Interpretable Face Verification Using Visual Explanations

Bernardo Manuel Marques Claro

Versão Final Após Defesa

Dissertação para obtenção do Grau de Mestre e
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor João Carlos Raposo Neves
Co-orientador: Prof. Doutor Hugo Pedro Martins Carriço Proença

Julho de 2025

Declaração de Integridade

Eu, Bernardo Manuel Marques Claro, que abaixo assino, estudante com o número de inscrição M13745 do Mestrado 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 Integridades 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 24/07/2025

Assinado por: **Bernardo Manuel Marques Claro**
Num. de Identificação: 15552900

Acknowledgments

I would like to express my gratitude to my advisor, Doctor João Neves, and my co-advisor, Doctor Hugo Proença, for their guidance, encouragement, and feedback throughout the course of this work. I am especially thankful for their high standards and demanding expectations, which consistently challenged me to improve and helped grow both academically and personally.

I also extend my appreciation to the members of the SociaLab, whose constructive discussions, collaborative mindset, and genuine willingness to offer support contributed meaningfully to this work. I am especially thankful to Diogo Paulo, Ana Dias, Sara Inácio and João Martins for their support during key stages of the project.

Finally, I would like to thank my family and close friends for their support and patience throughout this journey. Their encouragement has been a source of strength during both the challenges and accomplishments of this work.

Resumo

Uma quantidade significativa de sistemas de autenticação baseia-se em sistemas de Verificação Facial (VF), devido à sua elevada precisão e escalabilidade. Estes sistemas demonstraram desempenhos elevados que já ultrapassam o desempenho da VF humana. No entanto, os resultados são frequentemente binários, não fornecem explicações e, consequentemente, funcionam como caixas pretas, faltando-lhes transparência nos seus processos de tomada de decisão. Esta falta de interpretabilidade suscita sérias preocupações no que diz respeito à confiança e à equidade em contextos sensíveis em que as decisões devem ser justificadas. Além disso, os métodos existentes para explicar os sistemas de VF podem ser melhorados em termos de precisão e clareza das explicações visuais.

Neste trabalho, propomos uma estratégia inovadora para explicação de decisões de verificação facial, baseada em perturbações realistas obtidas através da combinação de máscaras semânticas com um modelo de *inpainting*. Para cada par de faces, o método remove sistematicamente regiões da face específicas e reconstrói essas regiões de forma visualmente plausível. A comparação entre as pontuações de similaridade antes e depois da reconstrução permite quantificar a contribuição de cada região para a decisão da verificação, resultando num mapa de similaridade que indica quais as regiões da face foram mais relevantes para a previsão do modelo.

Propomos duas estratégias complementares: a *Single Inpaint* (S0), que avalia individualmente a influência de cada região semântica, e a *Greedy Inpaint* (S1), que analisa o impacto conjunto de diferentes combinações de regiões perturbadas na pontuação de similaridade.

Os resultados experimentais obtidos demonstram de forma clara a eficácia da abordagem proposta. No plano qualitativo, verifica-se que a utilização conjunta de máscaras semânticas e *inpainting* permite gerar mapas de similaridade com maior precisão, consistência visual, interpretabilidade e fidelidade, que superam os métodos do estado da arte que recorrem a oclusões aleatórias e pouco realistas. Esta superioridade é também evidenciada nos resultados quantitativos: o estudo de remoção realizado confirma que, embora as máscaras semânticas já tragam ganhos relevantes na identificação de regiões faciais importantes, é a sua integração com o inpainting que assegura explicações significativamente mais fiáveis e representativas das decisões do modelo.

Palavras-chave

Verificação Facial, Interpretabilidade, Explicações Visuais,

Resumo alargado

A VF é uma das tarefas mais relevantes e amplamente adotadas no domínio do reconhecimento facial. Devido à sua elevada precisão, escalabilidade e facilidade de integração em sistemas automatizados, tornou-se uma componente essencial em diversos contextos, incluindo segurança, autenticação digital, controlo de acessos e vigilância. Atualmente, os sistemas de VF baseados em redes neuronais profundas demonstram desempenhos excelentes, frequentemente superiores ao reconhecimento facial realizado por humanos em cenários controlados. No entanto, apesar da sua eficácia, estes sistemas funcionam, maioritariamente, como caixas-pretas, fornecendo decisões binárias sem qualquer justificação visual ou semântica compreensível para o utilizador final. Esta falta de interpretabilidade levanta preocupações sérias em contextos sensíveis, onde a transparência, a confiança e a equidade são requisitos fundamentais, como nos domínios jurídico ou forense.

