect when an audio was changed (*i.e.*, comparing to previous record sounds/timeframes). This is also the reason behind the models chosen for our experiments;
- **How to analyze audio anomaly:** Given the sequential analysis approach, our proposal is to create audio timeframe intervals containing slight overlap (sliding window with stride) for models to assess the location of anomalies. Note that this is different from standard video-based approaches, where the analysis is done at the frame level to detect anomaly, not necessarily localize them pixel wise. For comparison, in CPD [4], anomaly localization is done via patch occlusion to assess anomaly score variation, meaning that they are actually "detecting" anomaly as a whole to perceive the location (pixel-wise) where it occurred;
- **In what context should we analyze anomalies:** We propose assessing this in a challenging speech setting: ASD, where audio and image are used to assess who is talking in a scene, with potential audio overlap, background noise, and varying audio quality. In this context, anomaly models receive the audio encoding features to assess audio anomaly. Note that we use TalkNet for our experiments, but other ASD models can be used to explore the effect of different audio encoding/audio processing strategies to synergize with anomaly models;
- **What type of setups can be used to assess audio anomaly:** The choice of ASD context allows for different setup variations to assess pause/silence effect, data complement, and audio characteristics, namely: 1) naive anomaly inclusion *vs.* inclusion on talking portions; 2) audio similarity effect in anomaly selection; 3) using only talking audio *vs.* original audio; and 4) complement with additional data (image);
- **How to assess models performance:** Contrary to video-based anomaly approaches, where the focus is on anomaly detection, we assess models in **detecting** (audio has an anomaly) and **localizing** anomalies (identifying the timeframe where the anomaly occurs). Although there are available metrics for detection, we propose a

new metric for localization to suit the audio anomaly context: Anomaly Score. It uses models predictions of anomaly presence in audio intervals, and we consider two variations of the metric to assess overall and accurate anomaly localization (AS_{All} and AS_C, respectively).

Another important consideration of the **audio anomaly setup** used is that it **is significantly different from image/video-based**, since that for the latter, train/test contain similar characteristics (background not changing), while in our setup (speech) train/test are vastly different, with artifacts and challenges not easily quantified (background noise, talking overlap, pitch/volume variation) that difficult anomaly detection and localization. This partly justifies the increased number of works in video settings and less/absent on audio ones, highlighting the importance of our work to pioneer a setup for audio anomaly in speech.

VI. Conclusions

We define a methodology to perceive audio anomaly, assess anomaly models performance, and consider setup variations for anomaly detection and localization in the context of Active Speaker Detection. Using this proposal, we explore the performance of CPD, a density-based video anomaly model, and recurrent models (LSTM and RNN) throughout different settings and anomaly conditions, using the proposed Anomaly Score metric for anomaly localization. The results show the superiority of CPD for detection and recurrent for localization, anomaly inclusion in talking audio contributes to more resilient models, and anomaly size is more impactful than audio similarity for anomaly localization. With this work, we define an audio anomaly setup that serves as a reference for future works on speech anomaly detection.

References

[1] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A comprehensive review on vision-based violence detection in surveillance videos," ACM Computing Surveys, vol. 55, no. 10, pp. 1–44, 2023. 1

[2] Z. Ren, F. Fang, N. Yan, and Y. Wu, "State of the art in defect detection based on machine vision," International Journal of Precision Engineering and Manufacturing-Green Technology, vol. 9, no. 2, pp. 661–691, 2022. 1

[3] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection–a survey," ACM Computing Surveys (CSUR), vol. 54, no. 7, pp. 1–37, 2021. 1

[4] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 481–490. 1, 2, 3, 4, 5

[5] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 733–742. 2

[6] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6536–6545. 2

[7] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13 588–13 597. 2

[8] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14 372–14 381. 2

[9] F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," IEEE Access, vol. 8, pp. 88 170–88 176, 2020. 2

[10] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, "Video anomaly detection with sparse coding inspired deep neural networks," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 3, pp. 1070–1084, 2019. 2

[11] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12 742–12 752. 2

[12] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9737–9746. 2

[13] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," IEEE Signal Processing Letters, vol. 27, pp. 1705–1709, 2020. 2

[14] X. Huang, C. Zhao, and Z. Wu, "A video anomaly detection framework based on appearance-motion semantics representation consistency," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5. 2

[15] B. Li, S. Leroux, and P. Simoens, "Decoupled appearance and motion learning for efficient anomaly detection in surveillance video," Computer Vision and Image Understanding, vol. 210, p. 103249, 2021. 2

[16] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," arXiv preprint arXiv:2110.11334, 2021. 2

[17] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the mahalanobis distance," Journal of experimental social psychology, vol. 74, pp. 150–156, 2018. 2

[18] M. Turcotte, J. Moore, N. Heard, and A. McPhall, "Poisson factorization for peer-based anomaly detection," in 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016, pp. 208–210. 2

[19] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. Springer, 2019, pp. 3–17. 2

[20] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," arXiv preprint arXiv:1909.09347, 2019. 2

[21] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson et al., "An open access database for the evaluation of heart sound algorithms," Physiological measurement, vol. 37, no. 12, p. 2181, 2016. 2

[22] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 713–719, 2011. 2

[23] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," arXiv preprint arXiv:2006.15321, 2020. 2

[24] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," arXiv preprint arXiv:1710.02997, 2017. 2

[25] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi et al., "Ava active speaker: An audio-visual dataset for active speaker detection," in 2020 IEEE ICASSP. IEEE, 2020, pp. 4492–4496. 2, 3

[26] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3927–3935. 2, 4

[27] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE transactions on acoustics, speech, and signal processing, vol. 26, no. 1, pp. 43–49, 1978. 3

[28] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, vol. 11, no. 5, pp. 561–580, 2007. 3

WASD: A Wilder Active Speaker Detection Dataset

Tiago Roxo, Joana C. Costa, Pedro R. M. Inácio, *Senior Member, IEEE,* Hugo Proença, *Senior Member, IEEE*
Instituto de Telecomunicações, University of Beira Interior, Portugal
{tiago.roxo, joana.cabral.costa}@ubi.pt, {prmi, hugomcp}@di.ubi.pt

***Abstract*—This document provides additional details to support the main submission. We begin by describing talking and head bounding box annotations in Sections A and B, respectively. Furthermore, we provide visual examples of Wilder Active Speaker Detection (WASD) categories in Section C, describe the full list of analyzed features for video selection in Section D, and extends the cross-domain experiments of the main manuscript in Section E.**

***Index Terms*—Active speaker detection, body-based analysis, dataset, visual surveillance, wild conditions.**

A. Talking Annotations

WE design a custom Graphical User Interface (GUI) for active speaker annotations, as shown in Figure 1. While the video is running, visual sliders are displayed (one for each speaker), automatically filled with red to denote absence of talking. We select a speaker by pressing the corresponding number key (*e.g.*, third speaker is selected with "3" key) and change the speaking label (and slider color) via Ctrl key. To pause, rewind, and forward, we use space, left, and right arrows, respectively. Video time is rewinded and forwarded by 5 seconds with each key press. Prior to GUI launch, we manually set a variable regarding the number of speakers for each video.

AVA-ActiveSpeaker Format Conversion. While performing annotations (either speaking or body/face), we save them on a JavaScript Object Notation (JSON) file (1 file per video), containing all the information used (head/body bounding boxes coordinates, person id, and speaking label), grouped by frame name and person. To convert to AVA-ActiveSpeaker format, we follow the authors guidelines [8], where each line of the annotation file relates to a time frame of a person in a video. To obtain the time frames, we start at time 0 with increments of video duration per video frames. Each line has the entity id (video name with person id), time frame, face bounding box coordinates, and speaking label. For body annotations, we replace face bounding boxes coordinates for body ones. The custom annotations and all AVA-ActiveSpeaker Comma-Separated Value (CSV) files are available at https://github.com/Tiago-Roxo/WASD.

B. Head Bounding Box Annotations

The default approach for head bounding box annotation is based on pose data. Using Alphapose [3], [6], [11], we retrieve the x and y coordinates of the right and left ears, right and left eyes, and nose. The head's central point is calculated using the arithmetic mean of the eye-nose reference point (mean of eyes and nose positions) and ears coordinates. Head bounding boxes are centered in the head's central point, with height and width as a fraction of body silhouette height. This fraction is manually set for each video to ensure adequate head area capture. Figure 2 displays examples of reference points used and head bounding box drawing in different scenarios. In conditions where this approach was not entirely suitable (most Surveillance Settings videos), we annotated manually.

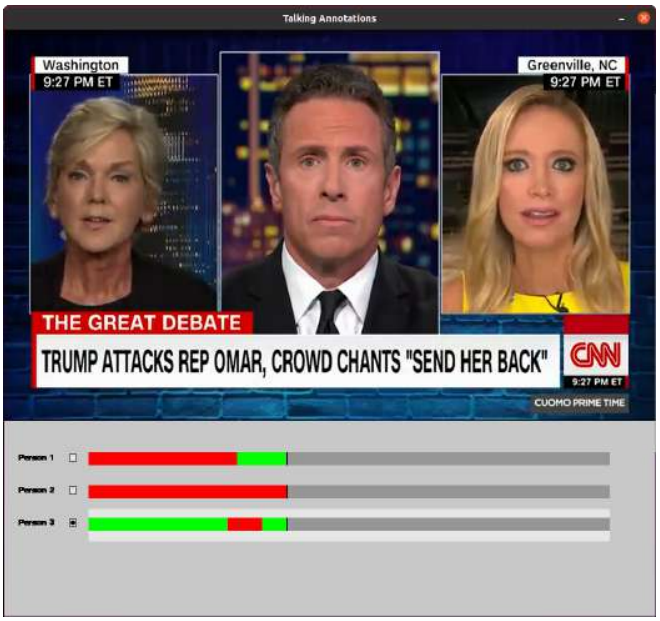


Fig. 1: GUI program used for talking annotations. Three speakers are represented by 3 sliders, below the video. Green color refers to speaking, while red represents silence. Leftmost speaker corresponds to the first slider.

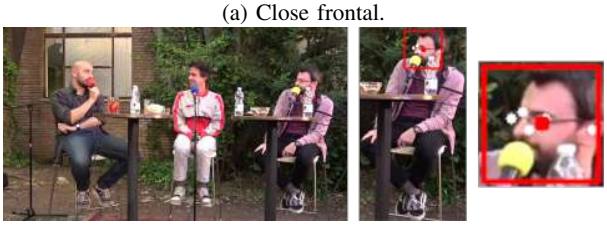
C. WASD Categories

We provide examples of scenarios considered for Wilder Active Speaker Detection (WASD) categories in Figure 3: *Optimal Conditions* mainly consists of interviews or people talking in an alternate manner, with cooperative poses; *Speech Impairment* refers to political debates, heated discussions, and online interviews/debates; *Face Occlusion* contains scenarios where subjects have partial face occlusion from the objects; *Human Voice Noise* relates to subjects reacting to a video, while it plays in the background, contributing to audio impairment; and *Surveillance Settings* are from surveillance interrogations, with variable audio and image quality (*i.e.*, face access and subject cooperation).

Manuscript received XX, 2023; revised XX, 2023.



(a) Close frontal.



(b) Far frontal.



(c) Close sideways.



(d) Far sideways.

Fig. 2: Head bounding box drawing in different scenarios. From left to right, all images contain the original scenario, head bounding box drawing, and zoom in for better visualization. White dots refer to the reference points used for head bounding box drawing, while the red dot is the head's central point. This approach is suitable for various conditions such as close or far frontal poses (a and b), even with facial occlusion, and in side poses at closer or farther positions (c and d, respectively).

D. WASD Features

The complete list of all the considered features, and their admissible values, are the following:

- **Facial Occlusion**: Yes or No;
- **Human Voice as Background Noise**: Yes or No;
- **Speech Overlap**: None-Low or Medium-High;
- **Delayed Speech**: Yes or No;
- **Surveillance Settings**: Yes or No;
- **Proximity to Camera**: Low, Medium, or High;
- **Audio Quality**: Low or High;
- **Face Availability**: Guaranteed or Non-Guaranteed;
- **Number of People**: from 2 to 7;
- **Number of White People**: from 0 to 5;
- **Number of Afro People**: from 0 to 4;
- **Number of Asian People**: from 0 to 5;
- **Language**: English, European, or Asian;
- **Number of Females**: from 0 to 4;
- **Number of Males**: from 0 to 5;

TABLE I: Main conditions presence in WASD. Num. Vids refers to the amount of videos that fulfill that condition and Hrs refers to Hours.

Condition	Presence	
	Num. Vids	Total (Hrs)
Face Availability	3004	23.2
Speech Overlap	932	7.02
Delayed Speech	551	4.49
Facial Occlusion	743	6.02
Human Voice as Background Noise	734	6.23
Surveillance Settings	720	6.00

- **Frames per Second (FPS)**: from 10 to 30;
- **Image Size**: Variable;
- **Video Location**: Indoor or Outdoor;
- **Body-Image Proportion**: Variable;
- **Head-Body Proportion**: Variable;
- **Speaking Percentage**: Variable;
- **Speaking Overlap**: Variable.

All features with predefined admissible values were attributed by human assessment. Although FPS has various admissible values, they mainly range from 24-30. All other features are continuous values, thus not having a strict set of possibilities: Head-Body Proportion refers to the proportion of head area (bounding box) relative to body area, with similar analogy for Body-Image Proportion; and Speaking Percentage and Speaking Overlap is the number of frames with talking and simultaneous talking, respectively. Regarding Proximity to Camera, we consider three values relating to the visible body area and subject proximity to camera. The CSV containing all the information for each WASD video is available at https://github.com/Tiago-Roxo/WASD. Regarding the considered languages, we group them as follows:

- **English**: English (USA and UK);
- **European**: Croatian, Dutch, French, German, Italian, Portuguese, Russian, and Spanish;
- **Asian**: Chinese, Japanese, Korean, Pakistanese, and Vietnamese.

The distribution of the main conditions of WASD categories is presented in Table I. The distribution shows that most conditions linked to specific categories are around 1/6 of the total dataset duration (30 hours). The exceptions are: 1) *Face Availability* (around 5/6 of the dataset), with non-available face (1/6) being linked to SS; 2) *Speech Overlap*, slightly over 1/6 since there are videos from other categories also containing overlap of communication (derived from normal/spontaneous interactions); and 3) *Delayed Speech* is not 1/6 of total dataset duration since it shares the "time" of Speech Impairment category with *Speech Overlap* (*i.e.*, part of the videos for this category have *Speech Overlap*, while others have *Delayed Speech* with/without *Speech Overlap*). For a complete description of the existent conditions per video we refer to the metadata CSV in our GitHub.

E. Cross-Domain Evaluation

To complement the cross-domain results of Table II from the main paper, we also assess the performance of models in



(a) Optimal Conditions.



(b) Speech Impairment.



(c) Face Occlusion.



(d) Human Voice Noise.



(e) Surveillance Settings.

Fig. 3: Different examples of the considered scenarios for WASD categories.

AVA-ActiveSpeaker, when trained in WASD, in Table II. The results show that, overall, training with AVA-ActiveSpeaker data translates into better results in WASD domains, relative to the inverse settings, meaning that the varying conditions of WASD (*e.g.*, facial occlusion, speech overlap/impairment) do not translate into better performance in more cooperative settings (AVA-ActiveSpeaker). When training in these cooperative conditions (AVA-ActiveSpeaker), models tend to better convey the main characteristics of active speaker detection in more challenging settings (WASD), although its performance in degrading settings is still subpar (columns Human Voice Noise and Surveillance Settings of Table III, from main paper).

We also assess the robustness of Light-ASD and TalkNet in Columbia, when trained in WASD and AVA-ActiveSpeaker, and compare them with the reported results in the Light-ASD paper, in Table III. The results show that neither AVA-ActiveSpeaker nor WASD data is more suited to prepare models to reliably perform in Columbia settings. Columbia data contains one hour and a half of press conference data, without significant data perturbation (*e.g.*, speech overlap, no face access), with subjects facing forward, which is an even more cooperative setting than AVA-ActiveSpeaker (which contains delayed audio and varying image/audio quality, derived from movie settings). As such, training in challenging ASD data may bias models to particular features that hinder its performance in cooperative conditions, although WASD is able to prepare models more evenly to deal with such settings (*i.e.*, AVA-training leads to a bigger difference in Columbia performance between models than WASD-training).

TABLE II: Cross-domain assessment of models when trained in AVA-ActiveSpeaker (AVA) and WASD.

	Trained AVA		**Trained WASD**	
	AVA	**WASD**	**WASD**	**AVA**
ASC [1]	83.6	74.6	85.7	62.3
MAAS [2]	82.0	70.7	86.4	63.3
ASDNet [5]	91.1	79.2	92.0	74.5
TalkNet [9]	91.8	85.0	92.3	75.2
TS-TalkNet [4]	92.7	85.7	93.1	76.3
Light-ASD [7]	93.4	86.2	93.7	78.6

REFERENCES

[1] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 3

TABLE III: Comparison of F1-Score (%) on the Columbia dataset. Models with $\star$ are trained in WASD, while the remaining are trained in AVA-ActiveSpeaker. Models with $*$ are the reported results from our experiments, while the first three results are from Light-ASD reported experiments.

Model	Speaker					
	Bell	**Boll**	**Lieb**	**Long**	**Sick**	**Avg**
TalkNet [9]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [10]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [7]	82.7	75.7	87.0	74.5	85.4	81.1
TalkNet$*$	61.3	32.6	31.1	52.3	75.6	50.6
Light-ASD$*$	82.5	61.2	65.6	73.4	73.1	71.1
TalkNet$\star$	68.7	37.0	61.9	84.8	72.2	64.9
Light-ASD$\star$	89.4	25.6	72.0	86.9	67.5	68.3

[2] Juan Léon Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 3

[3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1

[4] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li. Target active speaker detection with audio-visual cues. In *Proc. Interspeech*, 2023. 3

[5] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 3

[6] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 1

[7] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023. 3, 4

[8] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1

[9] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 3, 4

[10] Xizi Wang, Feng Cheng, Gedas Bertasius, and David Crandall. Loconet: Long-short context network for active speaker detection. *arXiv preprint arXiv:2301.08237*, 2023. 4

[11] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 1

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 13, 2022, accepted May 27, 2022, date of publication June 2, 2022, date of current version June 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3179692

Predicting CVSS Metric via Description Interpretation

JOANA CABRAL COSTA, TIAGO ROXO, (Member, IEEE), JOÃO B. F. SEQUEIROS, HUGO PROENÇA, (Senior Member, IEEE), AND PEDRO R. M. INÁCIO, (Senior Member, IEEE)

Department of Computer Science, Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal

Corresponding author: Joana Cabral Costa (joana.cabral.costa@ubi.pt)

This work was supported in part by the Fundação para a Ciência e a Tecnologia (FCT)/Programa Operacional Temático Competitividade e Internacionalização (COMPETE)/Fundo Europeu de Desenvolvimento Regional (FEDER) under the scope of Project SECURIoTESIGN under Grant POCI-01-0145-FEDER-030657; in part by the Portuguese FCT/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through National Funds and, when applicable, co-funded by EU funds under Project UIDB/50008/2020; in part by the FCT Doctoral Grant SFRH/BD/133838/2017, Grant 2020.09847.BD, and Grant 2021.04905.BD; in part by the C4—Competence Center in Cloud Computing co-financed by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica, Programas Integrados de Investigação Científica e esenvolvimento Tecnológico (IC&DT) under Project CENTRO-01-0145-FEDER-000019.

ABSTRACT Cybercrime affects companies worldwide, costing millions of dollars annually. The constant increase of threats and vulnerabilities raises the need to handle vulnerabilities in a prioritized manner. This prioritization can be achieved through Common Vulnerability Scoring System (CVSS), typically used to assign a score to a vulnerability. However, there is a temporal mismatch between the vulnerability finding and score assignment, which motivates the development of approaches to aid in this aspect. We explore the use of Natural Language Processing (NLP) models in CVSS score prediction given vulnerability descriptions. We start by creating a vulnerability dataset from the National Vulnerability Database (NVD). Then, we combine text pre-processing and vocabulary addition to improve the model accuracy and interpret its prediction reasoning by assessing word importance, via Shapley values. Experiments show that the combination of *Lemmatization* and *5,000-word addition* is optimal for DistilBERT, the outperforming model in our experiments of the NLP methods, achieving state-of-the-art results. Furthermore, specific events (such as an attack on a known software) tend to influence model prediction, which may hinder CVSS prediction. Combining *Lemmatization* with vocabulary addition mitigates this effect, contributing to increased accuracy. Finally, binary classes benefit the most from pre-processing techniques, particularly when one class is much more prominent than the other. Our work demonstrates that DistilBERT is a state-of-the-art model for CVSS prediction, demonstrating the applicability of deep learning approaches to aid in vulnerability handling. The code and data are available at https://github.com/Joana-Cabral/CVSS_Prediction.

INDEX TERMS Common vulnerability scoring system, deep learning, interpretability, natural language processing, security.

I. INTRODUCTION

Cyber threats force companies to increase their investments in security, which resulted in a $170 billion security aspects related market in 2015 [1]. These threats impact 556 million people annually, costing $3 trillion worldwide, with an expected increase to $10.5 trillion by 2025 [2]. Additionally, there was an increase of vulnerability entries in VulDB, with 61 new daily entries, in 2021, relative to the 41 reported in 2016 [3]. This tendency provides a clear picture of the increased risk of threats and cybercrime, raising concern among Information Technology (IT) administrators, which often lack the resources to handle all incoming threats [4]. Given this context, there is an inherent need to define which vulnerabilities should be tackled first.

To aid in the prioritization of vulnerability handling, experts typically use the Common Vulnerability Scoring System (CVSS) [5], a *de facto* standard, to accurately assign a score to a vulnerability. New vulnerability entries are enumerated via Common Vulnerability Enumeration

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

This work is licensed under a Creative Commons Attribution 4.0 Licence. For more information, see https://creativecommons.org/licenses/by/4.0/

(CVE) [6], with a unique identifier, description, and CVSS Base score metrics, the latter specified by the National Vulnerability Database (NVD).

The score metric assignment is performed manually from vulnerability description analysis, for which vendors do not always provide enough detail [7] for experts to accurately create these scores. Furthermore, some CVSS metrics are subjective [8], heavily relying on the previous experience at assigning CVSS metrics. The inherent problems of this process are exacerbated by the temporal mismatch of CVSS metric assignment and vulnerability finding: 19 days to populate a vulnerability with the respective CVSS and six days to find a new one [9]. Therefore, to reduce the time/cost spent while also mitigating the subjective aspect of score assignment, we explore the use of a deep learning approach to predict the CVSS metrics based on the vulnerability description.

We start by obtaining the vulnerabilities descriptions and respective CVSS metrics using the NVD Application Programming Interface (API). The collected data is processed for the most recent version of CVSS (version 3) and serves as input for the deep learning approach. We select the DistilBERT for sequence classification given its outperformance, in the created dataset, over other state-of-the-art Natural Language Processing (NLP) models. Since the vulnerability descriptions contain technical expressions and have reduced length size, we assess the effect of text pre-processing techniques and vocabulary addition. Our results show that text pre-processing improves the baseline model accuracy, exhibiting incremental performance with vocabulary addition.

One drawback of using a deep learning approach is that the reasoning behind their outputs is not easily disclosed. To overcome this limitation, we use the Shapley value [10], a game-theoretic approach to explain machine learning outputs, to perceive the correlation between description words and the predicted CVSS metric. This process allows us to understand the importance of each word towards the CVSS metric prediction, assessing their importance variance with text pre-processing and vocabulary addition.

The main contributions of our work are summarized as follows:

- We present a vulnerability dataset, derived from NVD data, with vulnerability descriptions and CVSS (version 3) metrics;
- We demonstrate the applicability of deep learning approaches to predict CVSS metrics, in combination with text pre-processing and vocabulary addition, achieving state-of-the-art results;
- We confer interpretability to model prediction by analyzing the importance of word descriptions, via Shapley value.

The remainder of this paper is organized as follows: Section II summarizes the most relevant CVSS-based works; Section III describes the methodologies used; Section IV describes the vulnerability dataset; and Section V discusses the results obtained. Finally, the main conclusions and future work are presented in Section VI.

II. RELATED WORK

A. CVSS APPLICABILITY

CVSS has been extensively analyzed and applied to multiple domains to prioritize or estimate security risks. Younis and Malaiya [11] compared the CVSS base metrics and the Microsoft rating system, declaring that both measures have a very high false-positive rate, with CVSS significantly affected by the software type. Joh [12] concluded that most vulnerabilities are compromised due to no authentication required systems, by analyzing the CVSS base scores for vulnerabilities of currently supported Windows operating systems, suggesting the addition of an authentication process in every system. CVSS base metrics have been used to assess cybersecurity risks in IT systems [4], using the risk formula, and calculating risk probability and impact. The same study reported that an identification of security properties in the early stages of development positively impacts the security of the systems. In the same context, Wirtz and Heisel [13] proposed a semi-automatic method to estimate security risks in the early stages of software development, using CVSS formulas to assess the threat severity. Since CVSS has already demonstrated its validity in typical IT systems, it was also adapted to calculate vulnerabilities regarding hybrid IT and IoT systems [14], [15] accurately. Following this idea, Mishra and Singh [16] proposed a taxonomy for Cloud-specific vulnerabilities, using the CVSS score to represent each major Cloud vulnerability severity. Finally, a guide for applying CVSS to medical devices was also proposed [17], consisting of questions that identify a value for a specific CVSS metric.

B. CVSS AND ARTIFICIAL INTELLIGENCE

The combination of Artificial Intelligence techniques and CVSS scores of individual vulnerabilities has also been reported. Sheehan *et al.* [18] proposed using Bayesian Networks to identify connected and autonomous vehicle cyber risks, using CVSS scores to predict knowledge gaps or potential new cyber vulnerabilities. Furthermore, Frigault *et al.* [19] employed Bayesian Networks and Attack Graphs to measure network security, using the CVSS scores as probabilities and considering metric values of each vulnerability to be independent. However, applying Bayesian Networks to assess CVSS scores has limitations [20], leading to the proposal of an approach that considers the dependency relationships between the CVSS base metrics, combining scores into three aspects: probability, effort, and skill. Allouzi and Khan [21] proposed using the Markov Chain to compute the probability distribution of Internet of Medical Things security threats, using CVSS scores to assign severity to the acknowledged vulnerabilities. One first attempt to predict CVSS final scores was made through the employment of fuzzy systems [22], outperforming Support Vector Machine (SVM) and Random-Forest. In this context,

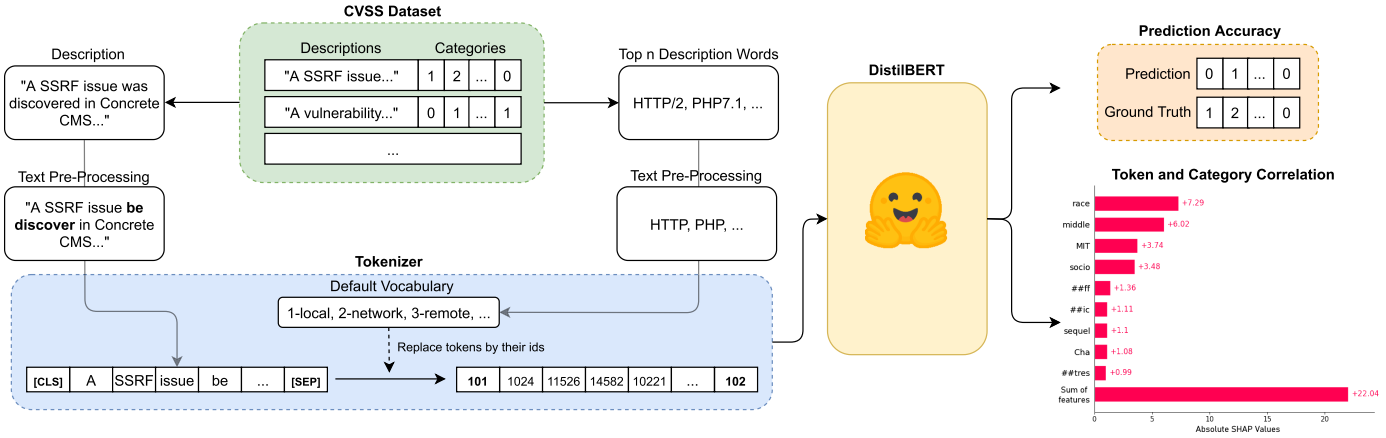


FIGURE 1. Overview of the methodology used to assess DistilBERT performance in vulnerability detection, using CVSS data descriptions and categories. We evaluate the model performance by varying two key aspects: 1) text pre-processing approaches; and 2) vocabulary addition. Furthermore, we evaluate the correlation of tokens and category, via Shapley value, to assess the tokens more influential towards each category prediction.

fuzzy CVSS [23] was used to calculate the final severity score for vulnerabilities, employing fuzzy theory to reduce the error rate. To predict CVSS values for base metrics, Elbaz *et al.* [24] propose a linear regression model, using a bag of words approach, with the removal of irrelevant words.

C. CVSS AND DEEP LEARNING

Deep learning is also known for its effectiveness in solving complex problems, with the drawback of time-costly training. Therefore, to resemble security experts decision-making [25], the usage of Neural Networks was proposed, automatically providing a vulnerability report through CVSS metrics. Deep reinforcement learning was also used to assess the cyber-physical security of electric power systems [26], which adapted CVSS to estimate the complexity of attack path. As a result, CVSS base metrics have been adopted as the guide for identifying and prioritizing threats among multiple systems. This indicates that correctly and swiftly predicting the metrics for CVSS is a valuable effort.

Sahin and Tosun [27] concluded that Long Short Term Memory (LSTM) was the most accurate model to predict CVSS final scores, when compared with Convolutional Neural Networks (CNN) and XGBoost. The two previously presented approaches gathered data from Open Source Vulnerability Database (OSVDB) and NVD, respectively, to train their models. Alternatively, Twitter discussions [28], with NVD as ground truth for CVSS scores, were fed to a Graph Convolutional Network with Attention-based input Embedding to predict the CVE severity scores. However, predicting CVSS final scores does not provide any insight to the experts about the values for the CVSS metrics.

D. VULNERABILITY INTERPRETABILITY

The analysis and interpretation of vulnerability descriptions is also reported in the literature. An empirical study based on the NVD vulnerability descriptions [29] concluded that information about the *asset*, *attack*, and *vulnerability* type is relevant to increase vulnerability scoring accuracy. Another work used the Local Interpretable Model-Agnostic (LIME) framework to explain the vulnerability descriptions [30], providing relevant words for a small number of vulnerabilities.

To the best of our knowledge, the work presented herein is the first to combine Deep Learning and NLP approaches to extract information from vulnerability descriptions and output CVSS metrics, while using interpretability to assess model predictions.

III. METHODOLOGY

The methodology used in our experiments is displayed in Fig. 1. We start by creating a CVSS dataset, using information from the NVD. Then, we vary two major performance-related aspects: 1) text pre-processing; and 2) vocabulary addition. Finally, we evaluate model accuracy and assess token correlation with category prediction, using Shapley value.

A. MODEL DETAILS AND EVALUATION METRICS

We used the following models in our experiments: BERT [31], DistilBERT [32], RoBERTa [33], ALBERT [34], and DeBERTa [35]. Our reasoning for model choice is linked to the importance of BERT for the NLP area. It is one of the most used models in NLP, in a variety of tasks, with proven quality. Then, we opted to choose other variations of BERT to assess what is the better model for CVSS metric prediction. Specifically, we choose ALBERT and DistilBERT for having fewer parameters than BERT and RoBERTa and DeBERTa for having more parameters than BERT. The chosen models belong to the BERT family while having specific characteristics, such as the number of parameters. As such, our work focused on finding the best performing state-of-the-art NLP models for CVSS metrics prediction.

We finetune each model following the authors' methodology: regarding the *learning rate*, RoBERTa was set to 1.5×10^{-5}, DistilBERT was set to 5×10^{-5}, while BERT,

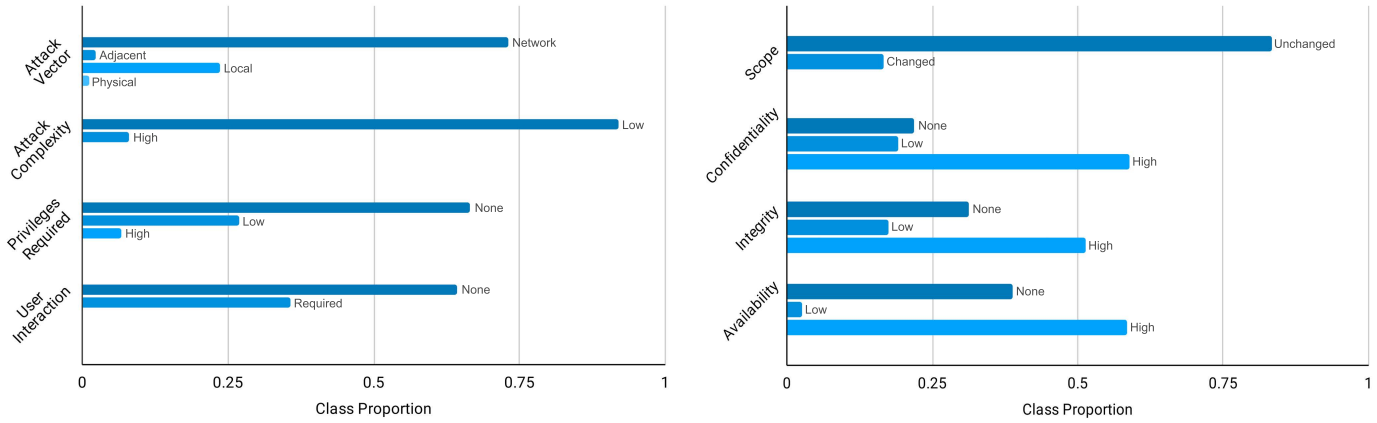


FIGURE 2. Bar plots of the eight categories analyzed of CVSS version 3, linked to vulnerability assessment, of the vulnerability dataset used. Each category displays the associated classes and respective class prior.