Embora tenham surgido diversas abordagens para explicar os modelos de reconhecimento facial, muitas destas são adaptações de técnicas genéricas de explicabilidade desenvolvidas, originalmente, para tarefas de classificação, como LIME [1], RISE [2] ou Grad-CAM [3]. Estas abordagens, quando adaptadas para tarefas de VF, tendem a gerar mapas de saliência dispersos e difíceis de interpretar. Além disso, também se baseiam em perturbações artificiais, como oclusões binarias aleatórias, que podem comprometer a coerência visual da imagem original e, consequentemente, produzir mapas de explicabilidade pouco confiáveis.

Com o objetivo de colmatar estas limitações, esta dissertação propõe uma abordagem inovadora e agnóstica ao modelo de VF para explicar decisões de verificação facial, baseada em reconstruções visuais de alta qualidade. A metodologia desenvolvida combina máscaras semânticas faciais, extraídas a partir de estimativas de pontos faciais e um modelo de segmentação semântica, com um modelo de *inpainting*. O processo consiste em perturbar sistematicamente diferentes regiões faciais, como os olhos, nariz, boca, cabelo, entre outras, e reconstruí-las de forma a que o resultado seja realista mas diferente do original. Em seguida, mede-se a variação na pontuação de similaridade entre o par de imagens original e o par modificado, utilizando um modelo de reconhecimento facial pré-treinado. Essa diferença permite quantificar o contributo de cada região para a decisão final, ao gerar, por fim, um mapa de similaridade que indica as áreas mais determinantes para a verificação positiva ou negativa da identidade.

Foram exploradas duas estratégias complementares para gerar mapas de explicabilidade: *Single Inpaint (S0)*, que avalia isoladamente o impacto de cada região semântica, o que permite uma análise independente da importância de cada área da face. *Greedy Inpaint (S1)*, que adota uma abordagem iterativa e cumulativa, que seleciona, a cada iteração, a região que mais impacta a pontuação de similaridade, de forma a captar efeitos conjuntos e interações entre regiões.

Os resultados obtidos demonstram, de forma qualitativa, que a abordagem proposta produz mapas de similaridade mais precisos, coerentes e interpretáveis, quando comparada com métodos do estado da arte. Em particular, as regiões destacadas pelo método estão em maior consonância com as diferenças visuais reais entre as imagens e são mais fáceis de interpretar.

Os métodos que foram usados como comparação revelam limitações, seja pela dispersão dos mapas gerados ou pela sua baixa robustez em casos mais desafiantes.
Adicionalmente, realizou-se um estudo de remoção quantitativo, baseado nas métricas *Deletion* e *Insertion*, especificamente desenhadas para tarefas de verificação facial. Os resultados confirmam que a combinação de máscaras semânticas com inpainting proporciona explicações mais fidedignas e exatas, em comparação com uma abordagem em que apenas se utilizem máscaras semânticas binárias.
Apesar dos contributos apresentados, a abordagem proposta apresenta algumas limitações. Nomeadamente, o custo computacional associado ao processo de inpainting é elevado, o que compromete a sua aplicação em tempo real. Para além disso, a qualidade das explicações depende da exatidão das máscaras semânticas utilizadas, o que torna o sistema sensível a erros na segmentação ou na deteção de pontos faciais.
Como trabalho futuro, sugere-se a investigação de modelos de inpainting que conciliem qualidade visual com desempenho computacional mais eficiente, bem como a adoção de modelos de inpainting controláveis, que permitam guiar o processo de reconstrução consoante objetivos específicos. Adicionalmente, seria relevante explorar abordagens que integrem modelos de linguagem-visão.

Abstract

A significant number of authentication systems rely on Face Verification (FV) due to its high accuracy, scalability and ease of integration in real-world applications. These systems often achieve performance levels that surpass human verification capabilities. However, they typically function as black boxes, offering binary outputs without insight into the decision-making process. This lack of interpretability raises serious concerns regarding trust, fairness and transparency, particularly in sensitive contexts where decisions must be explainable. Additionally, current explanation methods for FV still exhibit limitations in terms of precision, clarity and reliability of the visual explanations they generate.