ALBERT, and DeBERTa have all been set to 3×10^{-5}; for the *number of training epochs*, RoBERTa was trained for 2 epochs, DeBERTa for 10, and BERT, ALBERT, and DistilBERT for 3; regarding *batch size*, we used 8 for ALBERT and DistilBERT, and 4 for BERT, RoBERTa, and DeBERTa; finally, RoBERTa has a *weight decay* of 0.01, while the remaining models have the default value (0). We use the default losses and architectures of each model, from Hugging Face [36]. To obtain category classification, we use a *PyTorch Softmax* layer [37] on the model output.

To compare the performance of each model, we use the accuracy, F1 score, and balanced accuracy from the *scikit-learn* library [38]. To compare our results with state-of-the-art for CVSS metric inference, we use the accuracy metric.

B. TEXT PROCESSING AND VOCABULARY SELECTION

To assess the contribution of each word to the classification of the considered categories (discussed in section IV), we start by processing vulnerability descriptions. We use two pre-processing methods, namely, *Lemmatization* and *Stemming*. Finally, we tokenize the text to input to the model, evaluating its accuracy based on the pre-processing approach. Both text pre-processing approaches use Natural Language Toolkit (NLTK) methods [39], while tokenization is achieved using Transformers library, from Hugging Face [36]. We choose *Lemmatization* and *Stemming*, given their wide use as text pre-processing approaches in the NLP area. By using *Lemmatization* and *Stemming*, we intend to process text to maintain as much relevant data as possible while ignoring noisy data. This is achieved by ignoring variants of words that have the same ''base''. In the case of *Stemming* is the same stem, while in *Lemmatization* is the same lemma.

In our experiments, we also evaluate the effect of vocabulary addition. Moreover, we also assess this effect in conjunction with the best performing text pre-processing approach. We evaluate the accuracy of the used model when adding 5,000, 10,000, and 25,000 words to the default vocabulary of the tokenizer. To select the added words, we order them by frequency of appearance in the descriptions, choosing the top n words. To avoid redundancy, we only consider words that appear exclusively in the description and not in the default vocabulary.

Given the existence of software versions and code snippets in some data descriptions, we use regular expressions to filter digits and special characters. This approach reduces the ''noise'' of vocabulary addition, since this filtered data is not relevant to category classification and could potentially dissipate the importance of relevant added words.

C. SHAPLEY VALUE

Deep learning models have shown high performance in multiple tasks while providing little to no explanation for the reasoning for model prediction. To tackle this issue, we use Shapley value, an interpretability technique that allows us to interpret the reasoning of the model when providing predictions. The Shapley value, coined by Shapley in 1953 [40], is a cooperative game theory-based method used for assigning payouts to players, depending on their contribution towards the total payout. In the machine-learning context, the Shapley value is used to evaluate how each feature (player) of a given instance contributed (assigning payout) towards the model prediction of the instance (total payout).

The use of Shapley value in our experiments is linked to our interest in analyzing how each word contributed to category classification. For categories with more than n classes, and n higher than 2, we perform n Shapley value analysis, each considering a class versus the remaining classes of the category. The considered class is given the value 1, with the remaining receiving the value 0. If a word contributes positively, it means that it influences the considered class. The higher the absolute Shapley value is, the higher the feature influence. We use the SHapley Additive exPlanations (SHAP) framework [41] and the Explainer model, from a publicly available implementation in [42].

TABLE 1. Percentage of class prior of the eight vulnerability related categories for all dataset, train, and test set. Class prior displays the likelihood of an outcome in the dataset and each subset.

Category	Class Prior (%)		
	Train	Test	Dataset
Attack Vector			
Network	73.04	73.57	73.14
Adjacent	2.21	2.14	2.20
Local	23.62	23.18	23.53
Physical	1.13	1.11	1.13
Attack Complexity			
High	8.03	7.74	7.97
Low	91.97	92.26	92.03
Privileges Required			
High	6.70	6.32	6.62
Low	26.79	26.90	26.81
None	66.51	66.78	66.57
User Interaction			
Required	35.80	35.37	35.71
None	64.20	64.63	64.29
Scope			
Changed	16.72	16.36	16.65
Unchanged	83.28	83.64	83.35
Confidentiality			
High	58.94	59.15	58.98
Low	19.37	18.65	19.23
None	21.69	22.20	21.79
Integrity			
High	51.27	51.36	51.29
Low	17.52	17.32	17.48
None	31.21	31.32	31.23
Availability			
High	58.49	58.85	58.56
Low	2.60	2.71	2.63
None	38.91	38.44	38.81

IV. VULNERABILITY DATASET

The vulnerability dataset is based on NVD information, a United States government repository of standards-based vulnerability management data. We obtain the information through their API, starting from index 0 to 152,000, representing data collected until April 2021. Finally, we process the collected data to retrieve vulnerability descriptions and the classes for each of the eight categories analyzed: *Attack Vector*, *Attack Complexity*, *Privileges Required*, *User Interaction*, *Scope*, *Confidentiality*, *Integrity*, and *Availability*. Based on the CVSS documentation, these classes are grouped into Exploitability metrics (Attack Vector, Attack Complexity, Privileges Required, and User Interaction), Scope, and Impact metrics (Confidentiality, Integrity, and Availability). Tables and Figures throughout this paper consider this grouping. A visual representation of class proportions, for each category, of our dataset is displayed in Fig. 2.

Though the collected data corresponds to 152,000 vulnerability descriptions and categories, we only consider descriptions related to version 3 of CVSS in this work. For this reason, the total number of instances in our dataset is 79,810. We divide them into train and test sets, composed of 63,848 and 15,962 instances, respectively, corresponding to a 0.2 test ratio. The average description length is 43.85 and 44.55 words for train and test split, respectively. Each set follows a similar proportion of classes, exhibited in Fig 2, whose analytical values are shown in Table 1. The dataset is publicly available for repeatability purposes, and it serves as a basis for other models to evaluate their performance and compare with the proposed methodology.

V. EXPERIMENTS

A. MODEL COMPARISON

We start by comparing the performance of five different NLP methods in the proposed dataset. The accuracy, F1 score, and balanced accuracy for each of the eight categories are presented in Table 2. The results suggest that DistilBERT is the outperforming model for all the categories, in all the considered metrics. The method with the worst performance is ALBERT, which has the least number of parameters (11M), while DeBERTa, BERT, and RoBERTa, with over 100M parameters, also have worse performance than DistilBERT (65M). Since we intend to assess the class inference, given a vulnerability description, the number of parameters may be linked to the performance variance. In this case, too few parameters (ALBERT) are insufficient for the model to learn, and too many leads to poorer fine-tuning. The similarity of various accuracy values between BERT, ALBERT, and DeBERTa, for different categories, can be explained by dataset imbalance. In these cases, the values displayed represent a scenario where the models opted to achieve higher accuracy by outputting the same value in every instance. Thus, in cases of dataset imbalance, the use of accuracy can be deceptive, justifying the use of other metrics such as balanced accuracy.

In this experiment, we use the default pretraining weights (provided by HuggingFace [36]) and training parameters of every model. The models used are typically applied/evaluated in tasks where the association of two sentences is analyzed (*e.g.*, GLUE [43]) or the aim is finding answers in a text, given a question (*e.g.*, SQuAD [44]). These types of tasks differ from predicting a category given a vulnerability description (the aim of this work), which may justify the underperformance of state-of-the-art methods in our experiments. Based on the obtained results, we selected DistilBERT for continuing the experiments involving the usage of Deep Learning.

B. TEXT PRE-PROCESSING

We assess the performance of DistilBERT, for all eight considered categories, regarding different text pre-processing approaches. We present our results, using balanced accuracy, in Table 3, with *Baseline* referring to the condition where no pre-processing approach is used.

When comparing category-related performance variance, we observe that all categories benefit from pre-processing. Regarding processing-related performance variance, *Lemmatization* promotes better results than *Stemming*, for all categories. *Stemming* truncates words by chopping off letters from the end until the stem is reached. This is a more

TABLE 2. Model accuracy (Acc), F1 score, and Balanced Accuracy (BA) for each of the eight categories analyzed. The outperforming model for each metric and category is shown in bold.

Category	BERT			RoBERTa			ALBERT			DeBERTa			DistilBERT		
	Acc	F1	BA	Acc	F1	BA	Acc	F1	BA	Acc	F1	BA	Acc	F1	BA
Attack Vector	73.57	62.37	25.00	90.55	90.18	62.03	90.87	90.60	67.76	80.86	78.99	37.17	**91.04**	**90.85**	**70.60**
Attack Complexity	92.26	88.55	50.00	92.26	88.55	50.00	92.26	88.55	50.00	92.26	88.55	50.00	**95.16**	**94.32**	**70.74**
Privileges Required	66.78	53.48	33.33	66.86	83.49	67.61	66.78	53.48	33.33	66.86	63.65	41.15	**86.21**	**85.91**	**75.09**
User Interaction	64.63	50.75	50.00	82.73	92.57	91.26	64.63	50.75	50.00	82.73	82.85	81.90	**93.01**	**92.98**	**91.95**
Scope	83.64	76.19	50.00	96.02	95.91	90.20	83.64	76.19	50.00	83.64	76.19	50.00	**96.28**	**96.19**	**90.98**
Confidentiality	59.15	43.97	33.33	86.20	85.90	81.08	86.18	85.98	81.74	86.20	86.05	82.37	**86.29**	**86.16**	**82.67**
Integrity	61.52	50.10	55.84	86.93	86.85	84.41	51.36	34.85	33.33	51.37	34.87	33.34	**87.46**	**87.42**	**85.52**
Availability	58.85	43.61	33.33	83.50	87.89	67.49	58.85	43.61	33.33	83.50	82.47	57.54	**88.55**	**88.00**	**67.51**

TABLE 3. Category balanced accuracy of DistilBERT for baseline conditions (*Tokenization*), and using text pre-processing approaches (*Lemmatization* and *Stemming*). The outperforming approach for each category is shown in bold.

Category	Balanced Accuracy (%)		
	Baseline	Lemmatization	Stemming
Attack Vector	70.60	**71.03**	68.40
Attack Complexity	70.74	**71.45**	64.50
Privileges Required	75.09	**75.14**	73.83
User Interaction	91.95	**92.05**	91.75
Scope	90.98	**91.12**	90.40
Confidentiality	82.67	**83.12**	82.44
Integrity	85.52	**85.77**	84.98
Availability	67.51	**69.66**	68.69

crude approach than *Lemmatization*, which justifies the underperformance using this approach. Given the superiority displayed by *Lemmatization* over *Stemming*, this is the chosen pre-processing approach to use in the remaining experiments.

C. VOCABULARY ADDITION

We also evaluate the effect of vocabulary addition on prediction accuracy. Furthermore, we compare the vocabulary addition with its combination with a pre-processing approach. We display our results in Table 4.

Relative to the baseline, most variations of vocabulary addition translate into performance increase, for all categories. Regarding the vocabulary variations, 5,000-word addition was the condition with better results overall. This suggests that adding more words is beneficial to model accuracy improvement. However, subsequent vocabulary addition (10,000 and 25,000-word addition) does not promote incremental performance increase. Given that vocabulary addition is linked to word frequency in the description, adding more words may disperse the model attention towards less relevant words, hindering its performance. This aspect is more noticeable when 25,000-word addition has worse performance than baseline (*e.g.*, Attack Vector, Attack Complexity). For all categories, 25,000-word addition does not generally translate into performance improvement relative to 5,000-word addition, suggesting the existence of word importance redundancy with this approach.

Regarding the combination of vocabulary addition with *Lemmatization*, we observe that this approach generally improves the balanced accuracy, relative to vocabulary addition alone, for most vocabulary variations. This suggests that word importance may vary with processing approaches, which corroborates the importance of text pre-processing, even in the context of vocabulary addition. The results suggest that *5,000-word addition* with *Lemmatization* is the best approach for overall category prediction, exhibiting the importance of text processing and pertinent word addition in description-based classification.

D. STATE-OF-THE-ART COMPARISON

We compare DistilBERT, and its combination with pre-processing and vocabulary addition, with the state-of-the-art. To the best of our knowledge, only Ebalz *et al.* [24] evaluates class prediction accuracy in version 3 of CVSS. To compare our results with them, we also display the accuracy of *Baseline* and *5,000-word addition with Lemmatization*, whose balanced accuracy is presented in Table 4. Since the authors presented their results in a bar plot, not displaying the analytical values, we register the rounded values observed in said plot. We display the state-of-the-art comparison in Table 5.

Ebalz *et al.* use a bag of words approach, with the removal of irrelevant words, to input a regression model. Using DistilBERT, a deep learning approach, in conjunction with text pre-processing and vocabulary addition, we obtain substantial accuracy improvements in the majority of categories. The categories where Ebalz's approach was closer to ours were *Attack Complexity*, *User Interaction*, and *Scope*, which could be linked to these categories being two-classed. In these cases, the regression model used by Ebalz *et al.* can compete with deep learning approaches. However, for the remaining categories, with over two classes, the performance disparity is substantially larger, with up to a 28% accuracy increase. Furthermore, using the text pre-processing approach and adding vocabulary promotes an accuracy increase of DistilBERT, further enhancing its performance. The results suggest that DistilBERT is a state-of-the-art approach for vulnerability category prediction, particularly for multi-class categories.

TABLE 4. Category balanced accuracy of DistilBERT for baseline conditions (*Tokenization*), and with different vocabulary addition, assessing the effect of *Lemmatization*. *Base*, in each vocabulary column, refers to the vocabulary addition with *Tokenization*, without text pre-processing. The expression *w/ Lemm* refers to *Lemmatization* combination with vocabulary addition. The outperforming approach for each category is shown in bold.

Category	Balanced Accuracy (%)						
	Baseline	5,000 Words		10,000 Words		25,000 Words	
		Base	w/ Lemm	Base	w/ Lemm	Base	w/ Lemm
Attack Vector	70.60	70.79	**71.26**	71.05	71.19	68.66	69.83
Attack Complexity	70.74	71.42	**72.92**	68.73	68.78	65.59	68.06
Privileges Required	75.09	75.44	**75.81**	75.14	74.95	74.76	74.87
User Interaction	91.95	92.11	92.40	**92.48**	92.09	91.81	91.96
Scope	90.98	**91.60**	91.49	90.99	91.01	90.70	91.03
Confidentiality	82.67	82.89	**83.18**	82.82	82.66	82.73	82.38
Integrity	85.52	85.81	**85.97**	85.53	85.56	85.13	85.36
Availability	67.51	67.73	**69.35**	69.18	68.56	67.80	68.32

TABLE 5. Category accuracy of DistilBERT, DistilBERT-Enhanced (DistilBERT-E) and Ebalz's work [24]. DistilBERT-Enhanced refers to DistilBERT using *Lemmatization* and *5,000-word addition*. The outperforming approach for each category is shown in bold.

Category	Accuracy (%)		
	DistilBERT	DistilBERT-E	Ebalz [24]
Attack Vector	91.04	**91.41**	78.00
Attack Complexity	95.16	**95.20**	95.00
Privileges Required	86.21	**86.42**	79.00
User Interaction	93.01	**93.33**	89.00
Scope	96.28	**96.40**	96.00
Confidentiality	86.29	**86.71**	69.00
Integrity	87.46	**87.61**	63.00
Availability	88.55	**88.81**	60.00

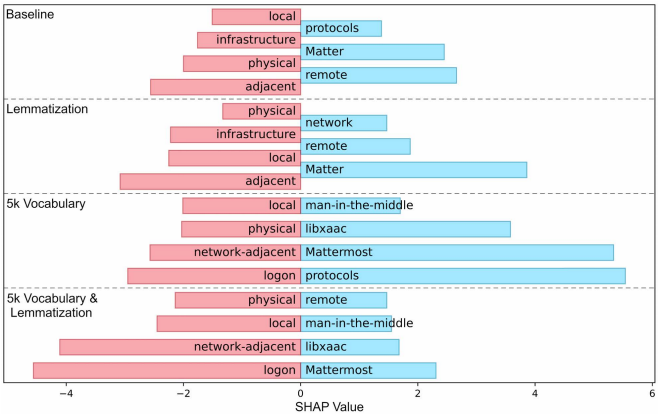
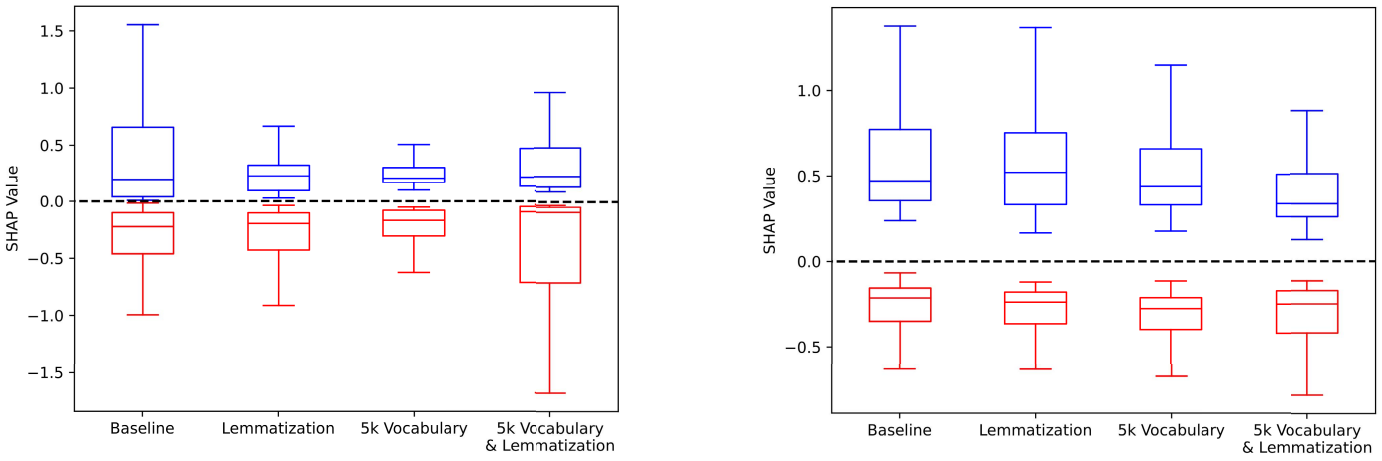


FIGURE 3. Bar plots of the SHAP value for Baseline, *Lemmatization*, 5,000-word addition (*5k Vocabulary*), and 5,000-word addition with *Lemmatization* (*5k Vocabulary & Lemmatization*) for the category *Attack Vector*, regarding *Network* class.

E. INTERPRETING CATEGORY CLASSIFICATION

We assess word importance in two distinct scenarios: 1) comparing the most relevant words, using different processing techniques, for a given category; and 2) assessing the variance of word importance towards/against binary and multi-class category prediction, given different processing techniques. In the first scenario, we compare word importance variance with text pre-processing and vocabulary addition in DistilBERT. Given the overall superiority of Lemmatization and 5,000-word addition (Table 4), these are the chosen approaches. We consider the four stages for comparison: 1) Baseline; 2) Lemmatization; 3) 5,000-word addition; and 4) 5,000-word addition with Lemmatization. For the remaining experiments, we will refer to each *word* of a description as a *token* to accurately represent the word translated into the tokenizer vocabulary. We evaluate token importance for the category *Attack Vector*, regarding the *Network* class. In this case, *Network* has a value of 1, and the remaining three classes have the value 0. Tokens with positive Shapley value influence *Network* classification, while negative ones are more relevant to the other three classes.

The results show a variance in token importance with text pre-processing and vocabulary addition. Starting in the **Baseline**, with no processing or vocabulary addition, *protocols*, *Matter*, and *remote* are tokens that, when in a description, influence the classification of the category towards Network. There is some logic behind said importance, given that *remote* and *protocols* are linked to network-related activities. The influence of *Matter* is linked to Mattermost, an open-source chat service, which was the target of multiple attacks. This shows that token importance might be influenced by specific network-related events. When we analyze tokens more associated with other classes (negative Shapley value), we observe that these are closely related to class definition (Local, Physical, and Adjacent) or associated with it (*infrastructure*). Adding **Lemmatization**, we observe the same tendency for tokens influential towards other classes but with increased importance. Furthermore, tokens linked to Network classification lose importance, aside from the specific network-related event of baseline (*Matter*). This suggests that token descriptions are more interpretably linked in not classifying Network than towards it, which could be due to class imbalance. Network is over 70% of Attack Vector classes, making it harder to distinguish tokens clearly associated with it, thus justifying the Lemmatization results. The addition of vocabulary (**5k Vocabulary**) heavily influences category classification, with new tokens being associated with Network classification: *libxaac*, *Mattermost*, and *man-in-the-middle*. *Libxaac* is an Android library with



(a) Binary Class Data Distribution

(b) Multi-Class Data Distribution

FIGURE 4. **Boxplots summarizing the effect of applying text pre-processing techniques (Lemmatization) and vocabulary addition (5,000-word addition) for (a) binary class (*Attack Complexity*, *User Interaction*, and *Scope*) and (b) multiple class problems.**

reported out-of-bound reading/writing errors, while *man-in-the-middle* is a type of network attack. The importance of these tokens is linked to specific network-related events (attacks, errors), which was also observed in the baseline. *Protocols* also increases in importance towards Network classification, which could be linked to their association with the added vocabulary. This shows that vocabulary addition shifts the focus of token importance heavily towards specific events, for Network classification. *Network-adjacent* (added by vocabulary addition) also gains importance in classifying other classes, given its relevance to dissociate Network from Adjacent. Complementing vocabulary addition with Lemmatization (**5k Vocabulary & Lemmatization**) diminishes the importance of tokens closely linked to Network (positive Shapley value), resurging the tendency observed with Lemmatization alone. The reduced importance of specific network-related events also greatly decreased token importance associated with it (*protocols*). Furthermore, the influence of added vocabulary was enhanced in *logon* (closely related to classes other than Network) and *network-adjacent*, while keeping high importance of tokens associated with other classes definition (*physical* and *local*). This result suggests that Lemmatization is necessary to obtain more coherent/explainable token importance, which ultimately translates into better model performance (as shown in Table 4).

The second considered scenario relates token importance when considering binary (*Attack Complexity*, *User Interaction*, and *Scope*) and multi-class categories, for the same processing approaches of the first scenario. For all categories, the highest proportion class per category was associated with the value 1, with the remaining being associated with 0. Fig. 4 displays the boxplots for the two cases considered, showing the data distribution (ignoring wildcard cases).

The analysis of binary boxplots indicates that using Lemmatization and vocabulary addition promotes a decrease in token importance variance in both towards (positive Shapley value) and against (negative Shapley value) the highest class. However, combining vocabulary addition with Lemmatization increases token importance variance, particularly for negative values. This translates into increased importance of tokens to categorize the least represented class. If almost all descriptions relate to a specific class, it may be more beneficial/discriminative to focus on tokens linked to the underrepresented class, which is the approach of the model in this case.

Analyzing multi-class boxplots shows that the variance of negative Shapley value remains nearly constant throughout the various text pre-processing methods. Comparatively to the binary classes, negative Shapley value refers to various classes and not simply one, which justifies the (low) variance observed for these cases. Relative to positive Shapley value, using vocabulary addition and its combination with Lemmatization tends to reduce the variance of token importance, achieving a similar variance to negative Shapley value tokens. In multi-class prediction, even when one class is more prevalent than others, the existence of tokens closely linked to specific categories is not as likely as in binary class prediction. For this reason, reducing the overall importance towards specific token importance classification translates into better results.

VI. CONCLUSION

The increasing number of threats and vulnerabilities in IT systems surpass the capability of professionals to handle them, potentially leading to company prejudice. This raises the need to prioritize vulnerabilities, typically achieved through CVSS metrics, via manual vulnerability description analysis. In this paper, we present a vulnerability dataset, from NVD data, and analyze the applicability of deep learning approaches, namely NLP methods, to aid in CVSS metric prediction via description interpretation. In our experiments, we also assess the importance of text processing and vocabulary addition in metric prediction while interpreting

it via Shapley value. Our results show that DistilBERT is a state-of-the-art model for CVSS metric prediction, with increased performance when combined with Lemmatization (text pre-processing) and 5,000 word-addition. Furthermore, this combination mitigates the effect of specific events in category prediction and leads to weighted word importance, particularly for binary categories, contributing to increased model accuracy. The presented dataset and model experiments serve as a comparable basis for future works in CVSS metric prediction, applicable for vulnerability handling/prioritization, which leads to increased usefulness and accuracy of the metric, benefiting system security and operational effectiveness.

REFERENCES

[1] P. Boden. (2016). *The Emerging Era of Cyber Defense and Cybercrime*. Accessed: Jul. 29, 2021. [Online]. Available: https://www.microsoft.com/security/blog/2016/01/27/the-emerging-era-of-cyber-defense-and-cybercrime/

[2] S. Morgan. (2020). *Cybercrime to Cost the World $10.5 Trillion Annually by 2025*. Accessed: Jul. 29, 2021. [Online]. Available: https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/

[3] vuldb.Com. (2021). *Vulnerability Database*. Accessed: Jul. 28, 2021. [Online]. Available: https://vuldb.com/?

[4] M. U. Aksu, M. H. Dilek, E. I. Tatli, K. Bicakci, H. I. Dirik, M. U. Demirezen, and T. Aykir, "A quantitative CVSS-based cyber security risk assessment methodology for IT systems," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–8.

[5] P. Mell, K. Scarfone, and S. Romanosky, "Common vulnerability scoring system," *IEEE Secur. Privacy*, vol. 4, no. 6, pp. 85–89, Nov./Dec. 2006.

[6] D. E. Mann and S. M. Christey, "Towards a common enumeration of vulnerabilities," in *Proc. 2nd Workshop Res. Secur. Vulnerability Databases*, West Lafayette, IN, USA: Purdue Univ., 1999, pp. 1–13.

[7] N. I. of Standards and Technology. (2021). *NVD—Vulnerability Metrics*. Accessed: Jul. 28, 2021. [Online]. Available: https://nvd.nist.gov/vuln-metrics/cvss

[8] P. Johnson, R. Lagerstrom, M. Ekstedt, and U. Franke, "Can the common vulnerability scoring system be trusted? A Bayesian analysis," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 6, pp. 1002–1015, Dec. 2018.

[9] A. Feutrill, D. Ranathunga, Y. Yarom, and M. Roughan, "The effect of common vulnerability scoring system metrics on vulnerability exploit delay," in *Proc. 6th Int. Symp. Comput. Netw. (CANDAR)*, Nov. 2018, pp. 1–10.

[10] L. S. Shapley, *A Value for N-Person Games*. Princeton, NJ, USA: Princeton Univ. Press, 2016, ch. 17.

[11] A. A. Younis and Y. K. Malaiya, "Comparing and evaluating CVSS base metrics and Microsoft rating system," in *Proc. IEEE Int. Conf. Softw. Qual., Rel. Secur.*, Aug. 2015, pp. 252–261.

[12] H. Joh, "Software risk assessment for windows operating systems with respect to CVSS," *Eur. J. Eng. Technol. Res.*, vol. 4, no. 11, pp. 41–45, Nov. 2019.

[13] R. Wirtz and M. Heisel, "CVSS-based estimation and prioritization for security risks," in *Proc. 14th Int. Conf. Eval. Novel Approaches Softw. Eng.*, 2019, pp. 297–306.

[14] A. Ur-Rehman, I. Gondal, J. Kamruzzuman, and A. Jolfaei, "Vulnerability modelling for hybrid IT systems," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2019, pp. 1186–1191.

[15] A. Ur-Rehman, I. Gondal, J. Kamruzzaman, and A. Jolfaei, "Vulnerability modelling for hybrid industrial control system networks," *J. Grid Comput.*, vol. 18, no. 4, pp. 863–878, Dec. 2020.

[16] N. Mishra and R. Singh, "Taxonomy & analysis of cloud computing vulnerabilities through attack vector, CVSS and complexity parameter," in *Proc. Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*, vol. 1, Sep. 2019, pp. 1–8.

[17] M. P. Chase and S. M. C. Coley, "Rubric for applying CVSS to medical devices," MITRE Corp., McLean, VA, USA, Tech. Rep. HHSM-500-2012-00008I, Oct. 2020.

[18] B. Sheehan, F. Murphy, M. Mullins, and C. Ryan, "Connected and autonomous vehicles: A cyber-risk classification framework," *Transp. Res. A, Policy Pract.*, vol. 124, pp. 523–536, Jun. 2019.

[19] M. Frigault, L. Wang, S. Jajodia, and A. Singhal, "Measuring the overall network security by combining CVSS scores based on attack graphs and Bayesian networks," in *Network Security Metrics*. Cham, Switzerland: Springer, 2017, pp. 1–23.

[20] P. Cheng, L. Wang, S. Jajodia, and A. Singhal, "Refining CVSS-based network security metrics by examining the base scores," in *Network Security Metrics*. Cham, Switzerland: Springer, 2017, pp. 25–52.

[21] M. Ali Allouzi and J. I. Khan, "Identifying and modeling security threats for IoMT edge network using Markov chain and common vulnerability scoring system (CVSS)," 2021, *arXiv:2104.11580*.

[22] A. Khazaei, M. Ghasemzadeh, and V. Derhami, "An automatic method for CVSS score prediction using vulnerabilities description," *J. Intell. Fuzzy Syst.*, vol. 30, no. 1, pp. 89–96, Aug. 2015.

[23] K. Gencer and F. Başçiftçi, "The fuzzy common vulnerability scoring system (F-CVSS) based on a least squares approach with fuzzy logistic regression," *Egyptian Informat. J.*, vol. 22, no. 2, pp. 145–153, Jul. 2021.

[24] C. Elbaz, L. Rilling, and C. Morin, "Fighting N-day vulnerabilities with automated CVSS vector prediction at disclosure," in *Proc. 15th Int. Conf. Availability, Rel. Secur.*, Aug. 2020, pp. 1–10.

[25] A. Beck and S. Rass, "Using neural networks to aid CVSS risk aggregation—An empirically validated approach," *J. Innov. Digit. Ecosyst.*, vol. 3, no. 2, pp. 148–154, 2016.

[26] X. Liu, J. Ospina, and C. Konstantinou, "Deep reinforcement learning for cybersecurity assessment of wind integrated power systems," *IEEE Access*, vol. 8, pp. 208378–208394, 2020.

[27] S. E. Sahin and A. Tosun, "A conceptual replication on predicting the severity of software vulnerabilities," in *Proc. Eval. Assessment Softw. Eng.*, Apr. 2019, pp. 244–250.

[28] H. Chen, J. Liu, R. Liu, N. Park, and V. S. Subrahmanian, "VASE: A Twitter-based vulnerability analysis and score engine," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 976–981.

[29] L. Allodi, S. Banescu, H. Femmer, and K. Beckers, "Identifying relevant information cues for vulnerability assessment using CVSS," in *Proc. 8th ACM Conf. Data Appl. Secur. Privacy*, Mar. 2018, pp. 119–126.

[30] K. B. Alperin, A. B. Wollaber, and S. R. Gomez, "Improving interpretability for cyber vulnerability assessment using focus and context visualizations," in *Proc. IEEE Symp. Vis. Cyber Secur. (VizSec)*, Oct. 2020, pp. 30–39.

[31] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 4171–4186.

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[35] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021.

[36] T. Wolf, L. Debut, V. Sanh, and J. Chaumond, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[37] A. Paszke, S. Gross, F. Massa, and A. Lerer, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12 no. 10, pp. 2825–2830, 2012.

[39] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, 2009.

[40] L. S. Shapley, "A value for n-person games," *Contrib. Theory Games*, vol. 2, no. 28, pp. 307–317, 1953.
[41] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
[42] S. E. A. Lundberg. (2021). *Shap (Shapley Additive Explanations)*. Accessed: Jul. 2, 2021. [Online]. Available: https://github.com/slundberg/shap
[43] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. (NLP)*, 2018, pp. 353–355.
[44] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2016, pp. 2383—2392.



JOANA CABRAL COSTA received the bachelor's and master's degrees in computer science and engineering from the Universidade da Beira Interior (UBI), in 2019 and 2021, respectively, where she is currently pursuing the Ph.D. degree with a Fundação para a Ciência e a Tecnologia (FCT) Scholarship, in the field of computer vision and adversarial attacks.



TIAGO ROXO (Member, IEEE) received the bachelor's degree in computer science and engineering from the Universidade da Beira Interior (UBI), in 2019, where he is currently pursuing the Ph.D. degree with a Fundação para a Ciência e a Tecnologia (FCT) Scholarship, in the field of computer vision and artificial intelligence.



JOÃO B. F. SEQUEIROS received the bachelor's and master's degrees in computer science and engineering from the Universidade da Beira Interior, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree under the title "Towards a Framework for System and Attack Modeling and Mapping of Security Requirements for the Internet of Things." His dissertation focused on the development of a box for automated network-based security assessments. He has authored or coauthored several journals and conference papers mainly in the security field. His main research interests include network and application security, cryptography, cybersecurity, and the IoT security.



HUGO PROENÇA (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Beira Interior, and has been researching mainly about biometrics and visual-surveillance. He is a member of the Editorial Boards of the *Image and Vision Computing*, IEEE Access, and *International Journal of Biometrics*. He served as a Guest Editor for special issues of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image and Video Processing* journals. He was the Co-ordinating Editor of the IEEE Biometrics Council Newsletter and the Area Editor (Ocular Biometrics) of the IEEE Biometrics Compendium journal.



PEDRO R. M. INÁCIO (Senior Member, IEEE) was born in Covilhã, Portugal, in 1982. He received the B.Sc. degree in mathematics/computer science and the Ph.D. degree in computer science and engineering from the Universidade da Beira Interior (UBI), Portugal, in 2005 and 2009, respectively. The Ph.D. work was performed in the enterprise environment of Nokia Siemens Networks Portugal S.A., through a Ph.D. Grant from the Portuguese Foundation for Science and Technology.

He has been a Professor of computer science at the UBI, since 2010, where he lectures subjects related with information assurance and security, programming of mobile devices and computer based simulation, to graduate and undergraduate courses, namely to the B.Sc., M.Sc. and Ph.D. programs in computer science and engineering. He is currently the Head of the Department of Computer Science, UBI. He is an Instructor of the UBI Cisco Academy. He is a Researcher of the Instituto de Telecomunicações (IT). He has about 40 publications in the form of book chapters and papers in international peer-reviewed books, conferences, and journals. He frequently reviews articles for IEEE, Springer, Wiley, and Elsevier journals. His main research interests include information assurance and security, computer-based simulation, and network traffic monitoring, analysis, and classification.

Dr. Inácio has been a member of the Technical Program Committee of International Conferences, such as the ACM Symposium on Applied Computing—Track on Networking. He was one of the chairs of WISARC 2016.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2024, accepted 25 April 2024, date of publication 29 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395118

SURVEY

How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses

JOANA C. COSTA, TIAGO ROXO, HUGO PROENÇA, (Senior Member, IEEE), AND PEDRO RICARDO MORAIS INÁCIO, (Senior Member, IEEE)
Instituto de Telecomunicações, sins-lab, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
Department of Computer Science, Universidade da Beira Interior, 6201-001 Covilhã, Portugal

Corresponding author: Joana C. Costa (joana.cabral.costa@ubi.pt)

This work was supported in part by Portuguese Fundação para a Ciência e Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior (FCT/MCTES) through National Funds and co-funded by European Union (EU) Funds under Project UIDB/50008/2020; and in part by the FCT Doctoral Grant 2020.09847.BD and Grant 2021.04905.BD.

ABSTRACT Deep Learning is currently used to perform multiple tasks, such as object recognition, face recognition, and natural language processing. However, Deep Neural Networks (DNNs) are vulnerable to perturbations that alter the network prediction, named adversarial examples, which raise concerns regarding the usage of DNNs in critical areas, such as Self-driving Vehicles, Malware Detection, and Healthcare. This paper compiles the most recent adversarial attacks in Object Recognition, grouped by the attacker capacity and knowledge, and modern defenses clustered by protection strategies, providing background details to understand the topic of adversarial attacks and defenses. The new advances regarding Vision Transformers are also presented, which have not been previously done in the literature, showing the resemblance and dissimilarity between this architecture and Convolutional Neural Networks. Furthermore, the most used datasets and metrics in adversarial settings are summarized, along with datasets requiring further evaluation, which is another contribution. This survey compares the state-of-the-art results under different attacks for multiple architectures and compiles all the adversarial attacks and defenses with available code, comprising significant contributions to the literature. Finally, practical applications are discussed, and open issues are identified, being a reference for future works.

INDEX TERMS Adversarial attacks, adversarial defenses, datasets, evaluation metrics, review, vision transformers.

I. INTRODUCTION

Machine Learning (ML) algorithms have been able to solve various types of problems, with Deep Neural Networks (DNNs) [1] achieving results similar to, or better than, humans in multiple tasks, such as Object Recognition [2], [3], Face Recognition [4], [5], and Natural Language Processing [6], [7]. These networks have also been employed in critical areas, such as Autonomous Driving [8], [9], Malware Detection [10], [11], and Healthcare [12], [13], whose application and impaired functioning can severely impact their users.

DNNs results suggested that these networks could correctly generalize in the local neighborhood of an input (image), which motivated their adoption and integration in real-time image analysis, making malicious entities target these techniques. Recently, Szegedy et al. [14] found that small perturbations in images can entirely alter DNNs prediction, without changing Human prediction and significantly increasing the confidence of erroneous output. Another early study [15] distinguishes four types of adversaries depending on the amount of information they can access, which was relevant to cluster the different types of attacks, and suggested that the strength of the attack is diminished in real scenarios, requiring queries that are usually limited or very expensive [16]. The perturbations effect has been analyzed with more emphasis on Object Recognition, being one of the inception areas for Deep Learning.

In this survey, we present the most recent adversarial attacks and defenses in the Object Recognition area. The

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

defense mechanisms [17], [18], [19], [20], which we clustered on six different domains, can reliably defend against legacy attacks [14], [21]. However, there is no assurance about the effective robustness of DNNs and if they can be trusted in critical areas, clearly raising the need to make the DNNs inherently robust or easy to be updated every time a new vulnerability is encountered.

Previous surveys [22], [23], [24] do not provide a state-of-the-art comparison, area-specific metrics, and architectures commonly used, hindering the initiation among new researchers. This survey provides a comprehensive review of the legacy and recent advances, such as Adversarial Purification and Adversarial Attacks in Vision Transformers (ViTs), presents datasets and metrics in this context, provides results for multiple datasets, using the *de facto* standard attack (Auto-Attack [25]), and summarizes approaches that have official or unofficial implementations (duly identified) to stimulate the research of new mechanisms. This survey intends to answer and was guided by the following research questions:

- **RQ1**: Which are the most recent developments regarding Vision Transformers?
- **RQ2**: Which metrics are used to evaluate the effect of the attacks and robustness of models/defenses?
- **RQ3**: What are the state-of-the-art results and architectures in Adversarial Attacks and Defenses?
- **RQ4**: Which practical applications and possible future directions in the topic of Adversarial Attacks and Defenses in Object Detection can we follow?

The papers considered in this study were all included in proceedings of international CORE A* and A conferences or published in Q1 and Q2 journals and the initial collection was built by conducting searches with the keywords "*Adversarial*", "*Attack*", "*Defense*", "*Adversarial Attack*", and "*Robustness*", starting from 2013 (inception of this topic). The search engines were IEEE Xplore, Google Scholar, ACM Digital Library, and Elsevier. The inclusion criteria admitted papers that cited the key works [14], [15], full papers published in journals and conferences, written in English, and articles that address an adversarial attack/defense on state-of-the-art datasets. The exclusion criteria ruled out duplicate papers, whitepapers, articles shorter than two pages, studies without full text available, no proposal of new attacks or defenses, no evaluation on the mentioned datasets, and containing only non-reproducible experiments. Papers that were not related to the topic in question were excluded by reading the abstract, introduction, and conclusion. The main contributions of this work can be summarized as follows:

- We present the most recent adversarial attacks grouped by the adversary capacity, accompanied by an illustration of the differences between black-box and white-box attacks;
- We propose six different domains for adversarial defense grouping, assisted by exemplary figures of each of these domains, and describe the effects of adversarial examples in ViTs;
- We detail the most widely used metrics and datasets, present state-of-the-art results on CIFAR-10, CIFAR-100, and ImageNet, and propose directions for future works.

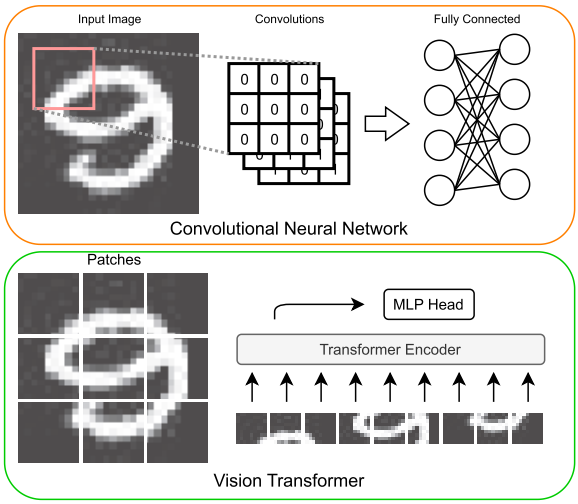


FIGURE 1. Schematic example of the convolutional neural networks (top) and vision transformers (bottom) to classify images.

The remaining of the paper is organized as follows: Section II provides background information; Section III compares this review with others; Section IV presents the set of adversarial attacks; Section V shows a collection of defenses to overcome these attacks; Section VI describes the advances in ViTs; Section VII displays the commonly used datasets; Section VIII lists and elaborates on metrics and presents state-of-the-art results; and Section IX presents future directions, with the concluding remarks included in Section X.

II. BACKGROUND FOR ADVERSARIAL ATTACKS

A. NEURAL NETWORK ARCHITECTURES

When an input image is fed into a CNN [26], it is converted into a matrix, or a set of matrices, containing the numeric values for each color channel. Then, the Convolutional layers apply filters to these matrixes and calculates a set of reduced-size features, which are fed into the Fully Connected layers that classify the provided image. ViT [27] gives a pre-processed image that is divided into Patches, which are crops of the original image, and their ordered position to the Transformer Encoder. The output resulting from this encoder is fed into the Multi-Layer Perceptron (MLP) Head that classifies the image. Figure 1 shows a schematic overview of CNNs and ViTs. For a complete explanation of the functioning of CNNs and ViTs we refer to [1] and [27], respectively.

B. ADVERSARIAL EXAMPLE

Misclassification can be justified if the object contained in the image is not visible even to Humans, yet adversarial examples

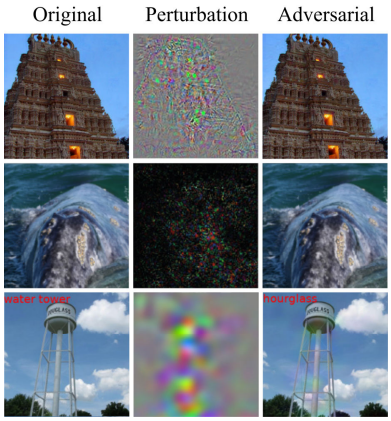


FIGURE 2. Adversarial examples created using different adversarial attacks. The images were withdrawn from [14], [28], and [29] and resized for better visualization.

do not fit this scope. These examples add a perturbation to an image that causes the DNNs to misclassify the object in the image, yet Humans can correctly classify the same object. The adversarial attacks focus on identifying the perturbations that make DNNs misclassify, with their calculation being an optimization problem formally defined as:

$$\arg\min_{\delta \mathbf{X}} \|\delta_{\mathbf{X}}\| \quad \mathbf{s.t.} \quad f(\mathbf{X}+\delta_{\mathbf{X}}) = \mathbf{Y}*, \tag{1}$$

where f is the classifier, $\delta_{\mathbf{X}}$ is the perturbation, $\mathbf{X}$ is the original/benign image, and $\mathbf{Y}*$ is the adversarial output. Furthermore, the adversarial example is defined as:

$$\mathbf{X}* = \mathbf{X} + \delta_{\mathbf{X}}, \tag{2}$$

where $\mathbf{X}*$ is the adversarial image. Figure 2 displays adversarial examples generated using different attacks. Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [14], DeepFool [28], and SmoothFool [29] correspond to the first, second, and third rows. The L-BFGS applies noise to almost the entirety of the adversarial image, the DeepFool only perturbs the area of the depicted whale but not all the pixels in that area, and the SmoothFool slightly disturbs the pixels in the area of the image, displaying the evolution of adversarial attacks in decreasing order of detectability. To limit the amount of noise, the adversarial attacks are divided into L_0, L_2, and L_p norms, known as *Vector Norms*, and common terminologies in the context of adversarial examples are defined in Table 1.

C. VECTOR NORMS AND ϵ CONSTRAINT

Vector Norms are functions that take a vector as input and output a positive value (scalar), which are essential to ML and allow the backpropagation algorithms to compute the loss value as a scalar. These functions are known as the *p-norm*, and the considered values are 0, 2, and ∞. L_0 norm consists of counting the number of non-zero elements in the vector, L_2 norm, also known as the Euclidean distance, measures the vector distance to the origin, and L_∞ norm represents the maximum hypothetical value that p can have and returns the absolute value of the element with the largest magnitude.

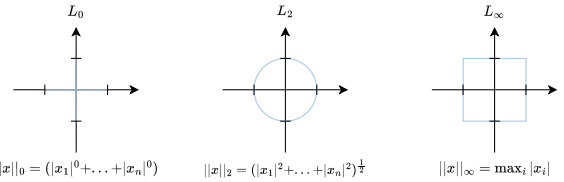


FIGURE 3. Geometric representation and formal definition of the L_0, L_2, and L_∞ norms, from left to right, respectively, with x_i referring to each element of the vector x.

TABLE 1. Common terminologies used in the context of adversarial attacks and their definition.

Terminology	Definition
Clean Example	Image without noise from a dataset
Adversarial Example	Image with a perturbation that fools a classifier
Perturbation	Set of changes/transformations applied to a image
Adversarial Attack	Technique used to calculate the perturbation
Transferability	Capability of an adversarial example being transferred from a known network to an unknown network

A geometric representation of the area of exploitation and a formal definition for the three considered p-norm is displayed in Figure 3.

One relevant property of the p-norm is: the higher p is, the more important the contribution of large errors; the lower p is, the higher the contribution of small errors. This translates into a large p benefiting small maximal errors (minimal perturbations along multiple pixels) and a small p encouraging larger spikes in fewer places (abrupt perturbations along minimal pixels). Therefore, L_2 and L_0 attacks have greater detectability than L_∞ attacks, with the latter being more threatening.

Another constraint normally seen is ϵ, which is a constant that limits the amount of noise that can be added to an image, and it is usually a tiny number that varies depending on the dataset, decreasing when the task increases in difficulty. According to the literature, for MNIST, $\epsilon = 0.1$, for CIFAR-10 and CIFAR-100, $\epsilon = 8/255$, and for ImageNet, $\epsilon = 4/255$.

D. ADVERSARY GOALS AND CAPACITY

Besides the restriction imposed by the different Vector Norms, the adversarial attacks are also divided by their impact on the networks, depending on the goals of the attacker, with the following designation: 1) **Confidence Reduction**, the classifier outputs the original label with less confidence; 2) **Untargeted**, the classifier outputs any class besides the original label; and 3) **Targeted**, the classifier outputs a particular class besides the original label.

As defined by Papernot et al. [15], the attackers can have access to: 1) training data and network architecture; 2) only network architecture; 3) only training data; 4) an oracle that replies to all the inputs given; and 5) only have pairs of input and corresponding output (samples). However, to simplify this classification, these capacities were divided into: 1) **White-box**, which considers that the attacker has access to both architecture and data; and 2) **Black-box**, when the attacker can only access samples from an oracle or pairs of input and output.

TABLE 2. Characteristics shown in state-of-the-art surveys on adversarial attacks.

Survey	Year	White & Black-Box	Survey Comparison	Grouping of Defenses	Future Directions	Datasets Overview	Metrics and Architectures	State-of-the-art Comparison	Vision Transformers
Akhtar and Mian [30]	2018	✓	×	✓	✓	×	×	×	×
Qiu *et al.* [31]	2019	✓	×	×	×	×	×	×	×
Serban *et al.* [32]	2020	✓	×	✓	✓	×	×	×	×
Xu *et al.* [33]	2020	✓	×	✓	×	×	×	×	×
Sadeghi *et al.* [34]	2020	✓	✓	✓	✓	✓	×	×	×
Chakraborty *et al.* [35]	2021	✓	×	×	×	×	×	×	×
Long *et al.* [22]	2022	✓	✓	×	✓	×	×	×	×
Liang *et al.* [23]	2022	✓	×	✓	✓	×	×	×	×
Zhou *et al.* [24]	2022	✓	×	✓	✓	✓	×	×	×
Mi *et al.* [36]	2023	✓	×	✓	×	×	×	×	×
Amritasree *et al.* [37]	2023	✓	×	✓	✓	×	×	×	×
This survey	2024	✓	✓	✓	✓	✓	✓	✓	✓

III. RELATED SURVEYS

Akhtar and Mian [30], initially surveyed adversarial attacks in computer vision, extensively referring to attacks for classification, providing a brief overview of attacks beyond the classification problem, presenting a set of attacks performed in the real world, providing insight into the existence of adversarial examples, and clustering defenses in three categories. From a broader perspective, Liu et al. [38] studies the security threats and possible defenses in ML scope, considering the different phases of an ML algorithm, and provides insight on the currently used techniques. Chakraborty et al. [35] provides insight into commonly used ML algorithms, presents the adversary capabilities and goals, cluster adversarial attacks based on the stage of the attack, and introduces relevant defenses used in adversarial settings.

Focusing more on the object recognition task, Serban et al. [32] extensively analyzed the adversarial attacks and defenses, providing conjectures for the existence of adversarial examples and evaluating the capacity of adversarial examples transferring between different DNNs. Qui et al. [31] explains background concepts in adversarial attacks, displays practical applications, presents the most relevant adversarial defenses, and illustrates a set of attacks divided by the stage in which they occur, referring to the most relevant attacks. Sadeghi et al. [34] summarizes the basic concepts, analyzes datasets and ML architectures, providing awareness of the adversary strategies and the defense responses to them, and proposes a comprehensive system-drive taxonomy for the latter. Long et al. [22] discusses a set of preliminary concepts of Computer Vision and adversarial context, providing a set of adversarial attacks grouped by adversary goals and capabilities, and propose a set of research directions that readers can use to continue the development of robust networks. Liang et al. [23] discuss the most significant attacks and defenses in the literature, with the latter being grouped by the underlying technique, finishing with the presentation of existing challenges in the adversarial context.

Xu et al. [33] also describes background concepts, summarizes the most relevant adversarial attacks at the time of that work, presenting physical world examples, classify defenses by the underlying methodology, and outline adversarial attacks in graphs, text, and audio networks, culminating in the possible applications of these attacks. More recently, Zhou et al. [24] provides insight into Deep Learning and Threat Models, focusing on the Cybersecurity perspective. Therefore, the authors identify multiple stages based on Advanced Persistent Threats, explain which adversarial attacks and defenses are employed in each stage, present the commonly used datasets in adversarial settings, and provide a set of future directions from a Cybersecurity perspective.

Mi et al. [36] and Amritasree et al. [37] evaluate the recent adversarial attacks and defenses, providing an extensive description of the different types of adversarial attacks, that specifically focus on Object Detection. The latter [37] includes a set of future research directions.

The literature clarifies that some of the specialized concepts of the area, such as adversary goals and capabilities and adversarial attacks and defenses, have already solidified and become standard. However, the datasets, architectures, and metrics usage still lack standardization. In an attempt to do so, this survey analyzes datasets, architectures, and metrics to provide insight to novice researchers. Additionally, it presents state-of-the-art results and provides recent advancements in adversarial attacks on ViTs, two contributions of this survey. Concluding with a set of future directions that researchers and practitioners can follow to start their work, which is the main contribution. A comparison between the several surveys discussed in this section (with the work reflected herein) is summarized in Table 2.

IV. ADVERSARIAL ATTACKS

Adversarial attacks are commonly divided by the amount of knowledge the adversaries have access to, white-box and black-box, as can be seen in Figure 4. Furthermore, the development of defenses is highly motivated by the proposal of new attacks, with the same being true for the attacks, suggesting that some defenses might have overcome some attacks and vice-versa. For example, Carlini and Wagner attack overcomes the Defensive Distillation methodology.

A. WHITE-BOX SETTINGS

Adversarial examples were first proposed by Szegedy et al. [14], which discovered that DNNs do not generalize well in the vicinity of an input, leading to the proposal of **L-BFGS**, which raised awareness in the scientific community for this generalization problem. L-BFGS is a computationally

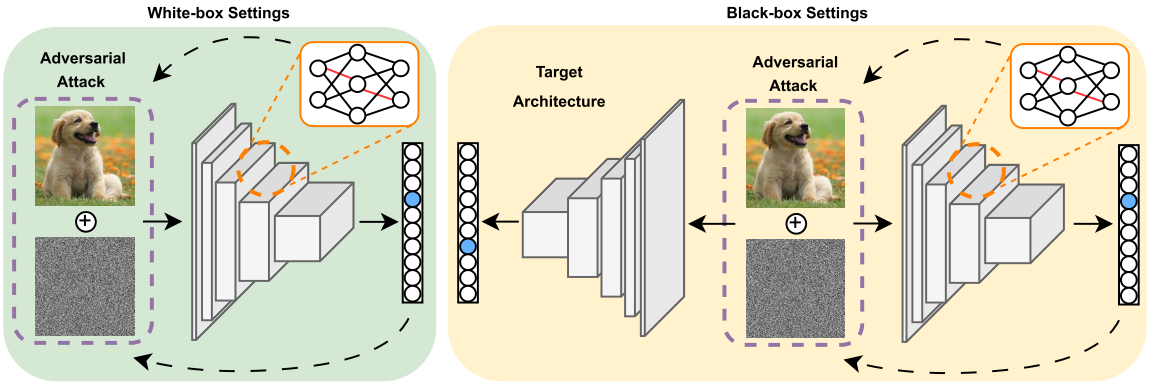


FIGURE 4. Schematic overview of an adversarial attack under white-box settings (left) and black-box settings (right). The first one uses the classifier predictions and network gradients to create perturbations (similar to noise), which can fool this classifier. These perturbations are added to the original images, creating adversarial images, which are fed to the network and cause misclassification. In the black-box settings, the same process is applied to a known classifier, and the obtained images are used to attack another classifier (represented as target architecture).

expensive method, which motivated **Fast Gradient Sign Method** (FGSM) [21], a one-step method to find adversarial examples, calculated using the model cost function, the gradient, and the radius epsilon. This attack is formally defined as:

$$\mathbf{x} - \epsilon \cdot \text{sign}(\nabla \text{loss}_{F,t}(\mathbf{x})), \tag{3}$$

where $\mathbf{x}$ is the original image, ϵ is the amount of changes to the image, and t is the target label. The value for ϵ should be very small to make the attack undetectable. **Jacobian-based Saliency Maps** (JSM) [15] explore the forward derivates to calculate the model gradients, replacing the gradient descent approaches, and discover which input regions are likely to yield adversarial examples. Then, it constructs the adversarial saliency maps, displaying the features the adversary must perturb, with only the adversarial examples correctly classified by humans being used to fool neural networks. **Gradient Aligned Adversarial Subspace** (GAAS) [39] estimates the dimensionality of the adversarial subspace using the first-order approximation of the loss function, proving to be the most successful at finding many orthogonal attack directions, which indicates that neural networks generalize linearly. Croce and Hein [40] noticed that perturbations generated by l_0 attacks are sparse and by l_∞ attacks are smooth on all pixels, proposing **Sparse and Imperceivable Adversarial Attacks** (SIAA), which creates sporadic and imperceptible perturbations by applying the standard deviation of each color channel in both axis directions, calculated using the two immediate neighboring pixels and the original pixel.

DeepFool [28] is an iterative attack that stops when the minimal perturbation, corresponding to the vector orthogonal to the hyperplane representing the decision boundary, is found. **SmoothFool** (SF) [29] is an iterative algorithm that uses DeepFool to calculate the initial perturbation and smoothly rectifies the resulting perturbation until the adversarial example fools the classifier, providing smoother perturbations that improve the transferability of adversarial examples, whose impact varies with the different categories in a dataset. **Projected Gradient Descent** (PGD) [41] is an iterative attack that uses saddle point formulation, an inner maximization problem to find an adversarial example that achieves high loss and an outer minimization problem to find model parameters that minimize the first, to find a strong perturbation. The saddle point problem used by PGD is defined as:

$$\min_{\theta} \rho(\theta), \text{ where } \rho(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\max_{\delta\in\mathcal{S}} L(\theta, \mathbf{x}+\delta, y)\Big], \tag{4}$$

where $\mathbf{x}$ is the original image, y is the corresponding label, and $\mathcal{S}$ is the set of allowed perturbations δ. Motivated by the inability to achieve a high transferability, the **Momentum Iterative FGSM** (MI-FGSM) [42] introduces momentum, a technique for accelerating gradient descent algorithms, into the Iterative FGSM (I-FGSM), showing that the attack transferability almost doubles.

Carlini and Wagner (C&W) [43] attack is one of the most powerful attacks, which uses three different vector norms: 1) L_2, a smoothing of clipped gradient descent approach, displaying low distortion; 2) L_0, an iterative algorithm that, at each iteration, fixes the pixels that do not have much effect on the classifier and finds the minimum amount of pixels that need to be altered; and 3) L_∞, an iterative algorithm with an associated penalty, penalizing every perturbation that exceeds a predefined value, formally defined as:

$$\min \quad c \cdot f(\mathbf{x}+\delta) + \sum_i [(\delta_i - \tau)^+], \tag{5}$$

where δ is the perturbation, τ is the penalty threshold (initially 1, decreasing in each iteration), and c is a constant. The value for c starts as a very low value (*e.g.*, 10^{-4}), and each time the attack fails, the value for c is doubled. If c exceeds a threshold (*e.g.*, 10^{10}), it aborts the search. **Meta Gradient Adversarial Attack** (MGAA) [44] can be integrated into any existing gradient-based attack method to improve cross-model transferability, consisting of multiple iterations, and various models that are used to generate adversarial perturbations, which are added to the previously

generated perturbations. Using multiple models simulates both white- and black-box settings, making the attacks more successful.

AdvGAN [45] uses Generative Adversarial Networks (GAN) [46] to create realistic adversarial examples that have high attack success rate. The generator receives the original instance and creates a perturbation, the discriminator distinguishes the original instance from the perturbed instance, and the target neural network is used to measure the distance between the prediction and the target class. Recently there was a shift into natural Unrestricted Adversarial Examples (UAEs), breaking the l_p perturbation bounds while remaining semantically plausible, considered by **AdvDiffuser** [47], which generates UAEs using diffusion models that are more natural and stronger than the state-of-the-art. In an attempt to exploit the vulnerabilities of detector networks, **Adversarial Manipulation Generation** (AMG) [48] is an attack that uses spatial and frequency features to generate adversarial examples that fool the detector networks, having high success rate.

Another topic that is being discussed under the context of white-box is the existence of backdoor attacks, by including poisoned data in the training process. Chen et al. [49] proposes a stealthier backdoor attack using input-instance-key and pattern-key strategies. Liu et al. [50] performs a backdoor attack not by manipulating the initial training but by doing an additional train that takes only a few hours. **BadNets** [51] considers the creation of DNNs from outsourcing and third-party cloud providers and Weng et al. [52] evaluates the trade-off between adversarial resilience and backdoor robustness, by conducting extensive experiments.

Summary: The white-box attacks assume that the attacker has access to information regarding the model, such as weights and data. Therefore, they are the most powerful type of attack and can, the majority of times, completely fool the models. However, these attacks are model-specific and do not transfer to different models, having to be executed again for each model.

B. UNIVERSAL ADVERSARIAL PERTURBATIONS

Moosavi-Dezfooli et al. [53] discovered image-agnostic perturbations, labeled as **Universal Adversarial Perturbations** (UAPs), that generalize well and cause misclassification with high probability, by searching for a vector of perturbations that causes misclassification in almost all the data drawn from a distribution of images. The optimization problem that Moosavi-Dezfooli et al. are trying to solve is the following:

$$\Delta v_i \leftarrow \arg\min_{\mathbf{r}} \|\mathbf{r}\|_2 \quad \text{s.t.} \quad \hat{k}(\mathbf{x}_i + v + \mathbf{r}) \neq \hat{k}(x_i), \qquad (6)$$

where Δv_i is the minimal perturbation to fool the classifier, $\hat{k}$ is the classifier, $\mathbf{r}$ is a perturbation vector that sends $\mathbf{x}_i + v$ to the decision boundary, v is the universal perturbation, and x_i is the original image. This optimization problem is calculated for each image in a dataset, and the vector containing the universal perturbation is updated.

The **Universal Adversarial Networks** (UAN) [54] are GANs that use UAPs as the training set and are capable of outputting perturbations for any given input that fool classifiers, demonstrating more outstanding results than the original UAPs. Based on UAP generation problems, gradient vanishing and local optima, **Stochastic Gradient Aggregation** (SGA) [55] uses multiple iterations of inner pre-search and aggregates all the inner gradients as a one-step gradient estimation. UAPs are highly dependent of prior data, leading to the proposal of data-free UAP methods that rely on random noises and maximize the CNN activation, **TRM-UAP** [56] is a data-free UAP generator that does not depend on any real data through truncated ratio maximization. Exploring the vulnerabilities of different downstream tasks, **AdvEncoder** [57] constructs a UAP that is able to fool all the downstream tasks, providing a new insight for how adversarial attacks can be applied.

Kurakin et al. [58] initially demonstrates that adversarial examples exist in the physical world, by generating adversarial examples and printing them on paper. **Basic Iterative Method** (BIM) applies the FGSM multiple times with a small step size between iterations and clips the intermediate values after each step, and **Iterative Least-likely Class Method** (ILCM) uses the least-likely class, according to the prediction of the model, as the target class and uses BIM to calculate the adversarial example that outputs the target class. **Adversarial Camouflage** (AdvCam) [59] crafts physical-world adversarial examples that are legitimate to human observers, using the target image, region, and style to perform a physical adaptation (creating a realistic adversarial example). Physical adversarial attacks also use printed adversarial patches/camouflages to alter the appearance of the target object, while remaining semantically plausible, **Reflected Light Attack** (RFLA) [60] generates adversarial examples by placing a color transparent plastic sheet and a paper cut aligned in front of a mirror to generate multiple colored geometries on the target object. Wei et al. [61] proposes a **Unified Adversarial Patch** that performs cross-modal, both visible and infrared, physical attacks with a single patch.

Summary: The goal of Universal Adversarial Perturbations is to create unique perturbations that have effect on multiple models and multiple images, and they are a subgroup of the black-box attacks. The same perturbation can be added to different images to perform an attack for different models, yet these perturbations are weaker than white-box but more transferable.

C. BLACK-BOX SETTINGS

Specifically considering the black-box setup, Limited Queries and Information Attack (LQIA) [62] proposes three different realistic approaches: 1) query-efficient algorithms, using Natural Evolutionary Strategies to estimate the gradients used to perform the PGD attack; 2) alternate between blending in the original image and maximizing the likelihood of the target class; and 3) include noise

robustness to mount a targeted attack. Under the same problem, **Subspace Activation Evolution Strategy** (SA-ES) [63] proposes an optimization method to capture the most promising direction, in a query-limited setting, generating a distribution of adversarial examples.

Feature-Guided Black-Box (FGBB) [64] views the creation of adversarial examples, using Scale Invariant Feature Transform, as a two-player game, where the first minimizes the distance and the second has multiple roles, leading to minimal adversarial examples. Another approach based on creating minimal adversarial examples using the decision of a model is the **Boundary Attack** [65], which repeatedly decreases the distance between the adversarial image and the target image. Similarly, **HopSkipJump** [66] is a decision-based attack that uses fewer queries, based on an optimized gradient estimation. The challenge of Black-box settings has led to multiple prospects, such as synthetic data [67], fluctuation phenomenon [68], and input transformation [69].

Square Attack [70] is an attack that uses a randomized search scheme that selects localized square-shaped updates in random positions, causing the perturbation to be situated at the decision boundaries. Further exploring the advantages of features, **Transferable Targeted Adversarial Attack** (TTAA) [71] captures the distribution information of the target class from label- and feature-wise perspective to generate highly transferable adversarial examples and **Style-Less perturbation** (StyLess) [72] uses stylized networks, encoding different style features, by perturbing an adaptive instance normalization. **Feature Importance-aware Attack** (FIA) [73] considers the object-aware features that dominate the model decisions and identifies the essential features, generating highly transferable adversarial examples when extracting the feature importance from multiple classification models.

Zanddizari et al. [74] sets a threshold to choose large sparse (limiting the amount of pixels that can be altered) components and perturb them either randomly or in the direction of another class. However, the limitation of sparse attacks causes the generation of evident perturbations, reducing the stealthiness of these attacks. To minimize this effect, **Multi-objective Sparse Attacks** (MSA) [75] efficiently minimizes the number of modified pixels and the size of perturbations, increasing the stealthiness and having state-of-the-art results.

Inspired on the difficulties of transferring adversarial examples from CNNs to ViTs and the success of ensemble attacks, **AdaEA** [76] adaptively controls the fusion of the outputs from each model toward the adversarial objective and **LEA**^{2} [77] considers three models with non-overlapping vulnerable frequency regions to cover a large vulnerable subspace. **Momentum Integrated Gradients** (MIG) [78] uses the Integrated Gradients and Momentum updating strategy to generate adversarial examples with high success against ViTs and transferability across the ViTs and CNNs.

Summary: There has been a recent focus on developing stronger Black-box attacks, since these are the more realistic ones, and assume that the attacker only has access to queries or pairs of input and respective prediction. Despite this, these attacks have lower impact on the performance of the model and can more easily be outperformed by defenses.

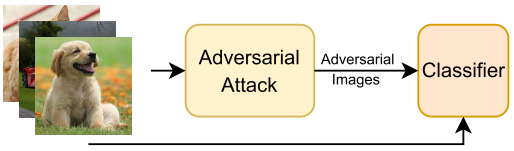


FIGURE 5. Schematic overview of adversarial training. A subset of the original images of a dataset is fed into an adversarial attack (*e.g.*, PGD, FGSM, or C&W), which creates adversarial images. Each batch contains original and adversarial images, with the Classifier being normally trained.

D. AUTO-ATTACK

Auto-Attack [25] is capable of testing adversarial robustness in a parameter-free, computationally affordable, and user-independent way, by proposing two variations of PGD that overcome sub optimal step sizes of the objective function, APGD-CE and APGD-DLR, with PGD using Cross-Entropy (CE) and Difference of Logits Ratio (DLR) loss, respectively. These new PGD variations are combined with Fast Adaptive Boundary (FAB) [79], which minimizes the norm of the adversarial perturbations, and Square Attack [70], a query-efficient black-box attack, to create Auto-Attack. Given the need for a computationally affordable approach, the FAB attack in Auto-Attack is a targeted version of FAB, which is both computationally and memory-wise plausible in datasets with an increased number of classes (*e.g.*, CIFAR-100 and ImageNet). Auto-Attack is an ensemble of attacks with important fundamental properties: APGD is a white-box attack aiming at any adversarial example within an L_p-ball, FAB minimizes the norm of the perturbation necessary to achieve a misclassification, and Square Attack is a score-based black-box attack for norm bounded perturbations which use random search and do not exploit any gradient approximation, competitive with white-box attacks [70].

V. ADVERSARIAL DEFENSES

The countermeasures are divided by their practical point of view, namely how quickly can they be adopted to the current models, and are an adaptation of previously presented taxonomies [32], [38].

A. ADVERSARIAL TRAINING

Szegedy et al. [14] proposed that training on a mixture of adversarial and clean examples could regularize a neural network, as shown in Figure 5. Goodfellow et al. [21] evaluated the impact of **Adversarial Training** as a regularizer by including it in the objective function, showing that this approach is a reliable defense that can be applied to every neural network. Some preliminary works evaluated how different image preprocessing techniques could influence the robustness, Gu and Rigazio [80] used noise injection, autoencoder, and denoising autoencoder and Zheng et al. [81]

used compression, rescaling, and cropping, denominated as **Image Processing**, discovering that adversarial noise mainly exists in high-frequency domain.

Kurakin et al. [82] performs adversarial training in more massive datasets (ImageNet), for the first time, displaying that the robustness significantly increases when using one-step methods. Training a model using one-step attacks and ground-truth labels leads to better accuracy on the adversarial images than on the clean images, an effect denominated as **Label Leaking**, suggesting that the adversarial training should not make use of the ground-truth labels. **Ensemble Adversarial Training** (EAT) [83] uses adversarial examples crafted for other pretrained models to augment training data to increase resilience to black-box attacks. To avoid the Label Leaking effect, Zhang and Wang [19] suggest that adversarial examples are generated through **Feature Scattering** (FScatter) in the latent space, by maximizing the feature-matching distance between the clean and perturbed examples, and the DNN performs standard adversarial training. To tackle the same issue, **Unsupervised Adversarial Training** (UAT) [84] suggests the training with unlabeled data considering two different approaches: (i) UAT with Online Target (UAT-OT) that minimizes a differentiable surrogate of the smoothness loss and (ii) UAT with Fixed Targets (UAT-FT) that trains an external classifier to predict the labels on the unsupervised data and uses its predictions as labels. Considering that disturbing labels can also improve accuracy, **Helper-based Adversarial Training** (HAT) [85] incorporates additional wrongly labeled examples during training, achieving a good trade-off between accuracy and robustness.

Shared Adversarial Training [86] uses gradients to create a shared perturbation for a mini-batch of images, which are accumulated and clipped to constrain them into a predefined magnitude, creating the adversarial examples used for adversarial training. Other approach that uses custom-made adversarial examples is **Defense against Occlusion Attacks** (DOA) [87], which is a defense mechanism that includes physically realizable attacks, ''normal'' in the real world (*e.g.*, eyeglasses and stickers), and applies standard adversarial training.

TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES) [88] is inspired by the presumption that robustness can be at odds with accuracy [89], [90], encouraging the natural error to be optimized and minimizing the distance between prediction of natural and adversarial examples. However, this approach still lacks differentiating misclassified from correctly classified examples, thus **Misclassification Aware adveRsarial Training** (MART) [91] explicitly differentiates these examples during training by introducing a misclassification aware regularization term.

Despite the popularity of adversarial training to defend models, it has a high cost of generating strong adversarial examples, **Free Adversarial Training** (Free-AT) [92] uses the gradient information when updating model parameters to generate the adversarial examples, eliminating the previously mentioned overhead. Another optimization regarding the computational costs, consists in only considering the first layer of a network for forward and backpropagation, effectively reducing the amount of propagation to one in each update, named as **You Only Propagate Once** (YOPO) [93], due to the adversary update being only related to the first layer. Exploring the different layers of a network, **Latent Adversarial Training** (LAT) [94] consists of finetuning adversarial-trained models to ensure robustness at the latent level, because the latent layer is significantly vulnerable to adversarial perturbations of small magnitude. Instead of using computational costly methods, **Cheap Adversarial Training** (Cheap-AT) [95] proposes the use of weaker and cheaper adversaries (FGSM) combined with random initialization to train robust networks effectively. Inspired by the results from Cheap-AT, **Fast Adversarial Training** (FAT) [96] performs randomized smoothing to optimize the inner maximization problem efficiently, and proposes a new initialization strategy, named backward smoothing, that helps to improve the stability and robustness of a model using single-step robust training methods.

The proposal of **Smooth Adversarial Training** (SAT) [97] considers the evolution normally seen in curriculum learning, where the difficulty increases with time (age), using two difficulty metrics based on the maximal Hessian eigenvalue (H-SAT) and the softmax Probability (P-SAT), which are used to stabilize the networks for large perturbations while having high clean accuracy. In the same context, **Friendly Adversarial Training** (Friend-AT) [98] minimizes the loss considering the least adversarial data (friendly) among the adversarial data that is confidently misclassified. **Oracle-Aligned Adversarial Training** (OA-AT) [99] creates a model that is robust to high perturbation bounds by aligning the network predictions with ones of an Oracle during adversarial training, using Learned Perceptual Image Patch Similarity [100] to generate Oracle-Invariant attacks and convex combination of clean and adversarial predictions as targets for Oracle-Sensitive samples.

Bilateral Adversarial Training (BAT) [101] proposes to perturb the images and labels (adversarial labels) during the training phase, which are generated from a one-step targeted attack and derived from a closed-form heuristic solution, respectively. Also considering perturbing components besides the images, **Adversarial Weight Perturbation** (AWP) [102] regularizes the flatness of weight loss landscape and robustness gap, using a double-perturbation mechanism that disturbs both inputs and weights. This defense boosts the robustness of multiple existing adversarial training methods, confirming that it can be applied to other methods. Since disturbing weights shows promising results, **Geometry-aware Instance-reweighted Adversarial Training** (GI-AT) [103] proposes using standard adversarial training considering that the weights are based on how difficult it is to attack a natural data point, *i.e.*, natural data points closer to class boundary are less robust.

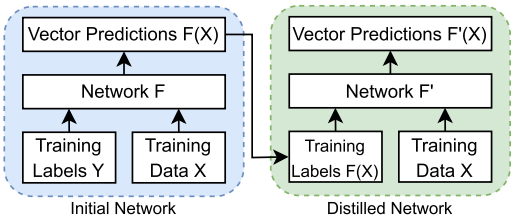


FIGURE 6. Method proposed by Defensive Distillation [104]. An Initial Network is trained on the dataset images and labels (discrete values). Then, the predictions given by the Initial Network are fed into another network, replacing the dataset labels. These predictions are continuous values, making the Distilled Network more resilient to adversarial attacks.

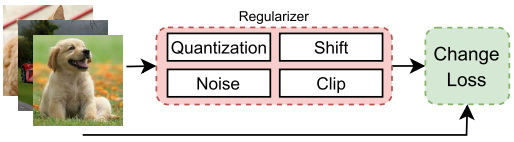


FIGURE 7. Schematic overview of the bit plane feature consistency [115]. This method applies multiple operations to input images, simulating adversarial images. Then, the loss is changed to include a regularizer (new term), which compares the original images with these manipulated images.

Summary: Adversarial Training is the most adopted approach to defend against adversarial attacks, demonstrating effective performance against specific attacks because they increase the boundaries for each class to include perturbed images. This type of defense consists in including adversarial examples in the training batches, using the original, other models, or heuristic labels, to train the robust model to recognize those images. Some works also suggest disturbing/changing the weights to be aware of the existence of adversarial examples. However, Adversarial Training using stronger attacks, such as PGD or C&W, is computationally expensive and lead to the search for more efficient approaches.

B. MODIFY THE TRAINING PROCESS

Defensive Distillation (DD) [104] uses knowledge distillation to train a resilient network by replacing the original labels with the labels of a previously trained network, as displayed in Figure 6, which is trained with original data and labels. To improve the results, Papernot and McDaniel [105] change the label vector used to train the distilled network by combining the original label with the first model uncertainty. Following the same context, **Robust Soft Label Adversarial Distillation** (RSLAD) [106] uses Robust Soft Labels (RSLs) produced by a teacher network to supervise the student training on natural and adversarial examples, without the latter having access to the original complex labels. **Low Temperature Distillation** (LTD) [107] considers low temperature in the teacher model using distillation to generate labels, by employing different fix temperatures for teacher and student models, which generates soft labels that can be integrated into existing works.

Robust Self-Training (RST) [108] uses a standard supervised approach to obtain pseudo-labels and feed them into a another network that targets adversarial robustness, bridging the gap between standard and robust accuracy using the unlabeled data. Exploring the results shown by not solely relying on original labels, **Self-Adaptive Training** (SAT) [109] uses as training labels a blend between the original labels and the model predictions, ignoring the first epochs due to instability, and updates their training labels throughout the training phase. In terms of changing the original data, **Visual Causal Feature Learning** [110] trains one network to predict causal effects and applies a manipulator function that return an image similar to the original one with the desired causal effect, performing data augmentation, which are fed to the classifier. **SENSEI** [111] and **SENSEI-SA** [111] use software testing methodologies to perform data augmentation, naively replacing each data point with a suitable variant or leaving it unchanged and selecting the opportunities suitable for skipping this augmentation process, respectively. Cutout [112] and Mixup [113] data augmentation techniques fail to prevent overfitting and make the model over-regularized, thus **Overfit Reduction** (OR) [114] uses a combination of early stopping and semi-supervised data augmentation to achieve substantial improvements.

Another considered approach is to add regularization terms, **Triplet Loss Adversarial** (TLA) [116] is a standard loss with an additional term that clusters natural and adversarial images of the same class and pushes the remaining classes further apart, using the Random Negative and the Switch Anchor techniques. **Curvature Regularization** (CURE) [117] introduces a regularizer term that penalizes large eigenvalues and minimizes the curvature of the loss surface, which is essential to achieve robustness. **Bit Plane Feature Consistency** (BPFC) [115] is a regularizer that uses the original and preprocessed images as shown in Figure 7, making the network give more importance to the higher bit planes. **Manifold Regularization for Locally Stable** (MRLS) [118] networks approximates the Laplacian graph when the data is sparse and exploits the continuous piece-wise linear nature of ReLU to create smooth decision boundaries.

Reducing the sensitive of neurons and their robust error are also valid approaches, **Controlling Neural Level Sets** (CNLS) [119] uses the level set values (*e.g.*, decision boundaries) from a sample network to relate the sample position with the network parameters, modifying the decision boundaries of the network. **Sensitive Neuron Stabilizing** (SNS) [120] includes a regularizer dedicated to maximizing the similarities of sensitive neuron behaviors when evaluation clean and adversarial examples, stabilizing the sensitive neurons that significantly contribute to predictions under adversarial settings. **Self-COnsistent Robust Error** (SCORE) [121] employs local equivariance to minimize the robustness error of a network, easing the reconciliation between robustness and accuracy and dealing with worst-case uncertainty. **Parametric Shifted Sigmoidal Linear Unit** (PSSiLU) [122] changes the activation function to create

high finite curvature and output positive results with negative inputs, creating a parameterized activation function, that is used with adversarial training.

Learning with a Strong Adversary (LSA) [123] uses adversarial learning, with the adversary applying perturbations to maximize the classification error and the classifier minimizing the misclassification error against the adversary, to significantly increase robustness while maintaining clean high accuracy. **Learnable Boundary Guided Adversarial Training** (LBGAT) [124] assumes that a model trained on clean data embeds the most discriminative features and constrains the logits from the robust model to make them similar to the first one. Recently, literature [125], [126], [127] demonstrated that neural Ordinary Differential Equations (ODE) are naturally more robust to adversarial attacks than vanilla networks. **Stable neural ODE for deFending against adversarial attacks** (SODEF) [128] uses optimization formulation to force the extracted feature points to be within the vicinity of Lyapunov-stable equilibrium points, suppressing the input perturbations.

Summary: This Modify the Training Process defense represents a wide variety of approaches that demonstrate increased robustness when applied to standard networks, which is divided into performing model distillation, augmenting data, changing the original labels, including a regularizer, constraining the weights, stabilizing neuron sensitive. The underlying idea is to modify the training process of the model, which requires that the training process is repeated to create a more robust model, translating into an additional effort to include this defenses in currently operating models.

C. USE OF SUPPLEMENTARY NETWORKS

The use of detector networks to avoid the transmission of adversarial examples to the classifier was one of the first defense mechanisms. **MagNet** [129] uses multiple detectors, trained on the reconstruction error, that are capable of detecting significantly and slightly perturbed examples based on probability divergence. In the same context, **Adversary Detection Network** (ADN) [130] is a network trained using adversarial images for a classifier that detects if the input example is adversarial or clean. Figure 8 displays a schematic overview of detector networks. Instead of training a network on adversarial examples, **Feature Squeezing** (FSqueez) [131] uses multiple squeezing techniques, which reduce the search space, before giving the image to the classifier and compares the predictions of standard and squeezed classifiers to detect an adversarial example. **Detector Graph** (DG) [132] constructs a Latent Neighborhood Graph for each original image and uses Graph Neural Networks (GNNs) [133] to distinguish between original and adversarial examples, using embeddings of image pixel values.

Removing the adversarial noise from the images is another approach introduced by **High-level representation Guided Denoiser** (HGD) [17] trains a Denoising AutoEncoder [134]

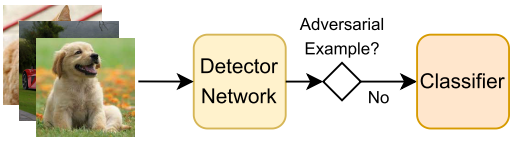


FIGURE 8. Schematic overview of the use of supplementary networks. The detector network was previously trained to detect adversarial images and is included between the input images and the classifier. This network receives the input images and determines if these images are adversarial or not. If they are not, they are redirected to the classifier; if they are, they are susceptible to human evaluation.

using the difference between original and adversarial images as the loss function and suppress the impact of adversarial examples. **Defense-GAN** [18] uses Wasserstein GANs (WGANs) [135] to accurately represent the set of clean training examples, which is attack-agnostic, and denoise the adversarial examples by minimizing the l_2 difference between the generator representation and the input image. **Embedding Regularized Classifier** (ER-Classifier) [136] uses an encoder and a discriminator to generate code vectors, reducing the dimensional space, and to distinguish these vectors from the ideal code vectors (sampled from a prior distribution), respectively, pushing the adversarial image into the clean image space.

Class Activation Feature-based Denoiser (CAFD) [137] is trained on samples generated by Class Activation Feature-based Attack to remove the noise from adversarial examples, by minimizing the distance between the class activation features and adversarial/natural examples. **Local Implicit Image Function** (LIIF) [138] trains an encoder to learn the construction of high resolution images, based on a coordinate and deep features around it, predicting the corresponding RGB value, and reduces the adversarial perturbation. **ADversarIal defenSe with local impliCit functiOns** (DISCO) [139] is an encoder that predicts the clean RGB value for a pixel, which receives an adversarial image and a query pixel location, by employing a local implicit module function. **Reverse Attacks** [140] uses a self-supervised approach to find the suitable additive perturbations that repair the adversarial example and reduce the perturbations during the testing phase.

Summary: The use of Supplementary Networks assume that additional networks/transforms are included between the input and classifier model to either detect or reduce the adversarial noise. The benefit from this type of approaches is not requiring any change to the original classifier nor retraining, which can be applied to in production models. These additional networks usually learn the representation space of the original images to remove adversarial noise or detect if an image differs from this space, making them suitable to deal with unknown adversarial attacks.

D. CHANGE NETWORK ARCHITECTURE

To create inherently robust networks, Guo et al. [141] proposes the use of Neural Architecture Search to select which networks are more robust to adversarial attacks, named as **RobNets**, discovering that densely connected

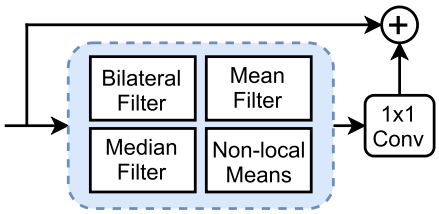


FIGURE 9. Overview of a feature denoising block [144], which can be included in the intermediate layers to make networks more robust. This method is an example of change network architecture.

patterns and convolution operations to direct connection edge are more robust. **HYDRA** [142] uses pruning techniques that are aware of the robust training objective, removing connections that negatively contribute to robustness, which creates compressed models with competitive standard and robust accuracy. **Dynamic Network Rewiring** (DNR) [143] creates pruned networks that have high robust and standard accuracy by unified constrained optimization formulation, using a hybrid loss function that merges model compression with robust adversarial training.

Zantedeschi et al. [145] discovered that Rectified Linear Units (ReLU) [146], [147] makes a small perturbation in the input accumulate with multiple layers and suggests the use of bounded ReLU (BReLU) [148] to prevent this accumulation. Benz et al. [149] found that BN [150] and other normalization techniques make CNNs more vulnerable to adversarial examples, suggesting that **Robust Features** are learned first and Non-Robust Features (which are the ones learned when using BN) afterward.

Feature Denoising (FD) [144] is a block that can be added to the intermediate layers of CNNs, consisting of feature-denoising operations, as described in Figure 9. **Input Random** [151] consists on adding layers at the beginning of CNNs that perform random resize, scaling the image to a random width and height, and random pad, padding with zeros around the resized image in a random manner. **Sparse Transformation Layer** (STL) [152] transforms the received images into a low-dimensional quasi-natural image space and can be included at the beginning of CNNs, reducing adversarial perturbations.

Summary: Different layers have distinct resilience to adversarial perturbations and some currently used techniques are more vulnerable. Multiple works propose the inclusion of specific layers to develop robust networks, yet require changes to the original architecture, hindering its application in production models.

E. PERFORM NETWORK VALIDATION

Datasets store their images using the Joint Photographic Experts Group (JPEG) [153] compression, Dziugaite et al. [154] evaluates the effect of JPEG, named as **JPG**, on the performance of FGSM, discovering JPEG compression can reverse smaller perturbations and does not have effect on perturbations with increased magnitude. Pang et al. [155] is the first to provide insights on implementation details influence the model robustness, herein named as **Bag of Tricks** (BT), concluding that: 1) the robustness is significantly affected by weight decay; 2) early stopping of the adversarial attacks may deteriorate worst-case robustness; and 3) smooth activation benefits lower capacity models. **Fixing Data Augmentation** (FDA) [156] focuses on reducing the robust overfitting by using different data augmentation techniques, demonstrating that model weight averaging combined with data augmentation schemes can significantly increase robustness, which is enhanced when using spatial composition techniques. Gowal et al. [157] systematically studied the effect of multiple training losses, model sizes, activation functions, the addition of unlabeled data, and other aspects, concluding that larger models with Swish/SiLU [158] activation functions and model weight averaging can reliably achieve state-of-the-art results in robust accuracy.

DeepXplore [159] uses multiple networks as cross-referencing oracles (to avoid manual checking) and evaluates inputs that trigger different behaviors and achieve high neuron coverage as a joint optimization problem solved by gradient-based search techniques, which is the first automated white-box testing framework. In the context of automated testing, **DeepGauge** [160] identifies a testbed containing multi-faceted representations using a set of multi-granularity testing criteria and evaluates the resilience of networks using two different strategies, primary function and corner-case behaviors, considering neuron- and layer-level coverage criteria. **Surprise Adequacy for Deep Learning Systems** (SADL) [161] introduces the Surprise of an input, consisting of the network behavior difference between an input and the learned data, which is used in the adequacy criterion (Surprise Adequacy), a metric to ensure the input surprise range coverage.

In the context of formal verification, a tool [162], using Satisfiability Modulo Theory (SMT), for automatic **Safety Verification** of the classification assumes that a decision is safe when, after applying transformations in the input, the model decision does not change. Similarly, multiple works have been inspired on the SMT with some variations, such as exploiting the piece-wise linear structure [163], [164] and extending the interval of the theory [165]. However, the previous works lack scalability for larger networks, leading to the proposal of [166], [167], and [168] that use LaGrangian relaxation, $l\infty$ norm-bounded ball, and includes other components besides ReLU. **AI**^{2} [169] and Mirman et al. [170] considered the abstract interpretation theory to represent the neural networks layers and activation functions, **PATROL** [171] uses game theory, and **Feature Partition Aggregation** [172] considers an ensemble of sub-models trained on different subsets.

Summary Software that is provided to the end consumer is often subject to a set of tests that validate its use in the real context, and it is a very important step to create a sense of security when using different devices. Similarly, the networks

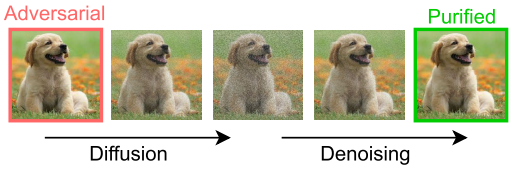


FIGURE 10. Overview of adversarial purification using denoising diffusion probabilistic models, adapted from [174]. The diffusion process is applied to an adversarial image, consisting of adding noise for a certain number of steps. During the denoising procedure, this noise is iteratively removed by the same amount of steps, resulting in a purified image (without perturbations).

should be subject to a set of tests that validate its performance in real context, using realistic attacks and images, ensuring their quality in real-world applications. These mechanisms provide a first insight on automated testing frameworks, formal verification, and implementation details that affect robustness, yet this topic remains highly unexplored.

F. ADVERSARIAL PURIFICATION

Adversarial Purification consists of defense mechanisms that remove adversarial perturbations using a generative model. **Improving Robustness Using Generated Data** (IRUGD) [173] explores how generative models trained on the original images can be leveraged to increase the size of the original datasets. Through extensive experiments, they concluded that Denoising Diffusion Probabilistic Models (DDPM) [174], a progression of diffusion probabilistic models [175], is the model that more closely resembles real data. Figure 10 presents the main idea behind the DDPM process.

Due to the great results in image synthesis displayed by the DDPM, Sehwag et al. [176] (**Proxy**) uses proxy distributions to significantly improve the performance of adversarial training by generating additional examples, demonstrating that the best generative models for proxy distribution are DDPM.

Inspired by previous works on adversarial purification [177], [178], **DiffPure** [179] uses DDPM for adversarial purification, receiving as input an adversarial example and recovering the clean image through a reverse generative process. Since this discovery, multiple improvements regarding the use of DDPM for Adversarial Purification have been studied. **Guided Diffusion Model for Adversarial Purification** (GDMAP) [180] receives as initial input pure Gaussian noise and gradually denoises it with guidance to an adversarial image.

DensePure [181] employs iterative denoising to an input image, with different random seeds, to get multiple reversed samples, which are given to the classifier and the final prediction is based on majority voting. Furthermore, Wang et al. [182] uses the most recent diffusion models [183] to demonstrate that diffusion models with higher efficiency and image quality directly translate into better robust accuracy.

Inspired by the need to evaluate the robustness of adversarial purification strategies, Lee et al. propose a gradual noise-scheduling strategy that improves the robustness of diffusion-based purification.

Summary: Recently, Diffusion Models have been proposed as a new technique to generate images based on their training data, that iteratively adds random noise to the clean image and applies reverse transformations to remove the noise and generate a new image. This process allows to reconstruct a new image (purified), removing the added adversarial perturbations, which leads to great performance under different adversarial attacks and can be applied to in production models. This process is called Adversarial Purification and its the most promising defense mechanism when it can be applied in real-time.

VI. ADVERSARIAL EFFECTS ON VISION TRANSFORMERS

ViTs are susceptible to adversarial perturbations that alter a patch in an image [184], demonstrating higher robustness, almost double, compared with ResNet-50 [185].

To further evaluate the **robustness of ViT** to adversarial examples, Mahmood et al. [186] used multiple adversarial attacks designed for CNNs and found out that ViT has increased robustness (compared with ResNet) for the majority of the attacks and has no resilience to the C&W and MI-FGSM attacks. Additionally, to complement the results obtained from the performance of ViTs, an extensive study [187] using feature maps and attention maps explains this performance visually. Bai et al. [188] suggests that ViTs and CNNs are being unfairly evaluated and provides a fair and in-depth comparison between them, indicating that ViTs are as vulnerable to adversarial perturbations as CNNs.

Attacks. The transferability of adversarial examples from CNNs to ViTs was also evaluated, suggesting that examples from CNNs do not instantly transfer to ViTs, led to the proposal of **Self-Attention blended Gradient Attack** (SAGA) [186] to misclassify both ViTs and CNNs. The **Pay No Attention** (PNA) [189] attack, which ignores the gradients of attention, and the **PatchOut** [189] attack, which randomly samples subsets of patches, demonstrate high transferability. **Architecture-oriented Transferable Attacking** (ATA) [190] is a framework that generates transferable adversarial examples by considering the common characteristics among different ViT architectures, discovering that attentional patch-wise regions significantly influence the model decision. **Patch-fool** [191] proposes a dedicated attack framework that fools the self-attention mechanism by attacking a single patch with multiple attention-aware optimization techniques.

Change Architecture. Multiple works have adopted techniques used in CNNs architecture to include in ViTs, such as pyramid structure [192], "soft" convolutional inductive bias [193], and dimension reduction principle [194]. Additionally, Tokens-To-Token ViT [195] propose a new way to analyze the input by recursively aggregating neighboring Tokens into one Token, and Swin Transformer [196] introduce the representation of images using shifted windows.



FIGURE 11. Images from MNIST [208] and Fashion-MNIST [209] in the first five and last five columns, respectively.

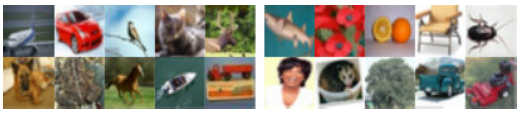


FIGURE 12. Images from CIFAR-10 [210] and CIFAR-100 [210] in the first five and last five columns, respectively.

Another issue with ViTs is the need to pretrain on large datasets and the time associated with it, which has been another research direction [197], [198]. **Robust Vision Transformer** (RVT) [199] reforms harmful components, proposing position-aware attention scaling and patch-wise augmentation, which can be applied to other ViTs. **Masked Autoencoder** (MAE) [200] partially masks the input tokens, with the visible patches being fed to the encoder and the masked patches added after the encoder and given to the decoder. Zhou et al. [201] discovered that self-attention promotes robustness through improved mid-level representations, proposing **Fully Attentional Networks** (FANs) which strengthen this capability. **Trade-off between Robustness and Accuracy of ViTs** (TORA-ViTs) [202] introduces two new components: a pair of accuracy and robustness adapters and a gated fusion module, which extract predictive and robust features and adjust their trade-off, respectively.

Adversarial Training. Smoothed ViTs [203] are trained using images composed of only one column of the original image, and the remaining are black, and the most predicted label is considered the correct one. **Pyramid Adversarial Training** (PyramidAT) [204] generates adversarial examples by perturbing the image at multiple scales and perform a ''matched'' Dropout and stochastic depth regularization, improving the robustness of ViTs. Gu et al. [205] evaluates the robustness of ViT to patch-wise perturbations, concluding that these models are more robust to naturally corrupted patches than CNNs while being more vulnerable to adversarially generated ones, proposing a simple **Temperature Scaling** based method that improves the robustness of ViTs. **Derandomized Smoothing** [206] uses a progressive smoothed image modeling task to train the ViTs, making them capture the more discriminating local context while preserving global semantic information, improving both robust and standard accuracy. **Token Gradient Regularization** (TGR) [207] reduces the variance of backpropagated gradient in internal blocks of ViTs in a token-wise manner and uses the regularized gradient to generate the adversarial examples.

Model Ensemble. PatchVeto [20] uses various transformers with different attention masks that output the encoding of the class and an image is only considered valid if all transformers reach a consensus in the voted class.

Summary: ViTs are a new technique, based on Transformers, to classify images that have similar or superior performance than CNNs and are more efficient. ViTs are also vulnerable to adversarial perturbations, yet there is not sufficient research on how susceptible are ViTs to the known adversarial perturbations and there are not many attacks designed specifically for ViTs. To the date it is known that the adversarial attacks designed for CNNs are not that easily transferred to ViTs, justifying further research on this topic, and changing the input and using techniques from CNN architecture display increased robustness. This section answers the proposed research question RQ1.

VII. DATASETS

One of the most used datasets is the **MNIST** [208] dataset, which contains images of handwritten digits collected from approximately 250 writers in shades of black and white, withdrawn from two different databases. Xiao et al. propose the creation of the **Fashion-MNIST** [209] dataset by using figures from a fashion website, which has a total size of 70,000 images, contains 10 classes, uses greyscale images, and each image has a size of 28 $\times$ 28. The MNIST and Fashion-MNIST datasets are divided into train and test sets, containing 60,000 and 10,000 examples, respectively.

Fig. 11 displays the 10 digits (from 0 to 9) from the MNIST dataset in the first five columns and the 10 fashion objects from Fashion-MNIST dataset in the last five columns. MNIST is one of the most widely studied datasets in the earlier works of adversarial examples, with defense mechanisms already displaying high robustness and accuracy on this dataset. The same does not apply to Fashion-MNIST, which requires further analysis, despite having similar characteristics to MNIST.

Another widely studied dataset is the CIFAR-10, which, in conjunction with the CIFAR-100 dataset, are subsets from a vast database containing 80 million tiny images [211], 32 $\times$ 32, and three color channels 75,062 different classes. **CIFAR-10** [210] contains only ten classes from this large database, with 6,000 images for each class, distributed into 50,000 training images and 10,000 test images. This dataset considers different objects, namely, animals and vehicles, usually found in different environments. **CIFAR-100** [210] contains 100 classes with only 600 images for each one with the same size and amount of color channels as the CIFAR-10 dataset. CIFAR-100 groups its 100 classes into 20 superclasses, located in different contexts/environments, making this dataset much harder to achieve high results.

Examples from the CIFAR-10 dataset are shown in Fig. 12 in the first five columns, and the remaining columns display examples of the superclasses from CIFAR-100. Due to the unsatisfactory results demonstrated by models trained on CIFAR-10, the CIFAR-100 dataset has not been included in most studies under the context of adversarial examples, suggesting that solving the issue of adversarial-perturbed images is still at its inception.

The **Street View House Numbers** (SVHN) [212] dataset provides the same challenge as MNIST: identifying which digits are present in a colored image, containing 10 classes



FIGURE 13. Images from street view house numbers [212] and german traffic sign recognition benchmark [213] in the first three and last three columns, respectively.

and an image size of 32 × 32 centered around a single character, with multiple digits in a single image. Regarding the dataset size, it has 630,420 digit images, but only 73,257 images are used for training, 26,032 images are used for testing, and the remaining 531,131 images can be used as additional training data. **German Traffic Sign Recognition Benchmark** (GTSRB) [213] is a dataset containing 43 classes of different traffic signs, demonstrating realistic scenarios, has 51,840 images, whose size varies from 15 × 15 to 222 × 193, divided into training, validation, and test sets with 50%, 25%, and 25%, respectively, of the total images.

The difficulties associated with the SVHN dataset are displayed in the first three rows of Fig. 13, showing unique digits that occupy the whole image and multiple digits on different backgrounds. The same figure presents the different types of traffic signs in the GTSRB dataset, such as prohibition, warning, mandatory, and end of prohibition. These datasets require further analysis, namely the SVHN, which has the same task as MNIST yet preliminaries results show that its performance is inferior. GTSRB has realistic scenarios that might already contain of realistic adversarial examples.

ImageNet [214] is one of the largest datasets for object recognition, containing 1,461,406 colored images and 1,000 classes, with images being resized to 224 × 224. This dataset collected photographs from Flickr, and other search engines, divided into 1.2 million training images, 50,000 validation images, and 100,000 test images. A possible alternative to ImageNet, when the dataset size is an important factor, is called **Tiny ImageNet** [215], a subset of ImageNet that contains fewer classes and images. This dataset contains only 200 classes (from the 1,000 classes in ImageNet), 100,000 training images, 10,000 validation images, and 10,000 test images. These classes include animals, vehicles, household items, insects, and clothing, considering the variety of contexts/environments that these objects can be found. Their images have a size of 64 × 64 and are colored.

ImageNet-A [216] is a subset of ImageNet, containing only 200 classes from the 1,000 classes, covering the broadest categories in ImageNet. ImageNet-A is a dataset composed of real-world adversarially filtered images, which were obtained by deleting the correctly predicted images by ResNet-50 classifiers, demonstrating transferability to unseen models, which makes this dataset suitable for evaluating the robustness of multiple classifiers. Two additional benchmarks, **ImageNet-C** [217] and **ImageNet-P** [217], were designed to evaluate the robustness of DNNs. The ImageNet-C standardizes and expands the corruption robustness topic, consisting of 75 corruptions applied to each image in the ImageNet validation set. ImageNet-P applies distortions to the images, though it differs from ImageNet-C because it contains perturbation sequences using only 10 common perturbations. Another benchmark is the **ImageNet-COLORDISTORT** (ImageNet-CD) [218], which considers multiple color distortions using different color space representations. This dataset contains the 1,000 classes from ImageNet, removing images without color channels, and the same image considers multiple color distortions under the Red Green Blue (RGB), Hue-Saturation-Value (HSV), CIELAB, and YCbCr color spaces considered common transformations used in image processing.

TABLE 3. Relevant characteristics to the context of adversarial examples of the mentioned datasets. #classes means the number of classes in the dataset. Empty color column means that the images in that dataset use greyscale or black and white shades. Datasets with ∗ are only used for testing.

Dataset	Size	#Classes	Classes	Color
MNIST	70,000	10	Digits	
Fashion-MNIST	70,000	10	Clothing	
CIFAR-10	60,000	10	Animals Vehicles	✓
SVHN	630,420	10	Digits	✓
GTSRB	51,840	43	Traffic Signs	✓
CIFAR-100	60,000	100	Household Items Outdoor Scenes	✓
Tiny ImageNet	120,000	200	Animals Household Items	✓
ImageNet-A*	7,500	200	Vehicles Food	✓
ImageNet-C*	3,750,000	200	Vehicles Food	✓
ImageNet-P*	15,000,000	200	Vehicles Food	✓
ImageNet	1,431,167	1,000	Vehicles Electronic devices	✓
ImageNet-CD*	736,515	1,000	Vehicles Electronic devices	✓

It is possible to observe a set of images withdrawn from ImageNet in the top left of Fig. 14. Additionally, some images misclassified by multiple classifiers (ImageNet-A) are shown in the top right of the same figure. The bottom represents the ImageNet with common corruptions and perturbations and is manipulated by multiple image techniques on the left and right, respectively. Table 3 summarizes the main characteristics of the datasets presented throughout this section. Recently, works are presenting their results in Tiny ImageNet dataset to estimate the performance of adversarial attacks on ImageNet dataset.

VIII. METRICS AND STATE-OF-THE-ART RESULTS

This section compiles the most important metrics that are applied in the context of adversarial attacks for Object Recognition and the state-of-the-art results for both CNNs and ViTs under multiple attack settings. Note that some of



FIGURE 14. Images from ImageNet [214] in the top left, from ImageNet-A [216] in the top right, from ImageNet-C and ImageNet-P datasets [217] in the bottom left, and ImageNet-COLORDISTORT [218] in the bottom right.

these metrics were not specifically designed for adversarial attacks, but they are used throughout the literature.

A. EVALUATION METRICS

1) ACCURACY

This metric measures the number of samples that are correctly predicted by the model, which is defined as:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{7}$$

where TP refers to True Positive, TN to True Negative, FP to False Positive, and FN to False Negative. When considering original images, this metric is denominated as *Clean Accuracy* and, when using adversarial images, is named as *Robust Accuracy*. Accuracy allows the comparison with previously proposed methods and clearly displays the performance of the model under clean and attack scenarios.

2) FOOLING RATE

The expected number of times the attack is able to flip the label, known as fooling rate, denoted by FR [219], measures the percentage of success for an adversarial attack, formally defined [220]:

$$\text{FR} = \frac{\sum_{i=1}^{N} (y(i) \neq \hat{y}(i))}{N}, \tag{8}$$

where $y(i)$ and $\hat{y}(i)$ refer to the pre-attack label and post-attack label associated with image i and N is the number of images. The Fooling Rate is an attack-specific metric that can measure how effectively the attack is on different images.

3) AVERAGE ROBUSTNESS

To objectively evaluate the robustness to adversarial perturbations of a classifier f, the average robustness $\hat{p}_{\text{adv}}(f)$ is defined as [28]:

$$\hat{p}_{\text{adv}}(f) = \frac{1}{\mathcal{D}} \sum_{x \in \mathcal{D}} \frac{\|\hat{r}(x)\|_2}{\|x\|_2}, \tag{9}$$

where $\hat{r}(x)$ is the estimated minimal perturbation obtained using the attack, and $\mathcal{D}$ denotes the test set.

TABLE 4. Accuracy comparison of different defense mechanisms on CIFAR-10 under auto-attack attack, l_∞ and $\epsilon = 8/255$, partially adapted from [139]. Clean and robust refers to accuracy without and with adversarial attacks, respectively.

Architecture	Defense Method	Year	Accuracy	
			Clean	Robust
WRN28-10	Input Random [151]	2017	94.3	8.6
	BAT [101]	2019	92.8	29.4
	FS [19]	2019	90.0	36.6
	Jpeg [154]	2016	83.9	50.7
	Pretrain [221]	2019	87.1	54.9
	UAT [84]	2019	86.5	56.0
	MART [91]	2020	87.5	56.3
	HYDRA [142]	2020	89.0	57.1
	RST [108]	2019	89.7	59.5
	GI-AT [103]	2020	89.4	59.6
	Proxy [176]	2021	89.5	59.7
	AWP [102]	2020	88.3	60.0
	FDA [156]	2021	87.3	60.8
	HAT [85]	2021	88.2	61.0
	SCORE [121]	2022	88.6	61.0
	PSSiLU [122]	2022	87.0	61.6
	Gowal *et al.* [157]	2020	89.5	62.8
	IRUGD [173]	2021	87.5	63.4
	Wang *et al.* [182]	2023	92.4	67.3
	STL [152]	2019	82.2	67.9
	DISCO [139]	2022	89.3	85.6
WRN34-10	Free-AT [92]	2019	86.1	41.5
	AT-PGD [41]	2018	87.1	44.0
	YOPO [93]	2019	87.2	44.8
	TLA [116]	2019	86.2	47.4
	LAT [94]	2019	87.8	49.1
	SAT [97]	2020	86.8	50.7
	FAT [96]	2022	85.3	51.1
	LBGAT [124]	2021	88.2	52.3
	TRADES [88]	2019	84.9	53.1
	SAT [109]	2020	83.5	53.3
	Friend-AT [98]	2020	84.5	55.5
	AWP [102]	2020	85.4	56.2
	LTD [107]	2021	85.2	56.9
	OA-AT [99]	2021	85.3	58.0
	Proxy [176]	2022	86.7	60.3
	HAT [85]	2021	91.5	62.8
WRN-70-16	SCORE [121]	2022	89.0	63.4
	IRUGD [173]	2021	91.1	65.9
	Gowal *et al.* [157]	2020	88.7	66.1
	FDA [156]	2021	92.2	66.6
	Wang *et al.* [182]	2023	93.3	70.7
	SODEF [128]	2021	93.7	71.3

4) DESTRUCTION RATE

To evaluate the impact of arbitrary transformations on adversarial images, the notion of destruction rate d is introduced and formally defined as [58]:

$$d = \frac{\sum_{k=1}^{n} C(\mathbf{X}^k, y^k_{\text{true}}) \neg C(\mathbf{X}^k_{\text{adv}}, y^k_{\text{true}}) C(T(\mathbf{X}^k_{\text{adv}}), y^k_{\text{true}})}{\sum_{k=1}^{n} C(\mathbf{X}^k, y^k_{\text{true}}) C(\mathbf{X}^k_{\text{adv}}, y^k_{\text{true}})}, \tag{10}$$

TABLE 5. Accuracy comparison of different defense mechanisms on CIFAR-10 under PGD attack, l_∞ and $\epsilon = 8/255$, results drawn from each paper. Clean and Robust refers to accuracy without and with adversarial attacks, respectively. Defenses with "–" on clean accuracy do not have a clean accuracy reported.

Defense Method	Year	Architecture	Accuracy	
			Clean	Robust
BPFC [115]	2020	ResNet-18	82.4	34.4
SNS [120]	2021	VGG-16	86.0	39.6
AT-MIFGSM [82]	2017	Inception v3	85.3	45.9
AT-PGD [41]	2018	ResNet-18	87.3	47.0
RobNets [141]	2020	RobNet-free	82.8	52.6
HGD [17]	2018	DUNET	92.4	53.1
RSLAD [106]	2021	ResNet-18	83.4	54.2
MART [91]	2020	WRN-28-10	83.1	55.6
TRADES [88]	2019	WRN-34-10	84.9	56.4
BagT [155]	2020	WRN-34-10	-	56.4
RO [114]	2020	ResNet-18	-	56.8
DOA [87]	2019	VGGFace	93.6	61.0
AWP [102]	2020	WRN-28-10	-	63.6
FS [19]	2019	WRN-28-10	90.0	68.6
CAFD [137]	2021	DUNET	91.1	87.2

TABLE 6. Accuracy comparison of different defense mechanisms on CIFAR-100 under auto-attack attack, l_∞ and $\epsilon = 8/255$, partially adapted from [139]. Clean and robust refers to accuracy without and with adversarial attacks, respectively.

Architecture	Defense Method	Year	Accuracy	
			Clean	Robust
WRN28-10	Input Random [151]	2017	73.6	3.3
	LIIF [138]	2021	80.3	3.4
	Bit Reduction [131]	2017	76.9	3.8
	Pretrain [221]	2019	59.2	28.4
	SCORE [121]	2022	63.7	31.1
	FDA [156]	2021	62.4	32.1
	Wang *et al.* [182]	2023	78.6	38.8
	Jpeg [154]	2016	61.9	39.6
	STL [152]	2019	67.4	46.1
	DISCO [139]	2022	72.1	67.9
WRN34-10	SAT [97]	2020	62.8	24.6
	AWP [102]	2020	60.4	28.9
	LBGAT [124]	2021	60.6	29.3
	OA-AT [99]	2021	65.7	30.4
	LTD [107]	2021	64.1	30.6
	Proxy [176]	2022	65.9	31.2
	DISCO [139]	2022	71.6	69.0
WRN-70-16	SCORE [121]	2022	65.6	33.1
	FDA [156]	2021	63.6	34.6
	Gowal *et al.* [157]	2020	69.2	36.9
	Wang *et al.* [182]	2023	75.2	42.7

where n is the number of images, $\mathbf{X}^k$ is the original image from the dataset, y^k_{true} is the true class of this image, $\mathbf{X}^k_{\text{adv}}$ is the adversarial image corresponding to that image, and T is an arbitrary image transformation. $\neg C(\mathbf{X}^k_{\text{adv}}, y^k_{\text{true}})$ is defined as the binary negation of $C(\mathbf{X}^k_{\text{adv}}, y^k_{\text{true}})$. Finally, the function $C(\mathbf{X}, y)$ is defined as [58]:

$$C(\mathbf{X}, y) = \begin{cases} 1, & \text{if image } \mathbf{X} \text{ is classified as } y; \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This metric can evaluate how resilient the perturbations are even when susceptible to multiple common transformations. This subsection answers the proposed research question RQ2.

B. DEFENSE MECHANISMS ROBUSTNESS

The most commonly used metric to evaluate models is accuracy, which evaluates the results on both original (**Clean Accuracy**) and adversarially perturbed (**Robust Accuracy**) datasets. One of the earliest and strongest adversarial attacks proposed was PGD, which is used by multiple defenses to evaluate their robustness. Table 5 displays defenses evaluated on CIFAR-10 under multiple steps PGD attack, ordered by increasing robustness. For the PGD attack, the best performing defenses are from approaches that use supplementary networks (CAFD) or modify the training process (FS and AWP). Overall, Wide ResNets [222] have better robust accuracy, due to high-capacity networks exhibiting greater adversarial robustness [41], [82], suggesting the usage of these networks in future developments of adversarial attacks and defenses.

To assess the robustness of defenses for white and black-box settings, Auto-Attack has gained increased interest over PGD in recent works. Tables 4, 6, and 7 present a set of defenses that are evaluated under Auto-Attack, on CIFAR-10, CIFAR-100, and ImageNet, respectively, ordered by increasing Robust Accuracy. The most used networks are Wide ResNets with different sizes, with the biggest Wide ResNet displaying better results overall, and the most resilient defense derives from the use of supplementary networks (DISCO), followed by modifying the train process (SODEF) and changing network architecture (STL). The results suggest that the inclusion of additional components to sanitize inputs of the targeted model (use of supplementary networks) is the most resilient approach for model robustness in white and black-box settings. The updated results for defenses under Auto-Attack can be found on the RobustBench [223] website.

We also provide results for ViTs on ImageNet-A dataset in Table 8, considering Vanilla ViTs and Robust ViTs, ordered by increasing Robust Accuracy. It is possible to observe that there is still a great margin of improvement for robustness in ViTs, achieving the best results when using a trade-off between accuracy and robustness (TORA-ViT-B/16) and partially occluding the input image (MAE-ViT-B). Finally, to provide a effortless introduction and exploration into the adversarial attacks topic, we collect the most used libraries in Python, implementations of the discussed attacks and defenses, presenting the respective GitHub URL, in Tables 9, 10, and 11. This subsection answers the proposed research question RQ3.

IX. PRACTICAL APPLICATIONS AND FUTURE DIRECTIONS

A. PRACTICAL APPLICATIONS

1) AUTONOMOUS DRIVING

Zhang et al. [225] proposes an end-to-end driving safety evaluation framework with designed performance metrics for the assessment of driving safety, suggesting that it is desirable to evaluate the robustness of deep learning models in terms of driving safety instead of model precision. Rossolini et al. [226] focuses on evaluating the robustness of segmentation models to real-world adversarial attacks and demonstrates that these already have a threshold of robustness. Wang et al. [227] explored how the manipulation of 3D point clouds can affect autonomous driving and a deeper analysis regarding adversarial reinforcement learning

TABLE 7. Accuracy comparison of different defense mechanisms on ImageNet under auto-attack attack, l_∞ and $\epsilon = 4/255$, partially adapted from [139]. Clean and robust refers to accuracy without and with adversarial attacks, respectively.

Architecture	Defense Method	Year	Accuracy	
			Clean	Robust
ResNet-18	Bit Reduction [131]	2017	67.6	4.0
	Jpeg [154]	2016	67.2	13.1
	Input Random [151]	2017	64.0	17.8
	Salman *et al.* [224]	2020	52.9	25.3
	STL [152]	2019	65.6	32.9
	DISCO [139]	2022	68.0	60.9
ResNet-50	Bit Reduction [131]	2017	73.8	1.9
	Input Random [151]	2017	74.0	18.8
	Cheap-AT [95]	2020	55.6	26.2
	Jpeg [154]	2016	73.6	33.4
	Salman *et al.* [224]	2020	64.0	35.0
	STL [152]	2019	68.3	50.2
	DISCO [139]	2022	72.6	68.2
WRN-50-2	Bit Reduction [131]	2017	75.1	5.0
	Input Random [151]	2017	71.7	23.6
	Jpeg [154]	2016	75.4	24.9
	Salman *et al.* [224]	2020	68.5	38.1
	DISCO [139]	2022	75.1	69.5

is required. For further details on attacks and defenses in the context of autonomous driving, we refer to a survey on this specific topic by Deng et al. [228].

2) MEDICAL IMAGING

Bortsova et al. [229] performed extensive experiments that show pretraining on ImageNet makes the models more vulnerable to transferability and the disparity between the data and architecture of the target and the substitute model can reduce the success of attacks. Tsai et al. [230] shows that all medical imaging is vulnerable to one-pixel attack and its effect would have devastating effects on the diagnosis of diseases. Kaviani et al. [231] and Apostolidis et al. [232] survey existing adversarial attacks and defenses in the context of medical imaging, emphasizing the urgent need to develop resistant mechanisms.

3) AUTHENTICATION

Currently, different methods are used to authenticate an user, giving it access to a device or location. However, these mechanisms also use artificial intelligence to evaluate if the person being authenticated has permissions or not. Tan et al. [233] evaluate how adversarially generated behavioral mouse dynamics can be used to bypass authentication mechanisms. Fingerprints are other example of authentication susceptible to adversarially crafted examples, using different methods [234], [235], [236]. Another daily-basis authentication mechanism is face recognition, which is vulnerable to more realistic attacks [237]. Wang et al. [238] provides a classification for attacks in authentication, reviewing common adversarial attacks and defenses.

4) OTHER AREAS

Natural Language Processing, which uses textual data, is also susceptible to adversarial attacks [239], with specific focus on *Social Network Applications* [240] and on Large Language Models, which have high relevance due to their broad scope.

TABLE 8. Accuracy comparison of different models and defense mechanisms on ImageNet-A, partially adapted from [202]. Clean and robust refers to accuracy without and with adversarial attacks, respectively.

Category	Model/Defense Method	Year	Accuracy	
			Clean	Robust
Vanilla ViT	ViT-B/16 [27]	2020	72.8	8.0
	ViT-B/16+MixUp [27]	2020	77.8	12.2
	ViT-B/16+CutMix [27]	2020	75.5	14.8
	ViT-B/16+AugReg [198]	2021	79.9	17.5
	PVT-Large [192]	2021	81.7	26.6
	DeiT-B/16 [197]	2021	82.0	27.4
	T2T-ViT$_{t-24}$ [195]	2021	82.6	28.9
	ConViT-B [193]	2021	82.4	29.0
	PiT-B [194]	2021	82.4	33.9
	Swin-B [196]	2021	83.4	35.8
Robust ViT	PyramidAT [204]	2022	81.7	23.0
	RVT-B [199]	2022	82.7	28.5
	FAN-L-ViT [201]	2022	83.9	34.2
	MAE-ViT-B [200]	2022	83.6	35.9
	TORA-ViT-B/16 [202]	2023	84.1	46.5

Financial Transactions are also a critical area that employs artificial intelligence, which is susceptible to adversarial attacks [241], [242]. *Smart Grids* that intelligently manage the distribution of electricity throughout a city are also vulnerable to adversarial attacks [243].

B. FUTURE DIRECTIONS

Following the *de facto* standards adopted by the literature, we suggest that future proposals of **defense mechanisms should be evaluated on Auto-Attack**, using the **robust accuracy** as a metric, and **Wide ResNets as standard architecture**. The adversarial defense that demonstrates better results is **Adversarial Training**, which should be a **requirement when evaluating attacks and defenses**. Additionally, we believe that standardization will directly degenerate into the increased efficacy of researching in the context of adversarial attacks.

The state-of-the-art results show that MNIST and CIFAR-10 datasets are already saturated. Other datasets should be further evaluated, namely: 1) **CIFAR-100 and ImageNet** since adversarial defenses do not achieve state-of-the-art clean accuracy (91% and 95%, respectively); 2) **GTSRB and SVHN**, depicting harder scenarios with greater variations of background, inclination, and luminosity; and 3) **Fashion-MNIST** that would allow better comprehension of which image properties influence DNNs performance. There should be an effort to **develop a metric/process to evaluate the amount of distortion** caused by new adversarial attacks, which would ease the expansion of adversarial attacks to other datasets that do not have a standardized ϵ value.

White-box attacks consider that the adversary has access to the network and training data, yet this is not feasible in real contexts, translating into the need of **focusing more on the development of black-box attacks**. A unique black-box set, **physical attacks, also require additional evaluation**, considering the properties of the real world and perturbations commonly found in it. Considering the increasing liberation of ML in the real world, end-users can partially control the training phase of DNNs, suggesting that **gray-box attacks**

TABLE 9. Most commonly used python libraries, containing multiple adversarial attacks and defenses, ordered by date of last update.

Name	Description	PT	TF	K	MX	Scikit	JAX	Last Update	Code
Torchattacks	Adversarial Attacks	✓						October 2023	URL
RobustBench	Datasets, Models, and Adversarial Attacks	✓						September 2023	URL
Adversarial Robustness Toolbox	Metrics, Models, and Adversarial Attacks	✓	✓	✓		✓		September 2023	URL
Advbox	Adversarial Attacks and Defenses	✓	✓	✓	✓			August 2022	URL
Foolbox	Adversarial Attacks	✓	✓				✓	May 2022	URL
advertorch	Adversarial Attacks and Defenses	✓						May 2022	URL

TABLE 10. Official and unofficial (identified with *) source code of discussed adversarial attacks.

Type	Name	PT	TF	Code
White-box Attacks	FGSM* [21]	✓		URL
	DeepFool [28]	✓		URL
	SF [29]	✓		URL
	SIAA [40]	✓	✓	URL
	PGD [41]		✓	URL
				URL
	MI-FGSM [42]		✓	URL
				URL
	C&W [43]		✓	URL
	MGAA [44]		✓	URL
	AdvGAN* [45]	✓		URL
	BIM* [58]	✓		URL
Universal Adversarial Perturbations	UAP [53]		✓	URL
	UAN [54]	✓		URL
	SGA [55]	✓		URL
	TRM-UAP [56]	✓		URL
	AdvEncoder [57]	✓		URL
	AdvCam [59]		✓	URL
Black-box Attacks	Auto-Attack [25]	✓		URL
	LQIA [62]		✓	URL
	GRA [68]		✓	URL
	SIA [69]	✓		URL
	Square Attack [70]	✓		URL
	StyLess [72]	✓		URL
	FIA [73]		✓	URL
	MSA [75]	✓		URL
	AdaEA [76]	✓		URL

TABLE 11. Official and unofficial (identified with *) source code of discussed adversarial defenses.

Type	Name	PT	TF	K	MX	Code
Adversarial Training	FScatter [19]	✓				URL
	RobDNN* [80]	✓				URL
	AdvScale* [82]		✓			URL
	EAT* [83]		✓			URL
	UAT [84]		✓			URL
	HAT [85]	✓				URL
	DOA [87]	✓				URL
	TRADES [88]	✓				URL
	MART [91]	✓				URL
	Free-AT [92]		✓			URL
		✓				URL
	YOPO [93]	✓				URL
	LAT [94]		✓			URL
	Cheap-AT [95]	✓				URL
	Friend-AT [98]	✓				URL
	AWP [102]	✓				URL
	GI-AT [103]	✓				URL
Modify the Training Process	DD* [104]	✓				URL
	E-DD* [105]	✓				URL
	RSLAD [106]	✓				URL
	RST [108]	✓				URL
	SAT [109]	✓				URL
	SENSEI [111]		✓			URL
	OR [114]	✓				URL
	TLA [116]		✓			URL
	CR* [117]	✓				URL
	BPFC [115]	✓				URL
	MRLS [118]	✓				URL
	CNLS [119]	✓				URL
	SCORE [121]	✓				URL
	LSA [123]				✓	URL
	LBGAT [124]	✓				URL
	SODEF [128]	✓				URL
Use of Supplementary Networks	HGD [17]	✓				URL
	Defense-GAN [18]			✓		URL
	MagNet* [129]			✓		URL
	FSqueez [131]		✓	✓		URL
	LIIF [138]	✓				URL
	DISCO [139]	✓				URL
	ReverseA [140]	✓				URL
Change Network Architecture	RobNets [141]	✓				URL
	HYDRA [142]	✓				URL
	DNR [143]	✓				URL
	FD [144]		✓			URL
	EffAdv* [145]			✓		URL
	Benz *et al.* [149]	✓				URL
	InputRand [151]		✓			URL
	STL [152]			✓		URL
Perform Network Validation	BT [155]	✓				URL
	FDA [156]	✓				URL
	Gowal *et al.* [157]	✓				URL
	DeepXplore [159]		✓			URL
	DeepGauge [160]		✓			URL
	SADL [161]			✓		URL
	SafeVeri [162]			✓		URL
Adversarial Purification	IRUGD [173]	✓				URL
	Proxy [176]	✓				URL
	SOAP [177]	✓				URL
	Yoon *et al.* [178]	✓				URL
	DiffPure [179]	✓				URL
	DensePure [181]	✓				URL
	Wang *et al.* [182]	✓				URL
	EDM [183]	✓				URL

will intensify (access only to network or data). Recent developments [47] point out to **UAPs that break the l_p bound but are semantically realistic** should be further analyzed.

The different network architectures are designed to increase the clean accuracy of DNNs in particular object recognition datasets, yet there should be **further evaluation on the impact of the different layers and their structure**. Additionally, there should be a focus on ensuring robustness through the **application of formal verification techniques**. ViTs introduce a new paradigm in image analysis and are more robust against natural corruptions, suggesting that **building ViT inherently robust to adversarial examples** might be a possible solution.

One of the widely adopted approach is the usage of supplementary networks that detect the existence of adversarial attacks, yet a **greater focus on evaluating the robustness of detector networks** is another issue that hinders the deployment of these defenses. DDPM are generative models that perform adversarial purification of images, but they can not be applied in real-time since they take up to dozens of seconds to create a single purified image. Therefore, an effort on **developing close to real-time adversarial purification strategies** is a viable strategy for future works.

The common factor among all the scenarios in adversarial attacks is the existence of an attacker and the **removal/withdrawal of attackers** from the network [244] would reduce the chance of security breaches. This section answers the proposed research question RQ4.

X. CONCLUSION

Networks are vulnerable to a set of inputs, denominated as adversarial examples, that modify their output and are constructed by adding a perturbation to the original image. This survey presents background concepts, such as adversary capacity and vector norms, and provides a comparison with existing surveys in the literature. Adversarial attacks are organized based on the adversary knowledge, highlighting the emphasis toward white box settings, and adversarial defenses are clustered into six domains, with adversarial training being the most popular strategy. We also present the latest developments of adversarial settings in ViTs, describe the commonly used datasets and metrics, providing results in CIFAR-10, CIFAR-100, and ImageNet, and provide the source code for adversarial attacks and defenses, finishing with a set of open issues that remain less explored.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[3] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[4] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 471–478.

[5] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45.

[7] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[8] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.

[9] A. Ndikumana, N. H. Tran, D. H. Kim, K. T. Kim, and C. S. Hong, "Deep learning based caching for self-driving cars in multi-access edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 5, pp. 2862–2877, May 2021.

[10] Z. Yuan, Y. Lu, Z. Wang, and Y. Xue, "Droid-sec: Deep learning in Android malware detection," in *Proc. ACM Conf. SIGCOMM*, NY, NY, USA, Aug. 2014, pp. 371–372.

[11] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," *IEEE Access*, vol. 7, pp. 46717–46738, 2019.

[12] X. Zhou, W. Liang, K. I. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[13] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with EMRs," in *Proc. IEEE Int. Conf. (BIBM)*, Nov. 2014, pp. 556–559.

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[15] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Mar. 2016, pp. 372–387.

[16] *Vertex AI Pricing*, Google, Los Angeles, CA, USA, 2022.

[17] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1778–1787.

[18] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17.

[19] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Proc. NIPS*, 2019, pp. 1–11.

[20] Y. Huang and Y. Li, "Zero-shot certified defense against adversarial patches with vision transformers," 2021, *arXiv:2111.10481*.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[22] T. Long, Q. Gao, L. Xu, and Z. Zhou, "A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions," *Comput. Secur.*, vol. 121, Oct. 2022, Art. no. 102847.

[23] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "Adversarial attack and defense: A survey," *Electronics*, vol. 11, no. 8, p. 1283, Apr. 2022.

[24] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–39, Aug. 2023.

[25] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2206–2216.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Lake Tahoe, NV, USA: Curran Associates, 2012, pp. 1–10.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[28] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[29] A. Dabouei, S. Soleymani, F. Taherkhani, J. M. Dawson, and N. M. Nasrabadi, "Smoothfool: An efficient framework for computing smooth adversarial perturbations," in *Proc. IEEE WACV*, Oct. 2020, pp. 2654–2663.

[30] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[31] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, p. 909, Mar. 2019.

[32] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–38, May 2021.

[33] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, Apr. 2020.

[34] K. Sadeghi, A. Banerjee, and S. K. S. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 4, pp. 450–467, Aug. 2020.

[35] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 25–45, Mar. 2021.

[36] J.-X. Mi, X.-D. Wang, L.-F. Zhou, and K. Cheng, "Adversarial examples based on object detection tasks: A survey," *Neurocomputing*, vol. 519, pp. 114–126, Jan. 2023.

[37] L. Amritasree and C. P. Prathibhamol, "Exploring adversarial attacks and countermeasures in image classification and object detection: A survey," *SSRN Electron. J.*, Jan. 2023, doi: 10.2139/ssrn.4607819.

[38] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[39] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," 2017, *arXiv:1704.03453*.

[40] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4723–4731.

[41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.

[42] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. CVPR*, Jun. 2018, pp. 9185–9193.
[43] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[44] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan, "Meta gradient adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7748–7757.
[45] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3905–3911.
[46] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–9.
[47] X. Chen, X. Gao, J. Zhao, K. Ye, and C.-Z. Xu, "Advdiffuser: Natural adversarial example synthesis with diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4562–4572.
[48] P. Zhu, G. Osada, H. Kataoka, and T. Takahashi, "Frequency-aware GAN for adversarial manipulation generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4315–4324.
[49] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.
[50] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. 25th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.
[51] C.-H. Weng, Y.-T. Lee, and S.-H. B. Wu, "On the trade-off between adversarial and backdoor robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Curran Associates, 2020, pp. 11973–11983.
[52] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.
[53] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 86–94.
[54] J. Hayes and G. Danezis, "Learning universal adversarial perturbations with generative models," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 43–49.
[55] X. Liu, Y. Zhong, Y. Zhang, L. Qin, and W. Deng, "Enhancing generalization of universal adversarial perturbation through gradient aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4435–4444.
[56] Y. Liu, X. Feng, Y. Wang, W. Yang, and D. Ming, "TRM-UAP: Enhancing the transferability of data-free universal adversarial perturbation via truncated ratio maximization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4762–4771.
[57] Z. Zhou, S. Hu, R. Zhao, Q. Wang, L. Y. Zhang, J. Hou, and H. Jin, "Downstream-agnostic adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4345–4355.
[58] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Artif. Intell. Saf. Secur.* Boca Raton, FL, USA: CRC Press, 2018, pp. 99–112.
[59] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 997–1005.
[60] D. Wang, W. Yao, T. Jiang, C. Li, and X. Chen, "RFLA: A stealthy reflected light adversarial attack in the physical world," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4455–4465.
[61] X. Wei, Y. Huang, Y. Sun, and J. Yu, "Unified adversarial patch for cross-modal attacks in the physical world," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 4445–4454, 2023.
[62] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. ICML*, 2018, pp. 2137–2146.
[63] Z. Li, H. Cheng, X. Cai, J. Zhao, and Q. Zhang, "SA-ES: Subspace activation evolution strategy for black-box adversarial attacks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 780–790, Jun. 2023.
[64] M. Wicker, X. Huang, and M. Kwiatkowska, "Feature-guided black-box safety testing of deep neural networks," in *Proc. TACAS*, 2018, pp. 408–426.
[65] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.
[66] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Sep. 2020, pp. 1277–1294.
[67] X. Yuan, L. Ding, L. Zhang, X. Li, and D. O. Wu, "ES attack: Model stealing against deep neural networks without data hurdles," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1258–1270, Oct. 2022.
[68] H. Zhu, Y. Ren, X. Sui, L. Yang, and W. Jiang, "Boosting adversarial transferability via gradient relevance attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4741–4750.
[69] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4607–4619.
[70] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 484–501.
[71] Z. Wang, H. Yang, Y. Feng, P. Sun, H. Guo, Z. Zhang, and K. Rent, "Towards transferable targeted adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20534–20543.
[72] K. Liang and B. Xiao, "StyLess: Boosting the transferability of adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8163–8172.
[73] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7619–7628.
[74] H. Zanddizari, B. Zeinali, and J. M. Chang, "Generating black-box adversarial examples in sparse domain," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 4, pp. 795–804, Aug. 2022.
[75] P. N. Williams and K. Li, "Black-box sparse adversarial attack via multi-objective optimisation CVPR proceedings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12291–12301.
[76] B. Chen, J. Yin, S. Chen, B. Chen, and X. Liu, "An adaptive model ensemble adversarial attack for boosting adversarial transferability," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4489–4498.
[77] Y. Qian, S. He, C. Zhao, J. Sha, W. Wang, and B. Wang, "LEA2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4510–4521.
[78] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4630–4639.
[79] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2196–2205.
[80] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," 2014, *arXiv:1412.5068*.
[81] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proc. CVPR*, Jun. 2016, pp. 4480–4488.
[82] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–17.
[83] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–11.
[84] J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli, "Are labels required for improving adversarial robustness?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.
[85] R. Rade and S.-M. Moosavi-Dezfooli, "Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off," in *Proc. ICML*, 2021, pp. 1–10.
[86] C. K. Mummadi, T. Brox, and J. H. Metzen, "Defending against universal perturbations with shared adversarial training," in *Proc. IEEE/CVF (ICCV)*, Oct. 2019, pp. 4927–4936.
[87] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–26.
[88] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

[89] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–23.

[90] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy?—A comprehensive study on the robustness of 18 deep image classification models," in *Proc. ECCV*, 2018, pp. 631–648.

[91] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.

[92] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[93] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, "You only propagate once: Accelerating adversarial training via maximal principle," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[94] N. Kumari, M. Singh, A. Sinha, H. Machiraju, B. Krishnamurthy, and V. N. Balasubramanian, "Harnessing the vulnerability of latent layers in adversarially trained models," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2779–2785.

[95] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[96] J. Chen, Y. Cheng, Z. Gan, Q. Gu, and J. Liu, "Efficient robust training via backward smoothing," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 6, pp. 6222–6230.

[97] C. Sitawarin, S. Chakraborty, and D. Wagner, "SAT: Improving adversarial training via curriculum-based loss smoothing," 2020, *arXiv:2003.09347*.

[98] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *Proc. ICML*, 2020, pp. 11278–11287.

[99] S. Addepalli, S. Jain, G. Sriramanan, S. Khare, and V. B. Radhakrishnan, "Towards achieving adversarial robustness beyond perceptual limits," in *Proc. ICML Workshop Adversarial Mach. Learn.*, 2021, pp. 1–24.

[100] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. (CVPR)*, Jun. 2018, pp. 586–595.

[101] J. Wang and H. Zhang, "Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks," in *Proc. IEEE/CVF (ICCV)*, Oct. 2019, pp. 6629–6638.

[102] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *Proc. NIPS*, 2020, pp. 2958–2969.

[103] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–29.

[104] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[105] N. Papernot and P. McDaniel, "Extending defensive distillation," 2017, *arXiv:1705.05264*.

[106] B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang, "Revisiting adversarial robustness distillation: Robust soft labels make student better," in *Proc. ICCV*, Oct. 2021, pp. 16443–16452.

[107] E.-C. Chen and C.-R. Lee, "LTD: Low temperature distillation for robust adversarial training," 2021, *arXiv:2111.02331*.

[108] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[109] L. Huang, C. Zhang, and H. Zhang, "Self-adaptive training: Beyond empirical risk minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19365–19376.

[110] K. Chalupka, P. Perona, and F. Eberhardt, "Visual causal feature learning," in *Proc. UAI*, 2015, pp. 1–15.

[111] X. Gao, R. K. Saha, M. R. Prasad, and A. Roychoudhury, "Fuzz testing based data augmentation to improve robustness of deep neural networks," in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. (ICSE)*, Oct. 2020, pp. 1147–1158.

[112] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[113] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[114] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 8093–8104.

[115] S. Addepalli, B. S. Vivek, A. Baburaj, G. Sriramanan, and R. V. Babu, "Towards achieving adversarial robustness by enforcing feature consistency across bit planes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1017–1026.

[116] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[117] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proc. IEEE/CVF Conf. (CVPR)*, Jun. 2019, pp. 9078–9086.

[118] C. Jin and M. Rinard, "Manifold regularization for locally stable deep neural networks," 2020, *arXiv:2003.04286*.

[119] M. Atzmon, N. Haim, L. Yariv, O. Israelov, H. Maron, and Y. Lipman, "Controlling neural level sets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[120] C. Zhang, A. Liu, X. Liu, Y. Xu, H. Yu, Y. Ma, and T. Li, "Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity," *IEEE Trans. Image Process.*, vol. 30, pp. 1291–1304, 2021.

[121] T. Pang, M. Lin, X. Yang, J. Zhu, and S. Yan, "Robustness and accuracy could be reconcilable by (proper) definition," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17258–17277.

[122] S. Dai, S. Mahloujifar, and P. Mittal, "Parameterizing activation functions for adversarial robustness," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2022, pp. 80–87.

[123] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvari, "Learning with a strong adversary," 2015, *arXiv:1511.03034*.

[124] J. Cui, S. Liu, L. Wang, and J. Jia, "Learnable boundary guided adversarial training," in *Proc. IEEE/CVF (ICCV)*, Jun. 2021, pp. 15721–15730.

[125] Y. Hanshu, D. Jiawei, T. Vincent, and F. Jiashi, "On robustness of neural ordinary differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[126] E. Haber and L. Ruthotto, "Stable architectures for deep neural networks," *Inverse Problems*, vol. 34, no. 1, Jan. 2018, Art. no. 014004.

[127] X. Liu, T. Xiao, S. Si, Q. Cao, S. Kumar, and C.-J. Hsieh, "How does noise help robustness? Explanation and exploration under the neural sde framework," in *Proc. IEEE/CVF Conf. (CVPR)*, 2020, pp. 282–290.

[128] Q. Kang, Y. Song, Q. Ding, and W. P. Tay, "Stable neural ODE with Lyapunov-stable equilibrium points for defending against adversarial attacks," in *Proc. NeurIPS*, vol. 34, 2021, pp. 14925–14937.

[129] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017.

[130] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–12.

[131] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.

[132] A. Abusnaina, Y. Wu, S. Arora, Y. Wang, F. Wang, H. Yang, and D. Mohaisen, "Adversarial example detection using latent neighborhood graph," in *Proc. IEEE/CVF (ICCV)*, Oct. 2021, pp. 7687–7696.

[133] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[134] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, 2008, pp. 1096–1103.

[135] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[136] Y. Li, M. R. Min, T. Lee, W. Yu, E. Kruus, W. Wang, and C.-J. Hsieh, "Towards robustness of deep neural networks via regularization," in *Proc. IEEE/CVF (ICCV)*, Oct. 2021, pp. 7496–7505.

[137] D. Zhou, N. Wang, C. Peng, X. Gao, X. Wang, J. Yu, and T. Liu, "Removing adversarial noise in class activation feature space," in *Proc. IEEE/CVF (ICCV)*, Oct. 2021, pp. 7878–7887.

[138] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8624–8634.

[139] C.-H. Ho and N. Vasconcelos, "Disco: Adversarial defense with local implicit functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. New Orleans, LA, USA: Curran Associates, 2022, pp. 23818–23837.

[140] C. Mao, M. Chiquier, H. Wang, J. Yang, and C. Vondrick, "Adversarial attacks are reversible with natural supervision," in *Proc. IEEE/CVF (ICCV)*, Oct. 2021, pp. 661–671.
[141] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When NAS meets robustness: In search of robust architectures against adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 628–637.
[142] V. Sehwag, S. Wang, P. Mittal, and S. Jana, "Hydra: Pruning adversarially robust neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19655–19666.
[143] S. Kundu, M. Nazemi, P. A. Beerel, and M. Pedram, "DNR: A tunable robust pruning framework through dynamic network rewiring of DNNs," in *Proc. 26th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2021, pp. 344–350.
[144] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–509.
[145] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 39–49.
[146] A. Fred Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.
[147] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, Jun. 2000.
[148] S. S. Liew, M. Khalil-Hani, and R. Bakhteri, "Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems," *Neurocomputing*, vol. 216, pp. 718–734, Dec. 2016.
[149] P. Benz, C. Zhang, and I. S. Kweon, "Batch normalization increases adversarial vulnerability: Disentangling usefulness and robustness of model features," 2020, *arXiv:2010.03316*.
[150] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[151] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.
[152] B. Sun, N.-H. Tsai, F. Liu, R. Yu, and H. Su, "Adversarial defense by stratified convolutional sparse coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11439–11448.
[153] W. F. Good, G. S. Maitz, and D. Gur, "Joint photographic experts group (JPEG) compatible data compression of mammograms," *J. Digit. Imag.*, vol. 7, no. 3, pp. 123–132, Aug. 1994.
[154] G. Karolina Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of JPG compression on adversarial images," 2016, *arXiv:1608.00853*.
[155] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.
[156] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Fixing data augmentation to improve adversarial robustness," 2021, *arXiv:2103.01946*.
[157] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," 2020, *arXiv:2010.03593*.
[158] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.
[159] K. Pei, Y. Cao, J. Yang, and S. S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *Proc. 26th Symp. Operating Syst. Princ.*, 2017, pp. 1–18.
[160] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu, J. Zhao, and Y. Wang, "DeepGauge: Multi-granularity testing criteria for deep learning systems," in *Proc. 33rd IEEE/ACM Int. Conf. ASE*, Sep. 2018, pp. 120–131.
[161] J. Kim, R. Feldt, and S. Yoo, "Guiding deep learning system testing using surprise adequacy," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng. (ICSE)*, May 2019, pp. 1039–1049.
[162] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *Proc. CAV*, 2017, pp. 3–29.
[163] R. R. Bunel, I. Turkaslan, P. Torr, P. Kohli, and P. K. Mudigonda, "A unified view of piecewise linear neural network verification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.
[164] V. Tjeng, K. Y. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–21.
[165] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *Computer Aided Verification*. Heidelberg, Germany: Springer, 2017, pp. 97–117.
[166] K. Dvijotham, R. Stanforth, S. Gowal, T. A. Mann, and P. Kohli, "A dual approach to scalable verification of deep networks," in *Proc. UAI*, vol. 1, 2018, p. 3.
[167] S. Gowal, K. D. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli, "Scalable verified training for provably robust image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4842–4851.
[168] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.
[169] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "AI2: Safety and robustness certification of neural networks with abstract interpretation," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 3–18.
[170] M. Mirman, T. Gehr, and M. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3578–3586.
[171] W. Guo, X. Wu, L. Wang, X. Xing, and D. Song, "Patrol: Provable defense against adversarial policy in two-player games," in *Proc. 32nd USENIX Conf. Secur. Symp.*, 2023, pp. 3943–3960.
[172] Z. Hammoudeh and D. Lowd, "Provable robustness against a union of L_0 adversarial attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 21134–21142.
[173] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann, "Improving robustness using generated data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 4218–4233.
[174] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
[175] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML*, 2015, pp. 2256–2265.
[176] V. Sehwag, S. Mahloujifar, T. Handina, S. Dai, C. Xiang, M. Chiang, and P. Mittal, "Robust learning meets generative models: Can proxy distributions improve adversarial robustness?" in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–30.
[177] C. Shi, C. Holtz, and G. Mishne, "Online adversarial purification based on self-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–15.
[178] J. Yoon, S. J. Hwang, and J. Lee, "Adversarial purification with score-based generative models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12062–12072.
[179] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 16805–16827.
[180] Q. Wu, H. Ye, and Y. Gu, "Guided diffusion model for adversarial purification from random noise," 2022, *arXiv:2206.10875*.
[181] Z. Chen, K. Jin, J. Wang, W. Nie, M. Liu, A. Anandkumar, B. Li, and D. Song, "DensePure: Understanding diffusion models towards adversarial robustness," in *Proc. NIPS*, 2022, pp. 1–20.
[182] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, "Better diffusion models further improve adversarial training," 2023, *arXiv:2302.04638*.
[183] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds. 2022, pp. 26565–26577.
[184] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.
[185] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.
[186] K. Mahmood, R. Mahmood, and M. van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7838–7847.

[187] A. Aldahdooh, W. Hamidouche, and O. Deforges, "Reveal of vision transformers robustness against adversarial attacks," 2021, *arXiv:2106.03734*.

[188] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than CNNs?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26831–26843.

[189] Z. Wei, J. Chen, M. Goldblum, Z. Wu, T. Goldstein, and Y.-G. Jiang, "Towards transferable adversarial attacks on vision transformers," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 2668–2676.

[190] Y. Wang, J. Wang, Z. Yin, R. Gong, J. Wang, A. Liu, and X. Liu, "Generating transferable adversarial examples against vision transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 5181–5190.

[191] Y. Fu, S. Zhang, S. Wu, C. Wan, and Y. Lin, "Patch-fool: Are vision transformers always robust against adversarial perturbations?" in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–18.

[192] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[193] S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2286–2296.

[194] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 11936–11945.

[195] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. ICCV*, Oct. 2021, pp. 558–567.

[196] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[197] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[198] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? Data, augmentation, and regularization in vision transformers," *Trans. Mach. Learn. Res.*, May 2022.

[199] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12032–12041.

[200] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[201] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Alvarez, "Understanding the robustness in vision transformers," in *Proc. ICML*, 2022, pp. 27378–27394.

[202] Y. Li and C. Xu, "Trade-off between robustness and accuracy of vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7558–7568.

[203] H. Salman, S. Jain, E. Wong, and A. Madry, "Certified patch robustness via smoothed vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15116–15126.

[204] C. Herrmann, K. Sargent, L. Jiang, R. Zabih, H. Chang, C. Liu, D. Krishnan, and D. Sun, "Pyramid adversarial training improves ViT performance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13409–13419.

[205] J. Gu, V. Tresp, and Y. Qin, "Are vision transformers robust to patch perturbations?" in *Computer Vision—ECCV*. Tel Aviv-Yafo, Israel: Springer, 2022, pp. 404–421.

[206] Z. Chen, B. Li, J. Xu, S. Wu, S. Ding, and W. Zhang, "Towards practical certifiable patch defense with vision transformer," in *Proc. IEEE/CVF Conf. (CVPR)*, Oct. 2022, pp. 15148–15158.

[207] J. Zhang, Y. Huang, W. Wu, and M. R. Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 16415–16424.

[208] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[209] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[210] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[211] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[212] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS*, 2011, pp. 1–9.

[213] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Netw.*, vol. 32, pp. 323–332, Aug. 2012.

[214] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[215] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[216] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE/CVF Conf. (CVPR)*, pp. 15262–15271, Jun. 2021.

[217] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–16.

[218] K. De and M. Pedersen, "Impact of colour on robustness of deep neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 21–30.

[219] Z. Huan, Y. Wang, X. Zhang, L. Shang, C. Fu, and J. Zhou, "Data-free adversarial perturbations for practical black-box attack," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Springer, 2020, pp. 127–138.

[220] K. R. Mopuri, V. Shaj, and R. V. Babu, "Adversarial fooling beyond 'flipping the label,'" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 778–779.

[221] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 2712–2721.

[222] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–15.

[223] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," in *Proc. 34th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–29.

[224] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do adversarially robust ImageNet models transfer better?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3533–3545.

[225] J. Zhang, Y. Lou, J. Wang, K. Wu, K. Lu, and X. Jia, "Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3443–3456, Mar. 2022.

[226] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, and G. Buttazzo, "On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 2, 2023, doi: 10.1109/TNNLS.2023.3314512.

[227] X. Wang, M. Cai, F. Sohel, N. Sang, and Z. Chang, "Adversarial point cloud perturbations against 3D object detection in autonomous driving systems," *Neurocomputing*, vol. 466, pp. 27–36, Nov. 2021.

[228] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, and Q.-L. Han, "Deep learning-based autonomous driving systems: A survey of attacks and defenses," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 7897–7912, Dec. 2021.

[229] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. W. Pluim, M. Veta, C. I. Sánchez, and M. de Bruijne, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102141.

[230] M.-J. Tsai, P.-Y. Lin, and M.-E. Lee, "Adversarial attacks on medical image classification," *Cancers*, vol. 15, no. 17, p. 4228, Aug. 2023.

[231] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," *Exp. Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116815.

[232] K. D. Apostolidis and G. A. Papakostas, "A survey on adversarial deep learning robustness in medical image analysis," *Electronics*, vol. 10, no. 17, p. 2132, Sep. 2021.

[233] Y. X. Marcus Tan, A. Iacovazzi, I. Homoliak, Y. Elovici, and A. Binder, "Adversarial attacks on remote user authentication using behavioural mouse dynamics," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–10.
[234] S. Karunaratne, E. Krijestorac, and D. Cabric, "Penetrating RF fingerprinting-based authentication with a generative adversarial attack," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[235] S. Marrone and C. Sansone, "On the transferability of adversarial perturbation attacks against fingerprint based authentication systems," *Pattern Recognit. Lett.*, vol. 152, pp. 253–259, Dec. 2021.
[236] S. Marrone and C. Sansone, "Adversarial perturbations against fingerprint based authentication systems," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–6.
[237] S. Joos, T. Van Hamme, D. Preuveneers, and W. Joosen, "Adversarial robustness is not enough: Practical limitations for securing facial authentication," in *Proc. ACM Int. Workshop Secur. Privacy Analytics*, Apr. 2022, pp. 2–12.
[238] X. Wang, Z. Yan, R. Zhang, and P. Zhang, "Attacks and defenses in user authentication systems: A survey," *J. Netw. Comput. Appl.*, vol. 188, Aug. 2021, Art. no. 103080.
[239] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran, "A survey of adversarial defenses and robustness in NLP," *ACM Comput. Surv.*, vol. 55, no. 14, pp. 1–39, Dec. 2023.
[240] I. Alsmadi, K. Ahmad, M. Nazzal, F. Alam, A. Al-Fuqaha, A. Khreishah, and A. Algosaibi, "Adversarial NLP for social network applications: Attacks, defenses, and research directions," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 6, pp. 3089–3108, Dec. 2023.
[241] I. Fursov, M. Morozov, N. Kaploukhaya, E. Kovtun, R. Rivera-Castro, G. Gusev, D. Babaev, I. Kireev, A. Zaytsev, and E. Burnaev, "Adversarial attacks on deep models for financial transaction records," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2868–2878.
[242] M. Goldblum, A. Schwarzschild, A. Patel, and T. Goldstein, "Adversarial attacks on machine learning systems for high-frequency trading," in *Proc. 2nd ACM Int. Conf. AI Finance*, Nov. 2021, pp. 1–9.
[243] J. Hao and Y. Tao, "Adversarial attacks on deep learning models in smart grids," *Energy Rep.*, vol. 8, pp. 123–129, May 2022.
[244] Z. Wang, S. Li, X. Zhang, and G. Feng, "Exploring abnormal behavior in swarm: Identify user using adversarial examples," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 250–260, Feb. 2023.



JOANA C. COSTA received the bachelor's and master's degrees in computer science and engineering from the Universidade da Beira Interior (UBI), in 2019 and 2021, respectively. She is currently pursuing the Ph.D. degree in computer vision and adversarial attacks with the Universidade da Beira Interior, funded by Fundação para a Ciência e a Tecnologia (FCT).



TIAGO ROXO received the bachelor's degree in computer science and engineering from the Universidade da Beira Interior (UBI), in 2019. He is currently pursuing the Ph.D. degree in computer vision and artificial intelligence with the Universidade da Beira Interior, funded by Fundação para a Ciência e a Tecnologia (FCT).



HUGO PROENÇA (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees, in 2001, 2004, and 2007, respectively. He is currently a Full Professor with the Department of Computer Science, University of Beira Interior, Portugal, has been researching mainly biometrics and visual-surveillance. He is a member of the Editorial Board of the *Image and Vision Computing*, IEEE Access, and *International Journal of Biometrics*. Also, he served as a Guest Editor for the Special Issue of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image and Video Processing*. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and an Area Editor (ocular biometrics) of the IEEE Biometrics Compendium.



PEDRO RICARDO MORAIS INÁCIO (Senior Member, IEEE) received the B.Sc. degree in mathematics/computer science and the Ph.D. degree in computer science and engineering from the University of Beira Interior (UBI), Portugal, in 2005 and 2009, respectively.

He is currently an Associate Professor with the Department of Computer Science, UBI, where he joined, in 2010. Lectures subjects related with information assurance and (cyber)security and computer-based simulation, to graduate and undergraduate courses, namely to the B.Sc., M.Sc., and Ph.D. programmes in computer science and engineering. He is currently the Pro-Rector of the Digital University and the Data Protection Officer of UBI. His Ph.D. work was performed in the enterprise environment of Nokia Siemens Networks Portugal S.A., through a Ph.D. grant from the Portuguese Foundation for Science and Technology. His main research interests include information assurance and security, computer-based simulation, network traffic monitoring, analysis, and classification. He has more than 70 publications in the form of book chapters and papers in international peer-reviewed books, conferences, and journals. He is an ACM Professional Member and a Researcher of the Instituto de Telecomunicaμes (IT). He frequently reviews articles of IEEE, Springer, Wiley, and Elsevier journals. He is a member of the Technical Program Committees of flagship national and international workshops and conferences, such as ACM SAC, IEEE NCA, IFIPSEC, or ARES.

Bibliography

[1] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935. xxviii, 15, 17, 83, 87, 88, 93, 95, 99, 103, 104, 105, 106, 120, 129, 130, 132, 133, 134, 135, 136, 137, 138

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. xxx, 70, 97, 116, 117, 118

[3] T. Roxo, J. C. Costa, P. R. M. Inácio, and H. Proença, "Wasd: A wilder active speaker detection dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 7, no. 1, pp. 61–70, 2025. xxxiii, xxxv, 16, 17, 45, 81, 97, 99, 103, 104, 105, 120, 121, 128, 129, 133, 134, 139

[4] T. Roxo and H. Proença, "Bias: A body-based interpretable active speaker approach," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 573–582, 2024. xxxiii, 15, 17, 97, 127, 129, 131, 134, 135, 136

[5] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850. xxxiii, 21

[6] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool, "P3depth: Monocular depth estimation with a piecewise planarity prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1610–1621. xxxiii, 30, 32

[7] K. Lewin, "Field theory in social science: selected theoretical papers (edited by dorwin cartwright.)." 1951. 9

[8] M. T. Anguera, "Observational methods (general)," *Encyclopedia of psychological assessment*, vol. 2, pp. 632–637, 2003. 9

[9] V. Tran, "Positive affect negative affect scale (panas)," *Encyclopedia of Behavioral Medicine*, pp. 1508–1509, 2013. 9

[10] Z. Hao, M. Liu, Z. Wang, and W. Zhan, "Human behavior analysis based on attention mechanism and lstm neural network," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. IEEE, 2019, pp. 346–349. 9, 10

[11] Z. Zhigang, D. Guangxue, L. Huan, Z. Guangbing, W. Nan, and Y. Wenjie, "Human behavior recognition method based on double-branch deep convolution neural network," in *2018 Chinese Control And Decision Conference (CCDC)*. IEEE, 2018, pp. 5520–5524. 9, 10

[12] S. Wang, J. Z. Gao, H. Lin, M. Shitole, L. Reza, and S. Zhou, "Dynamic human behavior pattern detection and classification," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2019, pp. 159–166. 9, 10

[13] Z. Shen, J. Cheng, X. Hu, and Q. Dong, "Emotion recognition based on multi-view body gestures," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3317–3321. 9, 11

[14] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8. 9, 11

[15] M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, and J. W. Brehm, "Attitude organization and change: An analysis of consistency among attitude components," *Yale Univer. Press*, 1960. 10, 40

[16] D. Katz and E. Stotland, "A preliminary statement to a theory of attitude structure and change," *Psychology: A study of a science*, vol. 3, no. 423-475, 1959. 10

[17] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013. 11

[18] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 509–30 555, 2020. 11

[19] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019. 11

[20] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230*, 2018. 11

[21] C. Hessler, M. Abouelenien, and M. Burzo, "A survey on extracting physiological measurements from thermal images," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, 2018, pp. 229–236. 11

[22] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: insights and new developments," *IEEE Transactions on Affective Computing*, 2019. 11

[23] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019. 11

[24] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2017. 11

[25] P. Santhiya and S. Chitrakala, "A survey on emotion recognition from eeg signals: Approaches, techniques & challenges," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*. IEEE, 2019, pp. 1–6. 11

[26] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic emotion recognition based on body movement analysis: a survey," *IEEE computer graphics and applications*, vol. 34, no. 6, pp. 35–45, 2014. 11

[27] F. Ahmed, A. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11 761–11 781, 2019. 11

[28] E. Aronson, *The social animal*. Oxford, England: Viking Press, 1972. 13

[29] R. Baumeister and M. Leary, "The need to belong: desire for interpersonal attachments as a fundamental human motivation." *Psychological bulletin*, vol. 117 3, pp. 497–529, 1995. 13

[30] S. Cohen and T. Wills, "Stress, social support, and the buffering hypothesis." *Psychological bulletin*, vol. 98 2, pp. 310–57, 1985. 13

[31] J. Park, J. Lee, H. Lee, and D. Truex, "Exploring the impact of communication effectiveness on service quality, trust and relationship commitment in it services," *International Journal of Information Management*, vol. 32, no. 5, pp. 459–468, 2012. 13

[32] T. Duncan and S. E. Moriarty, "A communication-based marketing model for managing relationships," *Journal of marketing*, vol. 62, no. 2, pp. 1–13, 1998. 13

[33] A. W. Gouldner, "The norm of reciprocity: A preliminary statement," *American sociological review*, pp. 161–178, 1960. 13

[34] T. Scott-Phillips, *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education, 2014. 13

[35] O. C. Schultheiss and J. C. Brunstein, "Inhibited power motivation and persuasive communication: A lens model analysis," *Journal of personality*, vol. 70, no. 4, pp. 553–582, 2002. 13

[36] B. Spengler, J. Hofer, and H. Busch, "Somebody hit the button! the implicit power motive and the frequency of verbal persuasion behavior in children," *Motivation and Emotion*, vol. 44, no. 5, pp. 695–703, 2020. 13

[37] C. Hill, A. Memon, and P. McGeorge, "The role of confirmation bias in suspect interviews: A systematic evaluation," *Legal and Criminological Psychology*, vol. 13, pp. 357 – 371, 12 2010. 13

[38] K. M. Conniff, J. M. Scarlett, S. Goodman, and L. D. Appel, "Effects of a pet visitation program on the behavior and emotional state of adjudicated female adolescents," *Anthrozoös*, vol. 18, no. 4, pp. 379–395, 2005. 13

[39] E. Friedmann, E. Galik, S. A. Thomas, S. Hall, J. Cheon, N. Han, H. J. Kim, S. McAtee, and N. R. Gee, "Relationship of behavioral interactions during an animal-assisted intervention in assisted living to health-related outcomes," *Anthrozoös*, vol. 32, no. 2, pp. 221–238, 2019. 13

[40] K. Hirsh-Pasek and R. Treiman, "Doggerel: motherese in a new context," *Journal of Child Language*, vol. 9, no. 1, p. 229–237, 1982. 13

[41] G. Melson, "Child development and the human-companion animal bond," *American Behavioral Scientist - AMER BEHAV SCI*, vol. 47, pp. 31–39, 09 2003. 13

[42] V. Lobue, M. E. B. Pickard, K. Sherman, C. Axford, and J. Deloache, "Young children's interest in live animals." *The British journal of developmental psychology*, vol. 31 Pt 1, pp. 57–69, 2013. 13

[43] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007. 14

[44] J. A. Bartz, K. Tchalova, and C. Fenerci, "Reminders of social connection can attenuate anthropomorphism: A replication and extension of epley, akalis, waytz, and cacioppo (2008)," *Psychological Science*, vol. 27, no. 12, pp. 1644–1650, 2016. 14

[45] L. Vygotski, "Thought and language cambridge," 1986. 14

[46] A. Baddeley, "Working memory and language: an overview," *Journal of Communication Disorders*, vol. 36, no. 3, pp. 189 – 208, 2003, aSHA 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021992403000194 14

[47] E. Saeki and S. Saito, "Verbal representation in task order control: An examination with transition and task cues in random task switching," *Memory & Cognition*, vol. 37, no. 7, pp. 1040–1050, 2009. 14

[48] A. J. Kirkham, J. M. Breeze, and P. Mari-Beffa, "The impact of verbal instructions on goal-directed behaviour," *Acta Psychologica*, vol. 139, no. 1, pp. 212 – 219, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001691811001867 14

[49] J. D. Mayer, P. Salovey, P. Salovey, and D. Sluyter, "Emotional development and emotional intelligence: Implications for educators," *What is emotional intelligence*, vol. 5, 1997. 14

[50] J. D. Mayer and G. Geher, "Emotional intelligence and the identification of emotion," *Intelligence*, vol. 22, no. 2, pp. 89–113, 1996. 14

[51] J. Froiland and M. Davison, "Emotional intelligence, listening comprehension, and reading comprehension among diverse adolescents," *Journal of Child and Family Studies*, vol. 29, 05 2020. 14

[52] S. F. Cramer, *Collaboration: A Success Strategy for Special Educators*. ERIC, 1998. 14

[53] T. Gordon and N. Burch, *Teacher effectiveness training: The program proven to help teachers bring out the best in students of all ages*. Three Rivers Press (CA), 2003. 14

[54] D. J. O'Shea, *Families and teachers of individuals with disabilities: Collaborative orientations and responsive practices*. Allyn & Bacon, 2001. 14

[55] C. Jonas-Simpson, G. J. Mitchell, A. Fisher, G. Jones, and J. Linscott, "The experience of being listened to: a qualitative study of older adults in long-term care settings," *Journal of gerontological nursing*, vol. 32, no. 1, pp. 46–53, 2006. 14

[56] L. McKenna, T. Brown, M. Boyle, B. Williams, C. Palermo, and E. Molloy, "Listening and communication styles in nursing students," *Journal of Nursing Education and Practice*, vol. 4, no. 11, p. 50, 2014. 14

[57] T. Sticht, J. James, and P. Pearson, "Handbook of reading research," *Listening and Reading*, pp. 293–317, 1984. 14

[58] R. Atkin, R. Bray, M. Davison, S. Herzberger, L. Humphreys, and U. Selzer, "Cross-lagged panel analysis of sixteen cognitive measures at four grade levels," *Child Development*, vol. 48, no. 3, pp. 944–952, 1977. 14

[59] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 4492–4496. 14, 15, 16, 17, 81, 82, 83, 84, 87, 88, 97, 99, 104, 120, 127, 129, 134

[60] Y. J. Kim, H.-S. Heo, S. Choe, S.-W. Chung, Y. Kwon, B.-J. Lee, Y. Kwon, and J. S. Chung, "Look who's talking: Active speaker detection in the wild," *arXiv preprint arXiv:2108.07640*, 2021. 15, 82, 84, 97, 99, 129

[61] J. S. Chung, "Naver at activitynet challenge 2019–task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019. 15, 83, 99, 129

[62] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, "Multi-task learning for audio-visual active speaker detection," *The ActivityNet Large-Scale Activity Recognition Challenge*, pp. 1–4, 2019. 15, 83, 99, 129

[63] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263. 15, 83, 99

[64] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *2019 IEEE ICASSP*. IEEE, 2019, pp. 3965–3969. 15, 83, 99

[65] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the 2006 IEEE Computer Society Conference on CVPR*, vol. 2. IEEE, 2006, pp. 1735–1742. 15, 83, 99

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017. 15, 83, 99, 102, 129

[67] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proceedings of the ECCV*. Springer, 2020, pp. 208–224. 15, 83, 99, 129

[68] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3884–3892. 15, 83, 99, 129

[69] J. L. Alcázar, F. Caba, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active speakers in context," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2020, pp. 12 465–12 474. 15, 17, 83, 87, 88, 99, 104, 105, 129, 135

[70] O. Köpüklü, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 1193–1203. 15, 17, 83, 87, 88, 99, 104, 105, 129, 130, 134, 135, 138

[71] Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, and X. Chen, "Unicon: Unified context network for robust active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3964–3972. 15, 17, 83, 99, 129, 135

[72] J. L. Alcázar, F. Caba, A. K. Thabet, and B. Ghanem, "Maas: Multi-modal assignation for active speaker detection," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 265–274. 15, 17, 82, 83, 84, 87, 88, 99, 104, 105, 129, 135, 138

[73] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the ICLR*, 2016. 15, 83, 99, 129

[74] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar, "Learning long-term spatial-temporal graphs for active speaker detection," *arXiv preprint arXiv:2207.07783*, 2022. 15, 83, 99, 129, 135

[75] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, "End-to-end active speaker detection," *arXiv preprint arXiv:2203.14250*, 2022. 15, 83, 99, 129, 130

[76] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, "A light weight model for active speaker detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 932–22 941. 15, 17, 83, 87, 88, 95, 99, 104, 105, 106, 107, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138

[77] Y. Jiang, R. Tao, Z. Pan, and H. Li, "Target active speaker detection with audio-visual cues," in *Proc. Interspeech*, 2023. 15, 17, 83, 87, 88, 129, 130, 134, 135

[78] T. Roxo, J. C. Costa, P. R. M. Inácio, and H. Proença, "Asdnb: Merging face with body cues for robust active speaker detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2025, pp. X–X. 15, 17

[79] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, "End-to-end active speaker detection," 2022. [Online]. Available: https://arxiv.org/abs/2203.14250 17, 105, 135

[80] J. Burgoon, D. Buller, and W. G. Woodall, *Nonverbal Communication: The Unspoken Dialogue*. New York: McGraw-Hill, 1996. 17

[81] D. Dew and P. Jensen, *Phonetic processing: The dynamics of speech*. Columbus OH: Merrill, 1977. 17

[82] R. Rubin, "Maternal touch." *Nursing outlook*, vol. 11, pp. 828–9, 1963. 17

[83] A. Montagu, *Touching : the human significance of the skin*, 3rd ed. New York: Perennial Library, 1986. 17

[84] M. Peláez-Nogueras, J. L. Gewirtz, T. Field, M. Cigales, J. Malphurs, S. Clasky, and A. Sanchez, "Infants' preference for touch stimulation in face-to-face interactions," *Journal of Applied Developmental Psychology*, vol. 17, no. 2, pp. 199 – 213, 1996. 17

[85] L. Brossard and T. Décarie, "Comparative reinforcing effect of eight stimulations on the smiling response of infants," *Journal of child psychology and psychiatry, and allied disciplines*, vol. 9, no. 1, p. 51—59, October 1968. 17

[86] M. Hertenstein and J. Campos, "Emotion regulation via maternal touch," *Infancy*, vol. 2, pp. 549 – 566, 10 2001. 17

[87] S. Weiss, P. Wilson, M. Seed, and S. Paul, "Early tactile experience of low birth weight children: Links to later mental health and social adaptation," *Infant and Child Development*, vol. 10, pp. 93 – 115, 09 2001. 17

[88] D. E. Smith, J. A. Gier, and F. N. Willis, "Interpersonal touch and compliance with a marketing request," *Basic and Applied Social Psychology*, vol. 3, no. 1, pp. 35–38, 1982. 17

[89] A. Crusco and C. G. Wetzel, "The midas touch: The effects of interpersonal touch on restaurant tipping," *Personality and Social Psychology Bulletin*, vol. 10, pp. 512 – 517, 1984. 17

[90] F. J. Willis and H. K. Hamm, "The use of interpersonal touch in securing compliance," *Journal of Nonverbal Behavior*, vol. 5, pp. 49–55, 1980. 18

[91] N. M. Henley, *Body politics: Power, sex, and nonverbal communication*. Englewood Cliffs, NJ: Prentice Hal, 1977. 18

[92] R. Heslin and D. Boss, "Nonverbal intimacy in airport arrival and departure," *Personality and Social Psychology Bulletin*, vol. 6, pp. 248 – 252, 1980. 18

[93] J. D. Fisher, M. Rytting, and R. Heslin, "Hands touching hands: Affective and evaluative effects of an interpersonal touch," *Sociometry*, pp. 416–421, 1976. 18

[94] S. Whitcher and J. Fisher, "Multidimensional reaction to therapeutic touch in a hospital setting," *Journal of personality and social psychology*, vol. 37, pp. 87–96, 02 1979. 18, 40

[95] Statista, "E-commerce share of total global retail sales from 2015–2021," 2019, [Online] https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide. 18

[96] S. Skrovan, "Why most shoppers still choose brick-and-mortar stores over e-commerce," 2017, [Online] https://www.retaildive.com/news/why-most-shoppers-still-choose-brick-andmortar-stores-over-e-commerce/436068. 18

[97] J. Peck and S. B. Shu, "The effect of mere touch on perceived ownership," *Journal of consumer Research*, vol. 36, no. 3, pp. 434–447, 2009. 18

[98] J. Peck and T. L. Childers, "To have and to hold: The influence of haptic information on product judgments," *Journal of Marketing*, vol. 67, no. 2, pp. 35–48, 2003. 18

[99] J. Peck and J. Wiggins, "It just feels good: Customers' affective response to touch and its influence on persuasion," *Journal of Marketing*, vol. 70, no. 4, pp. 56–69, 2006. 18

[100] J. Peck and T. L. Childers, "If i touch it i have to have it: Individual and environmental influences on impulse purchasing," *Journal of business research*, vol. 59, no. 6, pp. 765–769, 2006. 18

[101] C. Ringler, N. J. Sirianni, A. Gustafsson, and J. Peck, "Look but don't touch! the impact of active interpersonal haptic blocking on compensatory touch and purchase behavior," *Journal of Retailing*, vol. 95, no. 4, pp. 186 – 203, 2019. 18

[102] A. Beetz, K. Uvnäs-Moberg, H. Julius, and K. Kotrschal, "Psychosocial and psychophysiological effects of human-animal interactions: The possible role of oxytocin," *Frontiers in Psychology*, vol. 3, 2012. 18

[103] H. Brooks, K. Rushton, K. Lovell, P. Bee, L. Walker, L. Grant, and A. Rogers, "The power of support from companion animals for people living with mental health problems: A systematic review and narrative synthesis of the evidence," *BMC Psychiatry*, vol. 18, 12 2018. 18

[104] R. Pruchno, A. Heid, and M. Wilson-Genderson, "Successful aging, social support, and ownership of a companion animal," *Anthrozoös*, vol. 31, pp. 23–39, 01 2018. 18

[105] C. Hutt and R. Bhavnani, "Predictions from play," *Nature*, vol. 237, no. 5351, p. 171—172, 1972. 18

[106] A. Pellegrini, "The role of play in human development," *The Role of Play in Human Development*, pp. 1–288, 01 2010. 18

[107] R. Addison, "Music and play," *British Journal of Music Education*, vol. 8, no. 3, p. 207–217, 1991. 18

[108] Y. Xu, F. Wei, X. Sun, C. Yang, Y. Shen, B. Dai, B. Zhou, and S. Lin, "Cross-model pseudo-labeling for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2959–2968. 19

[109] J. Xiao, L. Jing, L. Zhang, J. He, Q. She, Z. Zhou, A. Yuille, and Y. Li, "Learning from temporal gradient for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3252–3262. 19

[110] X. Wang, S. Zhang, Z. Qing, M. Tang, Z. Zuo, C. Gao, R. Jin, and N. Sang, "Hybrid relation guided set matching for few-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 948–19 957. 19

[111] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 958–19 967. 19

[112] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165. 19

[113] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584. 19

[114] Y. Huang, Y. Sugano, and Y. Sato, "Improving action segmentation via graph-based temporal reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 024–14 034. 19

[115] S.-H. Gao, Q. Han, Z.-Y. Li, P. Peng, L. Wang, and M.-M. Cheng, "Global2local: Efficient structure search for video action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 805–16 814. 19

[116] J. Li and S. Todorovic, "Action shuffle alternating learning for unsupervised action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 628–12 636. 19

[117] L. Ding and C. Xu, "Weakly-supervised action segmentation with iterative soft boundary assignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6508–6516. 19

[118] M. Fayyaz and J. Gall, "Sct: Set constrained temporal transformer for set supervised action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 501–510. 19

[119] H. Ahn and D. Lee, "Refining action segmentation with hierarchical video representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 302–16 310. 19

[120] F. Yi, H. Wen, and T. Jiang, "Asformer: Transformer for action segmentation," *arXiv preprint arXiv:2110.08568*, 2021. 19

[121] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 156–10 165. 20

[122] Z. Yang, J. Qin, and D. Huang, "Acgnet: Action complement graph network for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 3090–3098. 20

[123] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7094–7103. 20

[124] J. Gao, M. Chen, and C. Xu, "Fine-grained temporal contrastive learning for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 999–20 009. 20

[125] B. He, X. Yang, L. Kang, Z. Cheng, X. Zhou, and A. Shrivastava, "Asm-loc: action-aware segment modeling for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 925–13 935. 20

[126] C. Zhang, T. Yang, J. Weng, M. Cao, J. Wang, and Y. Zou, "Unsupervised pretraining for temporal action localization tasks," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2022, pp. 14 031–14 041. 20

[127] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, "Learning to refactor action and co-occurrence features for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 884–13 893. 20

[128] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4405–4413. 20

[129] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211. 20, 21

[130] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 549–14 560. 20

[131] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=AhccnBXSne 20, 21

[132] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=UaXD4Al3mdb 20, 21

[133] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 668–14 678. 20, 21

[134] G. Singh, V. Choutas, S. Saha, F. Yu, and L. Van Gool, "Spatio-temporal action detection under large motion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6009–6018. 20

[135] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, and J. Malik, "On the benefits of 3d pose and tracking for human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 640–649. 20

[136] J. Zhao, Y. Zhang, X. Li, H. Chen, B. Shuai, M. Xu, C. Liu, K. Kundu, Y. Xiong, D. Modolo *et al.*, "Tuber: Tubelet transformer for video action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 598–13 607. 20, 21

[137] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022. 20

[138] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, 2024. 20

[139] M. S. Ryoo and J. Aggarwal, "Ut-interaction dataset, icpr contest on semantic description of human activities (sdha)," in *IEEE International Conference on Pattern Recognition Workshops*, vol. 2, 2010, p. 4. 20

[140] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056. 20, 21, 95, 103, 133

[141] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 21

[142] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015. 21

[143] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, and J. Malik, "On the benefits of 3d pose and tracking for human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 640–649. 21

[144] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 549–14 560. 21

[145] E. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge." *Annual Review of Psychology*, vol. 39, pp. 1–41, 1988. 21, 22

[146] C. Mears and H. Harlow, "Play: early and eternal." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72 5, pp. 1878–82, 1975. 21

[147] W. Cole, S. Robinson, and K. Adolph, "Bouts of steps: The organization of infant exploration," *Developmental Psychobiology*, vol. 58, no. 3, pp. 341–354, Apr. 2016. 21

[148] J. Hoch, S. O'Grady, and K. Adolph, "It's the journey, not the destination: Locomotor exploration in infants," *Developmental Science*, vol. 22, 09 2018. 22

[149] L. Smith, "Cognition as a dynamic system: Principles from embodiment," *Developmental Review*, vol. 25, pp. 278–298, 09 2005. 22

[150] A. Sheya and L. Smith, "Development through sensorimotor coordination," *J.Stewart, O.Gapenne and E.A.Di Paolo (Eds.)*, 11 2010. 22

[151] K. Gagnon, E. Cashdan, J. Stefanucci, and S. Creem-Regehr, "Sex differences in exploration behavior and the relationship to harm avoidance," *Human nature (Hawthorne, N.Y.)*, vol. 27, 12 2015. 22, 40

[152] F. Yang, X. Chang, S. Sakti, Y. Wu, and S. Nakamura, "Remot: A model-agnostic refinement for multiple object tracking," *Image and Vision Computing*, vol. 106, p. 104091, 2021. 22, 25

[153] M. Babaee, Z. Li, and G. Rigoll, "A dual cnn–rnn for multiple people tracking," *Neurocomputing*, vol. 368, pp. 69–83, 2019. 22

[154] N. L. Baisa, "Occlusion-robust online multi-object visual tracking using a gm-phd filter with cnn-based re-identification," *Journal of Visual Communication and Image Representation*, vol. 80, p. 103279, 2021. 22

[155] Q. Liu, D. Chen, Q. Chu, L. Yuan, B. Liu, L. Zhang, and N. Yu, "Online multi-object tracking with unsupervised re-identification learning and occlusion estimation," *Neurocomputing*, vol. 483, pp. 333–347, 2022. 22, 25

[156] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793. 22

[157] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412. 22

[158] S. Sun, N. Akhtar, X. Song, H. Song, A. Mian, and M. Shah, "Simultaneous detection and tracking with motion modelling for multiple object tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 626–643. 23

[159] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 7942–7951. 23

[160] X. Wan, S. Zhou, J. Wang, and R. Meng, "Multiple object tracking by trajectory map regression with temporal priors embedding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1377–1386. 23

[161] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detect and segment: An online multi-object tracker," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 352–12 361. 23

[162] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 145–161. 23

[163] X. Chen, S. M. Iranmanesh, and K.-C. Lien, "Patchtrack: Multiple object tracking using frame patches," *arXiv preprint arXiv:2201.00080*, 2022. 23, 25

[164] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, "Multiple object tracking with correlation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3876–3886. 23, 25

[165] X. Lin, Y.-a. Guo, and J. Wang, "Global correlation network: End-to-end joint multi-object detection and tracking," *arXiv preprint arXiv:2103.12511*, 2021. 23

[166] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-dense similarity learning for multiple object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 164–173. 23, 25

[167] A. Girbau, X. Giró-i Nieto, I. Rius, and F. Marqués, "Multiple object tracking with mixture density networks for trajectory estimation," *arXiv preprint arXiv:2106.10950*, 2021. 23, 25

[168] S. Guo, J. Wang, X. Wang, and D. Tao, "Online multiple object tracking with cross-task synergy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8136–8145. 23

[169] S. Wang, H. Sheng, Y. Zhang, Y. Wu, and Z. Xiong, "A general recurrent tracking framework without real data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 219–13 228. 23, 25

[170] G. Wang, Y. Wang, R. Gu, W. Hu, and J.-N. Hwang, "Split and connect: A universal tracklet booster for multi-object tracking," *IEEE Transactions on Multimedia*, 2022. 23, 25

[171] P. Dai, R. Weng, W. Choi, C. Zhang, Z. He, and W. Ding, "Learning a proposal classifier for multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2443–2452. 23, 25

[172] G. Wang, R. Gu, Z. Liu, W. Hu, M. Song, and J.-N. Hwang, "Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9876–9886. 23

[173] J. Xiang, G. Xu, C. Ma, and J. Hou, "End-to-end learning deep crf models for multi-object tracking deep crf models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 275–288, 2020. 23

[174] Y.-C. Yoon, D. Y. Kim, Y.-M. Song, K. Yoon, and M. Jeon, "Online multiple pedestrians tracking using deep temporal appearance matching association," *Information Sciences*, vol. 561, pp. 326–351, 2021. 23

[175] F. Bastani, S. He, and S. Madden, "Self-supervised multi-object tracking with cross-input consistency," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 695–13 706, 2021. 23, 25

[176] W.-C. Hung, H. Kretzschmar, T.-Y. Lin, Y. Chai, R. Yu, M.-H. Yang, and D. Anguelov, "Soda: Multi-object tracking with soft data association," *arXiv preprint arXiv:2008.07725*, 2020. 23

[177] Q. Liu, Q. Chu, B. Liu, and N. Yu, "Gsm: Graph similarity model for multi-object tracking." in *IJCAI*, 2020, pp. 530–536. 23

[178] J. He, Z. Huang, N. Wang, and Z. Zhang, "Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5299–5309. 23

[179] G. Wang, M. Song, and J.-N. Hwang, "Recent advances in embedding methods for multi-object tracking: a survey," *arXiv preprint arXiv:2205.10766*, 2022. 23

[180] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020. 23

[181] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942 23

[182] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831 23

[183] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009. 23

[184] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 23

[185] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2129–2137. 23

[186] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "Mot20: A benchmark for multi object tracking in crowded scenes," *arXiv:2003.09003[cs]*, Mar. 2020, arXiv: 2003.09003. [Online]. Available: http://arxiv.org/abs/1906.04567 23

[187] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 23

[188] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 17–24. 23

[189] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008. 23

[190] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*. Springer, 2016, pp. 17–35. 23

[191] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *International journal of computer vision*, vol. 129, pp. 548–578, 2021. 23, 24

[192] S. Karthik, A. Prabhu, and V. Gandhi, "Simple unsupervised multi-object tracking," *arXiv preprint arXiv:2006.02609*, 2020. 25

[193] W. Li, Y. Xiong, S. Yang, M. Xu, Y. Wang, and W. Xia, "Semi-tcl: Semi-supervised track contrastive representation learning," *arXiv preprint arXiv:2107.02396*, 2021. 25

[194] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, pp. 3069–3087, 2021. 25

[195] E. Yu, Z. Li, S. Han, and H. Wang, "Relationtrack: Relation-aware multiple object tracking with decoupled representation," *IEEE Transactions on Multimedia*, 2022. 25

[196] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda, "Transcenter: Transformers with dense queries for multiple-object tracking," *arXiv e-prints*, pp. arXiv–2103, 2021. 25

[197] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 708–13 715. 25

[198] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020. 25

[199] P. Tokmakov, J. Li, W. Burgard, and A. Gaidon, "Learning to track with object permanence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 860–10 869. 25

[200] G. Brasó and L. Leal-Taixé, "Learning a neural solver for multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6247–6257. 25

[201] C. Shan, C. Wei, B. Deng, J. Huang, X.-S. Hua, X. Cheng, and K. Liang, "Tracklets predicting based adaptive graph tracking," *arXiv preprint arXiv:2010.09015*, 2020. 25

[202] P.-L. Chen, W. Saleh, and C.-W. Pai, "Texting and walking: a controlled field study of crossing behaviours and inattentional blindness in taiwan," *Behaviour & Information Technology*, vol. 36, pp. 1–11, 10 2016. 25

[203] M. Neider, J. McCarley, J. Crowell, H. Kaczmarski, and A. Kramer, "Pedestrians, vehicles, and cell phones," *Accident; analysis and prevention*, vol. 42, pp. 589–94, 03 2010. 25

[204] W. Horrey and D. Simons, "Examining cognitive interference and adaptive safety behaviours in tactical vehicle control," *Ergonomics*, vol. 50, pp. 1340–50, 09 2007. 25

[205] L. Payne, B. Ainsworth, and G. Godbey, *Leisure, health, and wellness: Making the connections*. Venture Publishing State College, 2010. 25

[206] L. S. Anderson and L. A. Heyne, "Flourishing through leisure and the upward spiral theory of lifestyle change," *Therapeutic Recreation Journal*, vol. 50, no. 2, p. 118, 2016. 25

[207] C. Carruthers and C. Hood, "The power of the positive: Leisure and wellbeing," *Therapeutic recreation journal*, vol. 38, pp. 225–245, 01 2004. 25

[208] D. Kleiber, S. Hutchinson, and R. Williams, "Leisure as a resource in transcending negative life events: Self-protection, self-restoration, and personal transformation," *Leisure Sciences - LEISURE SCI*, vol. 24, pp. 219–235, 04 2002. 25

[209] Y. Iwasaki, C. Coyle, and J. Shank, "Leisure as a context for active living, recovery, health and life quality for persons with mental illness in a global context," *Health promotion international*, vol. 25, pp. 483–94, 12 2010. 26

[210] Y. Iwasaki, K. Mackay, and J. Mactavish, "Gender-based analyses of coping with stress among professional managers: Leisure coping and non-leisure coping," *Journal of Leisure Research*, vol. 37, pp. 1–28, 03 2005. 26

[211] S. Escobar-Chaves, S. Tortolero, C. Markham, S. Kelder, and A. Kapadia, "Violent behavior among urban youth attending alternative schools," *The Journal of school health*, vol. 72, pp. 357–62, 12 2002. 26

[212] L. N. Wallace, "Illicit juvenile weapon possession: The role of serious sanctioning in future behavior," *The Social Science Journal*, vol. 54, no. 3, pp. 319 – 328, 2017. 26

[213] A. Stickley, A. Koyanagi, R. Koposov, M. Blatnỳ, M. Hrdlička, M. Schwab-Stone, and V. Ruchkin, "Correlates of weapon carrying in school among adolescents in three countries," *American journal of health behavior*, vol. 39, no. 1, pp. 99–108, 2015. 26

[214] A. J. Lizotte, M. D. Krohn, J. C. Howell, K. Tobin, and G. J. Howard, "Factors influencing gun carrying among young urban males over the adolescent-young adult life course," *Criminology*, vol. 38, no. 3, pp. 811–834, 2000. 26

[215] U. Gatti, S. Haymoz, and H. M. Schadee, "Deviant youth groups in 30 countries: Results from the second international self-report delinquency study," *International Criminal Justice Review*, vol. 21, no. 3, pp. 208–224, 2011. 26

[216] M. A. Lane, S. D. Cunningham, and J. M. Ellen, "The intention of adolescents to carry a knife or a gun: a study of low-income african-american adolescents," *Journal of adolescent Health*, vol. 34, no. 1, pp. 72–78, 2004. 26

[217] A. Bolfing and A. Schwaninger, "Selection and pre-employment assessment in aviation security x-ray screening," *Proceedings ot the 43rd IEEE International Carnahan Conference on Security Technology*, vol. 43, 10 2009. 26

[218] J. McCarley, A. Kramer, C. Wickens, E. Vidoni, and W. Boot, "Visual skills in airport-security screening," *Psychological science*, vol. 15, pp. 302–6, 06 2004. 26

[219] T. Halbherr, A. Schwaninger, G. Budgell, and A. Wales, "Airport security screener competency: A cross-sectional and longitudinal analysis," *International Journal of Aviation Psychology*, vol. 23, pp. 113–129, 04 2013. 26

[220] S. Robers, *Indicators of school crime and safety: 2010*. DIANE Publishing, 2015. 26

[221] A. Rowhani-Rahbar, D. Azrael, V. Lyons, J. Simonetti, and M. Miller, "Loaded handgun carrying among us adults, 2015," *American journal of public health*, vol. 107, pp. e1–e7, 10 2017. 26

[222] R. Harper, G. Harper, and J. Stockdale, "The role of sentencing of women in drug trafficking crime," *Legal and Criminological Psychology*, vol. 7, pp. 101 – 114, 12 2010. 26

[223] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8359–8367. 27

[224] C. Gao, Y. Zou, and J.-B. Huang, "ican: Instance-centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018. 27

[225] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Visual-semantic graph attention networks for human-object interaction detection," in *2021 IEEE international conference on robotics and biomimetics (ROBIO)*. IEEE, 2021, pp. 1441–1447. 27

[226] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 319–13 327. 27, 29

[227] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594. 27

[228] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human–object interaction detection," *International Journal of Computer Vision*, vol. 129, pp. 1910–1929, 2021. 27

[229] Y. Liu, J. Yuan, and C. W. Chen, "Consnet: Learning consistency graph for zero-shot human-object interaction detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4235–4243. 27

[230] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu, "Pastanet: Toward human activity knowledge engine," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 382–391. 27

[231] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 617–13 626. 27

[232] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 482–490. 27, 29

[233] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4116–4125. 27

[234] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 498–514. 27

[235] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1291–1299. 27, 29

[236] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 74–83. 27

[237] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei *et al.*, "End-to-end human object interaction detection with hoi transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 825–11 834. 27

[238] S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 652–11 661. 27

[239] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9469–9478. 27

[240] L. V. Pato, R. Negrinho, and P. M. Aguiar, "Seeing without looking: Contextual rescoring of object detections for ap maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 610–14 618. 27

[241] E. Luna, J. C. San Miguel, D. Ortego, and J. M. Martínez, "Abandoned object detection in video-surveillance: survey and comparison," *Sensors*, vol. 18, no. 12, p. 4290, 2018. 27

[242] J. Iqbal, M. A. Munir, A. Mahmood, A. R. Ali, and M. Ali, "Orientation aware object detection with application to firearms," *arXiv preprint arXiv:1904.10032*, 2019. 27

[243] J. L. S. González, C. Zaccaro, J. A. Álvarez-García, L. M. S. Morillo, and F. S. Caparrini, "Real-time gun detection in cctv: An open problem," *Neural networks*, vol. 132, pp. 297–308, 2020. 27

[244] M. Antoun and D. Asmar, "Human object interaction detection: Design and survey," *Image and Vision Computing*, p. 104617, 2022. 27, 28

[245] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015. 27

[246] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018, pp. 381–389. 27

[247] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755. 28, 34, 38, 39

[248] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1017–1025. 28

[249] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 584–600. 29

[250] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 696–712. 29

[251] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, 2020. 29

[252] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 234–13 243. 29

[253] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, "Mining the benefits of two-stage and one-stage hoi detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 209–17 220, 2021. 29

[254] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 548–19 557. 29

[255] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 104–20 112. 29

[256] J. Chen and K. Yanai, "Qahoi: query-based anchors for human-object interaction detection," *arXiv preprint arXiv:2112.08647*, 2021. 29

[257] I. Hyman, S. Boss, B. WISE, K. McKENZIE, and J. CAGGIANO, "Did you see the unicycling clown? inattentional blindness while walking and talking on a cell phone," *Applied Cognitive Psychology*, vol. 24, pp. 597–607, 07 2009. 29

[258] V. Horner and A. Whiten, "Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens)," *Animal cognition*, vol. 8, pp. 164–81, 08 2005. 29

[259] C. Legare and M. Nielsen, "Imitation and innovation: The dual engines of cultural learning," *Trends in Cognitive Sciences*, vol. 19, 11 2015. 29

[260] M. Wilks, J. Kirby, and M. Nielsen, "Children imitate antisocial in-group members," *Developmental Science*, vol. 21, p. e12675, 04 2018. 29

[261] W. Mischel, Y. Shoda, and M. Rodriguez, "Delay of gratification in children," *Science*, vol. 244, no. 4907, pp. 933–938, 1989. 29

[262] S. Doebel and Y. Munakata, "Group influences on engaging self-control: Children delay gratification and value it more when their in-group delays and their out-group doesn't," *Psychological Science*, vol. 29, p. 095679761774736, 04 2018. 29

[263] Y. Iwasaki, "Hierarchical dimensions of leisure stress coping," *Journal of Leisure Research*, vol. 22, pp. 163–181, 01 2000. 29

[264] M. Sherif, "A study of some social factors in perception," *Archives of Psychology*, vol. 187, no. 60, 1935. 29

[265] S. E. Asch, "Effects of group pressure upon the modification and distortion of judgments," in *Groups, leadership and men; research in human relations*, H. Guetzkow, Ed. Oxford: Carnegie Press, 1951, pp. 177–190. 29

[266] L. Hawkley and J. Cacioppo, "Loneliness matters: A theoretical and empirical review of consequences and mechanisms," *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, vol. 40, pp. 218–27, 10 2010. 29

[267] T. Ford, H. Lee, and M. Jeon, "The emotional eating and negative food relationship experiences of obese and overweight adults," *Social work in health care*, vol. 56, pp. 1–17, 04 2017. 29

[268] L. Hawkley, R. Thisted, and J. Cacioppo, "Loneliness predicts reduced physical activity: Cross-sectional & longitudinal analyses," *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, vol. 28, pp. 354–63, 06 2009. 29

[269] M. Lim, D. Penn, N. Thomas, and J. Gleeson, "Is loneliness a feasible treatment target in psychosis?" *Social Psychiatry and Psychiatric Epidemiology*, vol. 55, 05 2019. 29

[270] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654. 30

[271] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6197–6206. 30

[272] B. Huang, J.-Q. Zheng, S. Giannarou, and D. S. Elson, "H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4460–4467. 30

[273] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2189–2199. 30

[274] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo," *arXiv preprint arXiv:2209.10248*, 2022. 30

[275] Y. Nam, M. Mostafavi, K.-J. Yoon, and J. Choi, "Stereo depth from events cameras: Concentrate and focus on the future," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6114–6123. 30

[276] S. N. Uddin, S. H. Ahmed, and Y. J. Jung, "Unsupervised deep event stereo for depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7489–7504, 2022. 30

[277] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical cutdepth," *arXiv preprint arXiv:2201.07436*, 2022. 30

[278] Y. Zhang, M. Gong, J. Li, M. Zhang, F. Jiang, and H. Zhao, "Self-supervised monocular depth estimation with multiscale perception," *IEEE transactions on image processing*, vol. 31, pp. 3251–3266, 2022. 30

[279] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019. 30, 32

[280] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018. 30, 32

[281] T.-W. Hui, "Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1675–1684. 30

[282] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 582–600. 30

[283] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5684–5693. 30, 32

[284] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494. 30

[285] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, 2021. 30

[286] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 1738–1764, 2020. 30

[287] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive, 1990, vol. 7. 30

[288] M. Swofford, J. C. Peruzzi, N. Tsoi, S. Thompson, R. Martín-Martín, S. Savarese, and M. Vázquez, "Improving social awareness through dante: A deep affinity network for clustering conversational interactants," *arXiv preprint arXiv:1907.12910*, 2019. 30

[289] H. Hung and B. Kröse, "Detecting f-formations as dominant sets," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 231–238. 30

[290] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015. 30

[291] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *ECCV*, vol. 7576, pp. 746–760, 2012. 31

[292] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. 31

[293] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*. IEEE, 2011, pp. 3457–3464. 31

[294] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011. 32

[295] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 179–12 188. 32

[296] M. Piacentini and G. Mailer, "Symbolic consumption in teenagers' clothing choices," *Journal of Consumer Behaviour: An International Research Review*, vol. 3, no. 3, pp. 251–262, 2004. 32

[297] M. Tiggemann and C. Lacey, "Shopping for clothes: Body satisfaction, appearance investment, and functions of clothing among female shoppers," *Body Image*, vol. 6, no. 4, pp. 285–291, 2009. 32

[298] T. S. Barton, "The body, in theory: Histories of cultural materialism," 1994. 32

[299] M. E. Mishkind, J. Rodin, L. R. Silberstein, and R. H. Striegel-Moore, "The embodiment of masculinity: Cultural, psychological, and behavioral dimensions," *American Behavioral Scientist*, vol. 29, no. 5, pp. 545–562, 1986. 32

[300] L. Veneroni, C. A. Clerici, T. Proserpio, C. Magni, G. Sironi, S. Chiaravalli, L. Roncari, M. Casanova, L. Gandola, M. Massimino *et al.*, "Creating beauty: the experience of a fashion collection prepared by adolescent patients at a pediatric oncology unit," *Tumori Journal*, vol. 101, no. 6, pp. 626–630, 2015. 32

[301] H. B. Baker, "The psychology of clothing as a treatment aid." *Mental hygiene*, vol. 39, no. 1, pp. 94–98, 1955. 32

[302] A. Furnham, P. S. Chan, and E. Wilson, "What to wear? the influence of attire on the perceived professionalism of dentists and lawyers," *Journal of applied social psychology*, vol. 43, no. 9, pp. 1838–1850, 2013. 32

[303] R. A. Gurung, L. Kempen, K. Klemm, R. Senn, and R. Wysocki, "Dressed to present: Ratings of classroom presentations vary with attire," *Teaching of Psychology*, vol. 41, no. 4, pp. 349–353, 2014. 32

[304] S. Murnen and L. Smolak, "I'd rather be a famous fashion model than a famous scientist: The rewards and costs of internalizing sexualization," *The sexualization of girls and girlhood: Causes, consequences, and resistance*, pp. 235–253, 2013. 32

[305] B. Hannover and U. Kühnen, ""the clothing makes the self" via knowledge activation 1," *Journal of Applied Social Psychology*, vol. 32, no. 12, pp. 2513–2525, 2002. 32

[306] N. Howlett, K. J. Pine, N. Cahill, İ. Orakçıoğlu, and B. C. Fletcher, "Unbuttoned: The interaction between provocativeness of female work attire and occupational status," *Sex Roles*, vol. 72, no. 3-4, pp. 105–116, 2015. 32

[307] P. Glick, S. Larsen, C. Johnson, and H. Branstiter, "Evaluations of sexy women in low-and high-status jobs," *Psychology of women quarterly*, vol. 29, no. 4, pp. 389–395, 2005. 32

[308] Y.-H. Kwon and J. Johnson-Hillery, "College students' perceptions of occupational attributes based on formality of business attire," *Perceptual and Motor Skills*, vol. 87, no. 3, pp. 987–994, 1998. 32

[309] H. Adam and A. D. Galinsky, "Enclothed cognition," *Journal of experimental social psychology*, vol. 48, no. 4, pp. 918–925, 2012. 32

[310] B. L. Fredrickson, T.-A. Roberts, S. M. Noll, D. M. Quinn, and J. M. Twenge, "That swimsuit becomes you: sex differences in self-objectification, restrained eating, and math performance." *Journal of personality and social psychology*, vol. 75, no. 1, p. 269, 1998. 32

[311] M. G. Frank and T. Gilovich, "The dark side of self-and social perception: Black uniforms and aggression in professional sports." *Journal of personality and social psychology*, vol. 54, no. 1, p. 74, 1988. 32

[312] P. G. Zimbardo, "The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos." in *Nebraska symposium on motivation*. University of Nebraska press, 1969. 32

[313] M. G. Haselton, M. Mortezaie, E. G. Pillsworth, A. Bleske-Rechek, and D. A. Frederick, "Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress," *Hormones and behavior*, vol. 51, no. 1, pp. 40–45, 2007. 32

[314] A. J. Elliot and D. Niesta, "Romantic red: Red enhances men's attraction to women." *Journal of personality and social psychology*, vol. 95, no. 5, p. 1150, 2008. 33

[315] L. S. Hone and M. E. McCullough, "Are women more likely to wear red and pink at peak fertility? what about on cold days? conceptual, close, and extended replications with novel clothing colour measures," *British Journal of Social Psychology*, p. e12371, 2020. 33

[316] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 229–245. 33

[317] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 172–180. 33

[318] Y. Hu, L. Xiao, Y. Wang, and Y. Jin, "Fashion pose machine for fashion landmark detection," in *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836. SPIE, 2018, pp. 186–190. 33

[319] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1913–1921. 33

[320] S. Qian, D. Lian, B. Zhao, T. Liu, B. Zhu, H. Li, and S. Gao, "Kgdet: Keypoint-guided fashion detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2449–2457. 33

[321] S. Lee, S. Oh, C. Jung, and C. Kim, "A global-local embedding module for fashion landmark detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0. 33

[322] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin, "Layout-graph reasoning for fashion landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2937–2945. 33

[323] Y. Li, S. Tang, Y. Ye, and J. Ma, "Spatial-aware non-local attention for fashion landmark detection," in *2019 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2019, pp. 820–825. 33

[324] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612. 33

[325] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–41, 2021. 33, 34

[326] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," *arXiv preprint arXiv:1703.01386*, 2017. 33

[327] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 125–143. 33

[328] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing, "Interpretable structure-evolving lstm," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1010–1019. 33

[329] J. Ye, Z. Feng, Y. Jing, and M. Song, "Finer-net: Cascaded human parsing with hierarchical granularity," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6. 33

[330] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *2012 IEEE Conference on Computer vision and pattern recognition*. IEEE, 2012, pp. 3570–3577. 33

[331] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016. 33

[332] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3182–3189. 33

[333] S. Liu, Z. Song, M. Wang, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1335–1336. 33

[334] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 2013, pp. 105–112. 33

[335] Z. Li, Y. Li, W. Tian, Y. Pang, and Y. Liu, "Cross-scenario clothing retrieval and fine-grained style recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2912–2917. 33

[336] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070. 33

[337] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y.-G. Jiang, "Matching user photos to online products with robust deep features," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, 2016, pp. 7–14. 33

[338] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, "Fashion retrieval via graph reasoning networks on a similarity pyramid," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3066–3075. 33

[339] K. Laenen, S. Zoghbi, and M.-F. Moens, "Web search of fashion items with multimodal querying," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 342–350. 33

[340] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in neural information processing systems*, vol. 27, 2014. 34

[341] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750. 34

[342] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 850–859. 34

[343] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019. 34

[344] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023. 34

[345] W. Gong and L. Khalid, "Aesthetics, personalization and recommendation: A survey on deep learning in fashion," *arXiv preprint arXiv:2101.08301*, 2021. 34

[346] Y. Deldjoo, F. Nazary, A. Ramisa, J. Mcauley, G. Pellegrini, A. Bellogin, and T. Di Noia, "A review of modern fashion recommender systems," *arXiv preprint arXiv:2202.02757*, 2022. 34

[347] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104. 34

[348] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5337–5345. 34, 35

[349] A. Sidnev, A. Krapivin, A. Trushkov, E. Krasikova, M. Kazakov, and M. Viryasov, "Deepmark++: Real-time clothing detection at the edge," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 2980–2988. 35

[350] H. J. Kim, D. H. Lee, A. Niaz, C. Y. Kim, A. A. Memon, and K. N. Choi, "Multiple-clothing detection and fashion landmark estimation using a single-stage detector," *IEEE Access*, vol. 9, pp. 11 694–11 704, 2021. 35

[351] A. Sidnev, A. Trushkov, M. Kazakov, I. Korolev, and V. Sorokin, "Deepmark: One-shot clothing detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 35

[352] D. Morris, *Manwatching : a field guide to human behaviour*. New York: Harry N. Abrams, Inc, 1977. 35

[353] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, 1969. 35

[354] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *Journal of Business and Psychology*, vol. 24, pp. 179–192, 06 2009. 35, 36

[355] R. Gifford, C. Ng, and M. Wilkinson, "Nonverbal cues in the employment interview. links between applicant qualities and interviewer judgments," *Journal of Applied Psychology*, vol. 70, pp. 729–736, 11 1985. 35, 36

[356] M. Helweg-Larsen, S. J. Cunningham, A. Carrico, and A. M. Pergram, "To nod or not to nod: An observational study of nonverbal communication and status in female and male college students," *Psychology of Women Quarterly*, vol. 28, no. 4, pp. 358–361, 2004. 35

[357] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinician—patient interaction," *Applied and preventive psychology*, vol. 4, no. 1, pp. 21–37, 1995. 35, 36

[358] N. Ambady, J. Koo, R. Rosenthal, and C. Winograd, "Physical therapists' nonverbal communication predicts geriatric patients' health outcomes," *Psychology and aging*, vol. 17, pp. 443–52, 10 2002. 35

[359] Z. Witkower and J. Tracy, "A facial-action imposter: How head tilt influences perceptions of dominance from a neutral face," *Psychological Science*, vol. 30, p. 095679761983876, 04 2019. 36

[360] C. F. Keating and D. L. Bai, "Children's attributions of social dominance from facial cues," *Child Development*, vol. 57, no. 5, pp. 1269–1276, 1986. 36

[361] H. Toscano, T. Schubert, and S. Giessner, "Eye gaze and head posture jointly influence judgments of dominance, physical strength, and anger," *Journal of Nonverbal Behavior*, 09 2017. 36

[362] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," *Nature neuroscience*, vol. 6, no. 12, pp. 1317–1322, 2003. 36

[363] S. Kelly, D. Barr, R. B. Church, and K. Lynch, "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory," *Journal of Memory and Language*, vol. 40, pp. 577–592, 1999. 36

[364] J. Iverson and S. Goldin-Meadow, "What's communication got to do with it?: Gesture in children blind from birth," *Developmental psychology*, vol. 33, pp. 453–67, 06 1997. 36

[365] M. Alibali and S. Goldin-Meadow, "Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind," *Cognitive Psychology*, vol. 25, pp. 468–523, 1993. 36

[366] S. Goldin-Meadow, D. Wein, and C. Chang, "Assessing knowledge through gesture: Using children's hands to read their minds," *Cognition and Instruction*, vol. 9, pp. 201–219, 1992. 36

[367] V. Southgate, C. Maanen, and G. Csibra, "Infant pointing: Communication to cooperate or communication to learn?" *Child development*, vol. 78, pp. 735–40, 05 2007. 36

[368] M. Tomasello, M. Carpenter, and U. Liszkowski, "A new look at infant pointing," *Child development*, vol. 78, pp. 705–22, 05 2007. 36

[369] S. Sporer and B. Schwandt, "Paraverbal indicators of deception: A meta-analytic synthesis," *Applied Cognitive Psychology*, vol. 20, pp. 421 – 446, 05 2006. 36

[370] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969. 36

[371] S. M. Schneider and B. L. Kintz, "The effect of lying upon foot and leg movement," *Bulletin of the Psychonomic Societ*, vol. 10, no. 6, pp. 451–453, 1977. 36

[372] E. T. Hall, *The Silent Language*. Doubleday, 1959, vol. 948. 36, 37

[373] A. Mehrabian, "Significance of posture and position in the communication of attitude and status relationships." *Psychological Bulletin*, vol. 71, no. 5, p. 359, 1969. 36

[374] C. Albas, "Proxemic behavior: A study of extrusion," *The Journal of Social Psychology*, vol. 131, no. 5, pp. 697–702, 1991. 37

[375] J. C. Peery and P. M. Crane, "Personal space regulation: Approach-withdrawal-approach proxemic behavior during adult-preschooler interaction at close range," *The Journal of Psychology*, vol. 106, no. 1, pp. 63–75, 1980. 37

[376] E. T. Hall, *The Hidden Dimension*. Doubleday, 1966. 37

[377] C. Albas, "Proxemic behavior: A study of extrusion," *The Journal of Social Psychology*, vol. 131, no. 5, pp. 697–702, 1991. 37

[378] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1944–1953. 37

[379] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "Tfpose: Direct human pose estimation with transformers," *arXiv preprint arXiv:2103.15320*, 2021. 37, 39

[380] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, Z. Wang, and A. v. den Hengel, "Poseur: Direct human pose regression with transformers," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. Springer, 2022, pp. 72–88. 37

[381] B. Artacho and A. Savakis, "Unipose: Unified human pose estimation in single images and videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7035–7044. 37

[382] Y. Zhang, Y. Wang, O. Camps, and M. Sznaier, "Key frame proposal network for efficient pose estimation in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*. Springer, 2020, pp. 609–625. 37

[383] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1107–1116. 37

[384] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 713–728. 37

[385] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 455–472. 37, 39

[386] C. Jiang, K. Huang, S. Zhang, X. Wang, and J. Xiao, "Pay attention selectively and comprehensively: Pyramid gating network for human pose estimation without pretraining," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2364–2371. 37

[387] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 525–534. 37

[388] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 37

[389] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-pcnn: Two stage human pose estimation with graph pose refinement," in *Computer Vision – ECCV 2020:*

16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI, 2020, p. 492–508. 37

[390] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5674–5682. 38

[391] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 2020, pp. 488–504. 38

[392] U. Rafi, A. Doering, B. Leibe, and J. Gall, "Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 36–52. 38

[393] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732. 38

[394] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, and L. Lin, "Lstm pose machines," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5207–5215. 38

[395] D. Zhang, G. Guo, D. Huang, and J. Han, "Poseflow: A deep motion representation for understanding human behaviors in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6762–6770. 38

[396] S. Chang, L. Yuan, X. Nie, Z. Huang, Y. Zhou, Y. Chen, J. Feng, and S. Yan, "Towards accurate human pose estimation in videos of crowded scenes," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4630–4634. 38

[397] M. Wang, J. Tighe, and D. Modolo, "Combining detection and tracking for human pose estimation in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 088–11 096. 38

[398] C. Zhou, Z. Ren, and G. Hua, "Temporal keypoint matching and refinement network for pose estimation and tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 680–695. 38

[399] R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain, "Multiview-consistent semi-supervised learning for 3d human pose estimation," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 6907–6916. 38

[400] U. Iqbal, P. Molchanov, and J. Kautz, "Weakly-supervised 3d human pose learning via multi-view images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5243–5252. 38

[401] H. Chen, P. Guo, P. Li, G. H. Lee, and G. Chirikjian, "Multi-person 3d pose estimation in crowded scenes based on multi-view geometry," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 541–557. 38

[402] T. Zhi, C. Lassner, T. Tung, C. Stoll, S. G. Narasimhan, and M. Vo, "Texmesh: Reconstructing detailed human texture and geometry from rgb-d video," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 492–509. 38

[403] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang, "Sequential 3d human pose and shape estimation from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7275–7284. 38

[404] Z. Zhang, C. Wang, W. Qin, and W. Zeng, "Fusing wearable imus with multi-view images for human pose estimation: A geometric approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2200–2209. 38

[405] F. Huang, A. Zeng, M. Liu, Q. Lai, and Q. Xu, "Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 429–438. 38

[406] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 267–281. 38

[407] M. Zhao, Y. Liu, A. Raghu, T. Li, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human mesh recovery using radio signals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 113–10 122. 38

[408] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2020, [Online] https://github.com/facebookresearch/detectron2. 38, 53, 71

[409] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, , and C. Lu., "Alphpose," 2020, [Online] https://github.com/MVIG-SJTU/AlphaPose. 38

[410] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose," 2020, [Online] https://github.com/CMU-Perceptual-Computing-Lab/openpose. 38

[411] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020. 38

[412] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2d human pose estimation: a survey," *Multimedia Systems*, pp. 1–24, 2022. 38

[413] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693. 38

[414] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," 2019. 39

[415] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7093–7102. 39

[416] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 718–734. 39

[417] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 269–286. 39

[418] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395. 39

[419] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," *arXiv preprint arXiv:1901.01760*, 2019. 39

[420] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3517–3526. 39

[421] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Computers & Graphics*, vol. 85, pp. 15–22, 2019. 39

[422] A.-S. Warren, K. Goldsmith, and K. Rimes, "Childhood gender□typed behavior and emotional or peer problems: a prospective birth□cohort study," *Journal of Child Psychology and Psychiatry*, vol. 60, 03 2019. 40

[423] G. Rieger and R. Savin-Williams, "Gender nonconformity, sexual orientation, and psychological well-being," *Archives of sexual behavior*, vol. 41, pp. 611–21, 06 2012. 40

[424] A. Roberts, M. Rosario, N. Slopen, J. Calzo, and S. B. Austin, "Childhood gender nonconformity, bullying victimization, and depressive symptoms across adolescence and early adulthood: An 11-year longitudinal study," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 52, pp. 143–52, 02 2013. 40

[425] P. Rodkin and C. Berger, "Who bullies whom? social status asymmetries by victim ender," *International Journal of Behavioral Development - INT J BEHAV DEV*, vol. 32, pp. 473–485, 11 2008. 40

[426] T. Farmer and P. Rodkin, "Antisocial and prosocial correlates of classroom social positions: The social network centrality perspective," *Social Development*, vol. 5, pp. 174 – 188, 04 2006. 40

[427] I. Bongers, J. Ende, and F. Verhulst, "Developmental trajectories of externalizing behaviors in childhood and adolescence," *Child development*, vol. 75, pp. 1523–37, 09 2004. 40

[428] E. Costello, S. Mustillo, A. Erkanli, G. Keeler, and A. Angold, "Prevalence and development of psychiatric disorders in childhood and adolescence," *Archives of general psychiatry*, vol. 60, pp. 837–44, 08 2003. 40

[429] L. Toussaint and J. Webb, "Gender differences in the relationship between empathy and forgiveness," *The Journal of Social Psychology*, vol. 145, pp. 673 – 685, 2005. 40

[430] Y. Deng, L. Chang, M. Yang, M. Huo, and R. Zhou, "Gender differences in emotional response: Inconsistency between experience and expressivity," *PLOS ONE*, vol. 11, no. 6, pp. 1–12, 06 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0158666 40

[431] F. Pazhoohi, C. Silva, J. Lamas, S. Mouta, J. Santos, and J. Arantes, "The effect of height and shoulder-to-hip ratio on interpersonal space in virtual environment," *Psychological Research*, vol. 83, 01 2018. 40, 41

[432] W. Pickett, W. Craig, Y. Harel, J. Cunningham, K. Simpson, M. Molcho, J. Mazur, S. Dostaler, M. D. Overpeck, C. E. Currie *et al.*, "Cross-national study of fighting and weapon carrying as determinants of adolescent injury," *Pediatrics*, vol. 116, no. 6, pp. e855–e863, 2005. 40

[433] R. Kozakai, S. Tsuzuku, K. Yabe, F. Ando, N. Niino, and H. Shimokata, "Age-related changes in gait velocity and leg extension power in middle-aged and elderly people," *Journal of epidemiology / Japan Epidemiological Association*, vol. 10, pp. S77–81, 05 2000. 40

[434] K. Biedenweg, H. Meischke, A. Bohl, K. Hammerback, B. Williams, P. Poe, and E. Phelan, "Understanding older adults' motivators and barriers to participating in

organized programs supporting exercise behaviors," *The journal of primary prevention*, vol. 35, 11 2013. 40

[435] C. Witcher, N. Holt, W. Young, C. Blanchard, D. Murnaghan, and J. Spence, "Physical activity perceptions and influences among older adults in rural nova scotia," *Canadian Journal on Aging / La Revue canadienne du vieillissement*, vol. -1, pp. 1–15, 01 2016. 40

[436] P. Sparling, B. Howard, D. Dunstan, and N. Owen, "Recommendations for physical activity in older adults," *BMJ (Clinical research ed.)*, vol. 350, p. h100, 01 2015. 40

[437] E.-M. Kessler and U. M. Staudinger, "Affective experience in adulthood and old age: The role of affective arousal and perceived affect regulation." *Psychology and aging*, vol. 24, no. 2, p. 349, 2009. 40

[438] D. Isaacowitz, H. Wadlinger, D. Goren, and H. Wilson, ""selective preference in visual fixation away from negative images in old age? an eye-tracking study": Correction." *Psychology and aging*, vol. 21, pp. 40–8, 04 2006. 40

[439] S. T. Charles, "Strength and vulnerability integration: a model of emotional well-being across adulthood." *Psychological bulletin*, vol. 136, no. 6, p. 1068, 2010. 40

[440] K. Livingstone and D. Isaacowitz, "Situation selection and modification for emotion regulation in younger and older adults," *Social Psychological and Personality Science*, vol. 6, 07 2015. 40

[441] L. Yang and L. Hasher, "Age differences in the automatic accessibility of emotional words from semantic memory," *Cognition and Emotion*, vol. 25, no. 1, pp. 3–9, 2011. 41

[442] W. H. Organization, "Obesity and overweight," 2020, [Online] https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight. 41

[443] R. Dobbs, C. Sawers, F. Thompson, J. Manyika, J. Woetzel, P. Child, S. McKenna, and A. Spatharou, "Overcoming obesity: An initial economic analysis. 2014," *McKinsey & Company: www. mckinsey. com/mgi*, pp. 1–106, 2014. 41

[444] A. Sutin, E. Robinson, M. Daly, and A. Terracciano, "Weight discrimination and unhealthy eating-related behaviors," *Appetite*, vol. 102, pp. 83–89, 2016. 41

[445] R. Puhl, M. Wall, C. Chen, S. B. Austin, M. Eisenberg, and D. Neumark-Sztainer, "Experiences of weight teasing in adolescence and weight-related outcomes in adulthood: A 15-year longitudinal study," *Preventive Medicine*, vol. 100, 04 2017. 41

[446] C. Otte, S. Gold, B. Penninx, C. Pariante, A. Etkin, M. Fava, D. Mohr, and A. Schatzberg, "Major depressive disorder," *Nature reviews. Disease primers*, vol. 2, p. 16065, 09 2016. 41

[447] F. N. Jacka, N. Cherbuin, K. J. Anstey, and P. Butterworth, "Dietary patterns and depressive symptoms over time: examining the relationships with socioeconomic position, health behaviours and cardiovascular risk," *PloS one*, vol. 9, no. 1, p. e87657, 2014. 41

[448] J. Tyrrell, A. Mulugeta, A. R. Wood, A. Zhou, R. N. Beaumont, M. A. Tuke, S. E. Jones, K. S. Ruth, H. Yaghootkar, S. Sharp *et al.*, "Using genetics to understand the causal influence of higher bmi on depression," *International journal of epidemiology*, vol. 48, no. 3, pp. 834–848, 2019. 41

[449] M. Mannan, A. Mamun, S. Doi, and A. Clavarino, "Prospective associations between depression and obesity for adolescent males and females-a systematic review and meta-analysis of longitudinal studies," *PloS one*, vol. 11, no. 6, p. e0157240, 2016. 41

[450] K.-K. Mak and S. H. Tan, "Underweight problems in asian children and adolescents," *European journal of pediatrics*, vol. 171, no. 5, pp. 779–785, 2012. 41

[451] J. D. Russell, M. Mira, B. J. Allen, P. M. Stewart, J. Vizzard, B. Arthur, and P. Beumont, "Protein repletion and treatment in anorexia nervosa," *The American journal of clinical nutrition*, vol. 59, no. 1, pp. 98–102, 1994. 41

[452] M. Kakizaki, S. Kuriyama, Y. Sato, T. Shimazu, K. Matsuda-Ohmori, N. Nakaya, A. Fukao, S. Fukudo, and I. Tsuji, "Personality and body mass index: A cross-sectional analysis from the miyagi cohort study," *Journal of psychosomatic research*, vol. 64, pp. 71–80, 02 2008. 41

[453] E. Sundstrom and I. Altman, "Interpersonal relationships and personal space: Research review and theoretical model," *Human Ecology*, vol. 4, no. 1, pp. 47–67, 1976. 41

[454] L. Ellis, *The high and the mighty among man and beast: How universal is the relationship between height (or body size) and social status?* Praeger, 01 1994, pp. 93–112. 41

[455] C. Batres, D. Re, and D. Perrett, "Influence of perceived height, masculinity, and age on each other and on perceptions of dominance in male faces," *Perception*, vol. 44, 08 2015. 41

[456] G. Stulp, A. P. Buunk, S. Verhulst, and T. Pollet, "Human height is positively related to interpersonal dominance in dyadic interactions," *PLoS ONE*, vol. 10, 2015. 41

[457] J. M. M. PhD, "The social significance of short stature," *Loss, Grief & Care*, vol. 3, no. 3-4, pp. 183–189, 1989. 41

[458] C. Watkins, P. Fraccaro, F. Smith, J. Vukovic, D. Feinberg, L. DeBruine, and B. Jones, "Taller men are less sensitive to cues of dominance in other men," *Behavioral Ecology*, vol. 21, pp. 943–947, 08 2010. 41

[459] H. Betancourt and S. Lopez, "The study of culture, ethnicity, and race in american psychology," *American Psychologist*, vol. 48, pp. 629–637, 06 1993. 41

[460] K. Deater-Deckard, K. Dodge, J. E. Bates, and G. Pettit, "Physical discipline among african american and european american mothers: Links to children's externalizing behaviors." *Developmental Psychology*, vol. 32, pp. 1065–1072, 1996. 41

[461] J. D. McLeod, C. Kruttschnitt, and M. Dornfeld, "Does Parenting Explain the Effects of Structural Conditions on Children's Antisocial Behavior? A Comparison of Blacks and Whites*," *Social Forces*, vol. 73, no. 2, pp. 575–604, 12 1994. 41

[462] K. Nomaguchi and A. N. House, "Racial-ethnic disparities in maternal parenting stress: The role of structural disadvantages and parenting values," *Journal of health and social behavior*, vol. 54, no. 3, pp. 386–404, 2013. 42

[463] E. Wood and M. Elliott, "Opioid addiction stigma: The intersection of race, social class, and gender," *Substance Use & Misuse*, vol. 55, no. 5, pp. 818–827, 2020. 42

[464] L. E. Glaze and D. Kaeble, "Correctional populations in the united states, 2013," 2014, [Online] http://www.antoniocasella.eu/nume/Glaze_Kaeble_dec14.pdf. 42

[465] A. Pinkham, N. Sasson, M. Calkins, J. Richard, P. Hughett, R. Gur, and R. Gur, "The other-race effect in face processing among african american and caucasian individuals with schizophrenia," *The American journal of psychiatry*, vol. 165, pp. 639–45, 06 2008. 42

[466] N. Toosi, L. Babbitt, N. Ambady, and S. Sommers, "Dyadic interracial interactions: A meta-analysis," *Psychological bulletin*, vol. 138, pp. 1–27, 11 2011. 42

[467] J. Sanchez-Burks, F. Lee, I. Choi, R. Nisbett, S. Zhao, and J. Koo, "Conversing across cultures: East-west communication styles in work and nonwork contexts." *Journal of personality and social psychology*, vol. 85 2, pp. 363–72, 2003. 42

[468] C. Prado, D. Mellor, L. K. Byrne, C. Wilson, X. Xu, and H. Liu, "Facial emotion recognition: a cross-cultural comparison of chinese, chinese living in australia, and anglo-australians," *Motivation and emotion*, vol. 38, no. 3, pp. 420–428, 2014. 42

[469] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proceedings of the IEEE ICCV*, 2017, pp. 531–540. 42, 50, 67

[470] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning." in *Proceedings of the IJCAI*, 2018, pp. 3177–3183. 42, 50, 68, 99, 129

[471] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the ICML*. PMLR, 2015, pp. 448–456. 42, 50, 51, 68

[472] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6. 42, 68

[473] J. Zhang, P. Ren, and J. Li, "Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints," *arXiv preprint arXiv:2011.06798*, 2020. 42, 68

[474] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," *arXiv preprint arXiv:1808.09102*, 2018. 42, 68

[475] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 111–115. 42, 50, 55, 68, 74, 75, 99, 129

[476] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019. 42, 50, 51, 55, 68, 74, 75, 99, 129

[477] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu, "Hierarchical feature embedding for attribute recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 055–13 064. 42, 68

[478] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," *arXiv preprint arXiv:2005.11909*, 2020. 42, 50, 55, 68, 74, 75, 99, 129

[479] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 055–12 062. 42, 44

[480] T. Roxo and H. Proença, "Is gender "in-the-wild" inference really a solved problem?" *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 573–582, 2021. 42, 47, 68, 100, 116, 129

[481] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the ECCV*, 2018, pp. 680–697. 42, 44, 50, 68, 99, 129

[482] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzalez, "Age and gender recognition in the wild with deep attention," *Pattern Recognition*, vol. 72, pp. 563–571, 2017. 42, 68

[483] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proceedings of the IEEE ICCV*, 2019, pp. 4997–5006. 42, 44, 50, 51, 55, 68, 74, 75, 99, 129

[484] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019, pp. 729–739. 42, 44, 50, 51, 55, 68, 74, 75, 99, 129

[485] M. Chen, G. Wang, J.-H. Xue, Z. Ding, and L. Sun, "Enhance via decoupling: Improving multi-label classifiers with variational feature augmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 1329–1333. 42, 44

[486] J. Jia, X. Chen, and K. Huang, "Spatial and semantic consistency regularizations for pedestrian attribute recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 962–971. 42, 44

[487] J. Jia, N. Gao, F. He, X. Chen, and K. Huang, "Learning disentangled attribute representations for robust pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 1069–1077. 42, 44

[488] A. Specker, M. Cormier, and J. Beyerer, "Upar: Unified pedestrian attribute recognition and person retrieval," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 981–990. 42, 44

[489] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017. 42, 49

[490] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep ror architecture," *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017. 42, 49, 61, 67

[491] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42. 42, 49, 61, 67

[492] R. Vera-Rodriguez, M. Blazquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, and H. Proença, "Facegenderid: exploiting gender information in dcnns face recognition systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 42

[493] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 1–5. 42

[494] J.-C. Xie and C.-M. Pun, "Chronological age estimation under the guidance of age-related facial attributes," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2500–2511, 2019. 42

[495] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Asian conference on computer vision*. Springer, 2014, pp. 144–158. 42

[496] D.-s. Lee, J.-s. Kim, S. C. Jeong, and S.-k. Kwon, "Human height estimation by color deep learning and depth 3d conversion," *Applied Sciences*, vol. 10, no. 16, p. 5531, 2020. 42

[497] S. Günel, H. Rhodin, and P. Fua, "What face and body shapes can tell us about height," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 1819–1827. 42

[498] F. Yin and S. Zhou, "Accurate estimation of body height from a single depth image via a four-stage developing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8267–8276. 42

[499] D. Bieler, S. Gunel, P. Fua, and H. Rhodin, "Gravity as a reference for estimating a person's height from video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8569–8577. 42

[500] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber, "Face-to-bmi: Using computer vision to infer body mass index on social media," *arXiv preprint arXiv:1703.03156*, 2017. 42

[501] M. Jiang and G. Guo, "Body weight analysis from human body images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2676–2688, 2019. 42

[502] A. Pantanowitz, E. Cohen, P. Gradidge, N. Crowther, V. Aharonson, B. Rosman, and D. M. Rubin, "Estimation of body mass index from photographs using deep convolutional neural networks," *arXiv preprint arXiv:1908.11694*, 2019. 42

[503] B. Hassan, E. Izquierdo, and T. Piatrik, "Soft biometrics: a survey: Benchmark analysis, open challenges and recommendations," *Multimedia Tools and Applications*, pp. 1–44, 2021. 43

[504] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Machine Vision and Applications*, vol. 31, pp. 1–13, 2020. 43

[505] X. Wang, S. Zheng, R. Yang, A. Zheng, Z. Chen, J. Tang, and B. Luo, "Pedestrian attribute recognition: A survey," *Pattern Recognition*, vol. 121, p. 108220, 2022. 43

[506] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792. 43, 51, 54, 67, 74, 99, 129

[507] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE ICCV*, 2017, pp. 350–359. 43, 51, 54, 67, 74, 99, 129

[508] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016. 43, 44, 51, 54, 55, 67, 74, 99, 129

[509] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE transactions on image processing*, vol. 28, no. 4, pp. 1575–1590, 2019. 43

[510] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–8. 43, 49, 65, 67

[511] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74. 43, 49, 65, 67

[512] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15. 43

[513] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in face detection and facial image analysis*. Springer, 2016, pp. 189–248. 43

[514] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 43

[515] M. Jiang, Y. Shang, and G. Guo, "On visual bmi analysis from facial images," *Image and Vision Computing*, vol. 89, pp. 183–196, 2019. 43

[516] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 341–345. 43

[517] J. Zhong, H. Qiao, L. Chen, M. Shang, and Q. Liu, "Improving pedestrian attribute recognition with multi-scale spatial calibration," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8. 44

[518] J. Jia, H. Huang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting," *arXiv preprint arXiv:2107.03576*, 2021. 44

[519] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *International conference on biometric authentication*. Springer, 2004, pp. 731–738. 47, 65

[520] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2015. 47, 65

[521] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of the IEEE/CVF ICCV*, 2017. 48, 52, 66, 69, 71, 87

[522] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," *arXiv preprint arXiv:1812.00324*, 2018. 48, 52, 66, 69, 71, 87

[523] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *Proceedings of the BMVC*, 2018. 48, 52, 66, 69, 71, 87

[524] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953. 48, 53

[525] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017. 49, 67

[526] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012. 49, 67

[527] A. Althnian, N. Aloboud, N. Alkharashi, F. Alduwaish, M. Alrshoud, and H. Kurdi, "Face gender recognition in the wild: An extensive performance comparison of deep-learned, hand-crafted, and fused features with deep and traditional models," *Applied Sciences*, vol. 11, no. 1, p. 89, 2021. 49

[528] J.-H. Lee, Y.-M. Chan, T.-Y. Chen, and C.-S. Chen, "Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 162–165. 49, 61

[529] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738. 49, 65, 67

[530] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017. 49, 67

[531] H. Ge, J. Dong, and L. Zhang, "Face attributes recognition based on one-way inferential correlation between attributes," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 253–265. 49

[532] T. Chong, N. Maudet, K. Harima, and T. Igarashi, "Exploring a makeup support system for transgender passing based on automatic gender recognition," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13. 49

[533] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018. 49, 65, 67

[534] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014. 49, 55, 61, 65

[535] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gait analysis for gender classification in forensics," in *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Springer, 2019, pp. 180–190. 50

[536] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015. 50

[537] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019. 50

[538] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 51, 55, 69, 101

[539] A. Morgand and M. Tamaazousti, "Generic and real-time detection of specular reflections in images," *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, vol. 1, pp. 274–282, 01 2014. 51, 73

[540] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 52, 87

[541] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 52

[542] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017. 53, 100, 116

[543] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020. 53

[544] S. e. a. Lundberg, "Shap (shapley additive explanations)," 2021, [Online] https://github.com/slundberg/shap. Last access: 14/03/2021. 53

[545] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700. 55

[546] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 55, 69

[547] S. C. Hung, J.-H. Lee, T. S. Wan, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 339–343. 61

[548] T. Roxo and H. Proença, "Yinyang-net: Complementing face and body information for wild gender recognition," *IEEE Access*, vol. 10, pp. 28 122–28 132, 2022. 65, 99, 117, 129

[549] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2597–2609, 2017. 67

[550] A. Dantcheva and F. Brémond, "Gender estimation based on smile-dynamics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 3, pp. 719–729, 2016. 67

[551] S. Lapuschkin, A. Binder, K.-R. Muller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1629–1638. 67

[552] D. Yaman, F. Irem Eyiokur, and H. Kemal Ekenel, "Multimodal age and gender classification using ear and profile face images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 67

[553] E. Yaghoubi, P. Alirezazadeh, E. Assunção, J. C. Neves, and H. Proença, "Region-based cnns for pedestrian gender recognition in visual surveillance environments," in *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2019, pp. 1–5. 68

[554] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gender classification on 2d human skeleton," in *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE, 2019, pp. 1–4. 68

[555] Y. Li, H. Xu, M. Bian, and J. Xiao, "Attention based cnn-convlstm for pedestrian attribute recognition," *Sensors*, vol. 20, no. 3, p. 811, 2020. 68

[556] H. Fan, H.-M. Hu, S. Liu, W. Lu, and S. Pu, "Correlation graph convolutional network for pedestrian attribute recognition," *IEEE Transactions on Multimedia*, 2020. 68

[557] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9275–9282. 68

[558] Z. Ji, Z. Hu, E. He, J. Han, and Y. Pang, "Pedestrian attribute recognition based on multiple time steps attention," *Pattern Recognition Letters*, vol. 138, pp. 170–176, 2020. 68

[559] L. Gao, D. Huang, Y. Guo, and Y. Wang, "Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1340–1348. 68

[560] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. 72

[561] Y. Yoo, D. Han, and S. Yun, "Extd: Extremely tiny face detector via iterative filter reuse," *arXiv preprint arXiv:1906.06579*, 2019. 72

[562] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212. 72

[563] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929. 77

[564] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gotcha-i: A multiview human videos dataset," in *Security in Computing and Communications*, S. M. Thampi, G. Martinez Perez, R. Ko, and D. B. Rawat, Eds. Singapore: Springer Singapore, 2020, pp. 213–224. 80

[565] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE TPAMI*, vol. 40, no. 5, pp. 1086–1099, 2017. 81, 84, 127

[566] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," *Proc. Interspeech*, pp. 371–375, 2019. 81, 127

[567] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *Proc. Interspeech*, pp. 299–303, 2020. 81, 127

[568] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Transactions on Multimedia*, vol. 24, pp. 942–954, 2021. 81, 127

[569] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target doa estimation with an audio-visual fusion mechanism," in *2021 IEEE ICASSP*. IEEE, 2021, pp. 4280–4284. 81, 127

[570] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proceedings of the ECCV*. Springer, 2016, pp. 285–301. 82, 84, 87, 99, 104, 128, 134

[571] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, 2021. 82, 84, 99, 129

[572] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Tenth IEEE ICCV*, vol. 2. IEEE, 2005, pp. 1424–1431. 83

[573] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009. 83

[574] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Visual voice activity detection in the wild," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 967–977, 2016. 83

[575] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van hamme, "Who's speaking? audio-supervised classification of active speakers in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 87–90. 83

[576] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I.-L. Moreno, "Personal vad: Speaker-conditioned voice activity detection," in *Proceedings of the Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 433–439. 83

[577] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 235–242. 84

[578] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE ICASSP*, vol. 2. IEEE, 2002, pp. II–2017. 84

[579] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection." in *Proceedings of the Interspeech*, 2017, pp. 1938–1942. 84

[580] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition," in *Proceedings of the Eighth International Conference on LREC*, 2012, pp. 1102–1107. 84

[581] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn—a multimodal lstm for speaker identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016. 84

[582] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy"–automatic naming of characters in tv video." in *BMVC*, vol. 2, 2006, p. 6. 84

[583] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1107–1110. 84

[584] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE CVPR*, 2017, pp. 6447–6456. 84

[585] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103. 84

[586] ——, "Lip reading in profile," *BMVC*, 2017. 84

[587] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018. 84

[588] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proceedings of the Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950 84

[589] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of the Interspeech 2018*, 2018, pp. 1086–1090. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929 84

[590] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Proceedings of the Interspeech 2020*, 2020, pp. 299–303. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2337 84, 99

[591] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE ICIP*. IEEE, 2017, pp. 3645–3649. 87

[592] A. Authors, "Anonymous title," *Anonymous Journal/Conference*, p. Anonymous Pages, Anonymous Year. 87

[593] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, TOsmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming, and T. Truong, "opencv/cvat: v1.1.0," Aug. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4009388 87

[594] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018. 93, 101

[595] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf 93

[596] X. Wang, F. Cheng, G. Bertasius, and D. Crandall, "Loconet: Long-short context network for active speaker detection," *arXiv preprint arXiv:2301.08237*, 2023. 95, 106, 136

[597] T. Roxo, J. C. Costa, P. R. Inácio, and H. Proença, "On exploring audio anomaly in speech," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2023, pp. 1–6. 99, 115, 129

[598] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE ICCV*, 2017, pp. 618–626. 100, 115, 117, 120

[599] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016. 100, 116

[600] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *Advances in neural information processing systems*, vol. 32, 2019. 100, 116, 120

[601] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017. 100, 116

[602] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the ICML*. PMLR, 2017, pp. 3319–3328. 100, 116

[603] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015. 100

[604] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of cnns via contrastive backpropagation," in *Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 119–134. 100, 116

[605] B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," in *Proceedings of the 2019 IEEE/CVF ICCVW*. IEEE, 2019, pp. 4176–4185. 100, 116

[606] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the ICML*. PMLR, 2017, pp. 3145–3153. 100, 116

[607] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE TPAMI*, vol. 41, no. 9, pp. 2131–2145, 2018. 100, 116

[608] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Advances in neural information processing systems*, vol. 30, 2017. 100, 116

[609] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018. 100, 116

[610] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 2950–2958. 100, 116

[611] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE ICCV*, 2017, pp. 3429–3437. 100, 116

[612] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 782–791. 100, 117

[613] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu, "Towards interpretable face recognition," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 9348–9357. 100, 117

[614] M. Winter, W. Bailer, and G. Thallinger, "Demystifying face-recognition with locally interpretable boosted features (libf)," in *2022 10th EUVIP*. IEEE, 2022, pp. 1–6. 100, 117

[615] J. Fu, J. Gao, and C. Xu, "Learning semantic-aware spatial-temporal attention for interpretable action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5213–5224, 2021. 100, 117

[616] K. Doshi and Y. Yilmaz, "Towards interpretable video anomaly detection," in *Proceedings of the IEEE/CVF WACV*, 2023, pp. 2655–2664. 100, 117

[617] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020. 103

[618] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. 103

[619] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on CVPR*, 2009, pp. 248–255. 103

[620] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015. 103

[621] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on CVPR*, 2015, pp. 3128–3137. 103

[622] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Text summarization branches out*, 2004, pp. 74–81. 104

[623] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72. 104

[624] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. 104

[625] N. Hingun, C. Sitawarin, J. Li, and D. Wagner, "Reap: a large-scale realistic adversarial patch benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4640–4651. 108

[626] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, vol. 12, pp. 61 113–61 136, 2024. 108

[627] Q. Wang, J. Wan, and A. B. Chan, "On diversity in image captioning: Metrics and methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1035–1049, 2020. 113

[628] T. Roxo, J. C. Costa, P. R. M. Inácio, and H. Proença, "How to squeeze an explanation out of your model," 2024. [Online]. Available: https://arxiv.org/abs/2412.05134 115

[629] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, pp. 239–249, 2020. 115

[630] E. Ning, C. Wang, H. Zhang, X. Ning, and P. Tiwari, "Occluded person re-identification with deep learning: a survey and perspectives," *Expert Systems with Applications*, p. 122419, 2023. 115

[631] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, vol. 12, pp. 61 113–61 136, 2024. 115

[632] J. C. Costa, T. Roxo, J. B. F. Sequeiros, H. Proença, and P. R. M. Inácio, "Predicting cvss metric via description interpretation," *IEEE Access*, vol. 10, pp. 59 125–59 134, 2022. 115

[633] L. Cascone, C. Pero, and H. Proença, "Visual and textual explainability for a biometric verification system based on piecewise facial attribute analysis," *Image and Vision Computing*, vol. 132, p. 104645, 2023. 115

[634] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019. 115

[635] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf 116

[636] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 116, 120

[637] Z. Chen, C. Wang, Y. Wang, G. Jiang, Y. Shen, Y. Tai, C. Wang, W. Zhang, and L. Cao, "Lctr: On awakening the local continuity of transformer for weakly supervised object localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 410–418. 116

[638] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2886–2895. 116

[639] S. Gupta, S. Lakhotia, A. Rawat, and R. Tallamraju, "Vitol: Vision transformer for weakly supervised object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4101–4110. 116

[640] Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, and D. Zhu, "Attcat: Explaining transformers via attentive class activation tokens," *Advances in neural information processing systems*, vol. 35, pp. 5052–5064, 2022. 116

[641] T. Yuan, X. Li, H. Xiong, H. Cao, and D. Dou, "Explaining information flow inside vision transformers using markov chain," in *eXplainable AI approaches for debugging and diagnosis.*, 2021. 117

[642] B. Pan, R. Panda, Y. Jiang, Z. Wang, R. Feris, and A. Oliva, "Ia-red2: Interpretability-aware redundancy reduction for vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 898–24 911, 2021. 117

[643] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe, "Saliency tubes: Visual explanations for spatio-temporal convolutions," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1830–1834. 117

[644] L. Hiley, A. Preece, Y. Hicks, S. Chakraborty, P. Gurram, and R. Tomsett, "Explaining motion relevance for activity recognition in video deep learning models," *arXiv preprint arXiv:2003.14285*, 2020. 117

[645] C. Roy, M. Nourani, S. Arya, M. Shanbhag, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate, "Explainable activity recognition in videos using deep learning and tractable probabilistic models," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 4, pp. 1–32, 2023. 117

[646] S. A. Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, and S. Sclaroff, "Excitation backprop for rnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1440–1449. 117

[647] Z. Li, W. Wang, Z. Li, Y. Huang, and Y. Sato, "Towards visually explaining video understanding networks with perturbation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1120–1129. 117

[648] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009. 120

[649] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008. 120

[650] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847. 120

[651] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–7. 120

[652] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021. 120

[653] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, "A light weight model for active speaker detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 932–22 941. 120, 121

[654] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 120, 125

[655] T. Roxo, J. C. Costa, P. Inácio, and H. Proença, "Asdnb: Merging face with body cues for robust active speaker detection," 2024. [Online]. Available: https://arxiv.org/abs/2412.08594 127

[656] S.-H. Kang and J.-H. Han, "Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction," *International Journal of Social Robotics*, vol. 15, no. 4, pp. 631–641, 2023. 127

[657] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021. 127

[658] J. Liao, H. Duan, X. Li, H. Xu, Y. Yang, W. Cai, Y. Chen, and L. Chen, "Occlusion detection for automatic video editing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2255–2263. 127

[659] H. Duan, J. Liao, L. Lin, and W. Cai, "Flad: a human-centered video content flaw detection system for meeting recordings," in *Proceedings of the 32nd Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2022, pp. 43–49. 127

[660] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241. 131

[661] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 192–201. 133

[662] G. Datta, T. Etchart, V. Yadav, V. Hedau, P. Natarajan, and S.-F. Chang, "Asd-transformer: Efficient active speaker detection using self and multimodal transformers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4568–4572. 135