This dissertation presents a novel model-agnostic framework for explaining FV decisions using realistic perturbations guided by semantically segmented face regions. The proposed approach combines semantic face masks with a state-of-the-art face inpainting model to reconstruct masked regions in a visually coherent manner. For each face pair, the system systematically masks individual semantic face regions, inpaints the occluded areas, and compares the similarity scores before and after modification. This process quantifies the contribution of each region to the final verification decision and produces a similarity map that highlights the most influential facial areas.

Two complementary perturbation strategies are introduced: *Single Inpaint (S0)*, which assesses the individual impact of each semantic region, and *Greedy Inpaint (S1)*, which incrementally evaluates combinations of regions to capture joint contributions. Extensive qualitative analysis shows that the proposed method produces more interpretable, precise, and visually coherent explanation maps than existing state-of-the-art techniques, which typically rely on random or unrealistic occlusions. This is further validated through a quantitative ablation study using Deletion and Insertion metrics, which confirms that the integration of semantic guidance with inpainting significantly improves the accuracy, reliability, and faithfulness of the resulting explanations.

Keywords

Face Verification, Interpretability, Visual Explanations,

Contents

List of Figures

List of Tables

Lista de Acrónimos

AI Artificial Intelligence

AUC Area Under Curve

AVG Average Removal / Aggeregation

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

CorrRISE Correlation-based Randomized Input Sampling for Explanation

DDPM Denoising Diffusion Probabilistic Model

D-HMs Dissimilarity Heat Maps

FI Face Identification

FR Face Recognition

FV Face Verification

GAN Generative Adversarial Network

Grad-CAM Gradient-weighted Class Activation Mapping

LIME Local Interpretable Model-agnostic Explanations

RISE Randomized Input Sampling for Explanation of Black-box Models

S-HMs Similarity Heat Maps

SEQ Sequential Removal/ Aggregation

SHAP Shapley Additive Explanations

SOTA state-of-the-art

VF Verificação Facial

VLM Vision Language Model

Chapter 1

Introduction

1.1 Motivation & Scope

Face Verification has become a crucial component of modern authentication systems. It can be used to unlock smartphones or even guarantee secure access to important facilities. Substantial advances in the area of deep learning have significantly improved accuracy and scalability of these systems [11–13], making them indispensable in various areas such as finance, health, and even border security. There is, however, a problem: these systems lack transparency and interpretability in the way they make their decisions [14], functioning mostly as black boxes, in other words, they produce superb results, even surpassing the human eye [15], but without revealing the underlying logic. This opacity can cause problems, since in some sensitive applications, trust and fairness are key components [**?**, 16, 17]. Face verification involves determining whether two face images belong to the same person, that is, a binary classification over image pairs. Yet, current systems typically output only a final decision (match or non-match) without indicating which visual cues influenced that outcome. The need to develop more interpretable facial verification systems arises from society's growing concern and distrust of potential biases and errors in the most sensitive contexts. Without knowledge and understanding of how these systems reach their decisions and outputs, it is difficult to identify and resolve these biases. Furthermore, in high-risk and important scenarios, such as a criminal investigation [18] or a national security matter, a lack of interpretability can hinder the ability to justify or challenge the system's decisions, which could potentially cause some lack of trust in the technology and its creators. This research is therefore motivated by the need to bridge the gap between the high performance that already exists in face verification / recognition systems and their lack of interpretability. By developing strategies that provide visual explanations, this work aims to explain the decision-making process of these systems in a way that allows users to understand which regions of an image pair contribute most significantly to a positive or negative classification. The need for better interpretable solutions also addresses potential emerging regulatory requirements about Artificial Intelligence (AI) transparency in biometric systems.

In summary, the main contributions of this dissertation are as follows. First, it introduces a novel model-agnostic explanation strategy for face verification systems based on realistic perturbations that combine semantic face masks with inpainting techniques. This strategy allows for the generation of reliable similarity maps that highlight the most influential face regions involved in each verification decision, thus addressing the lack of interpretability in existing models. Second, two complementary algorithms are developed for generating explanation maps: the Single Inpaint (S0) strategy, which individually assesses the contribution of each semantic region, and the Greedy Inpaint (S1) strategy, which incrementally evaluates

combinations of regions to capture joint effects. Third, the proposed method is validated through both qualitative and quantitative evaluations, which demonstrate that the combination of semantic masking with inpainting provides more accurate and faithful similarity maps. Finally, this work contributes to the growing field of explainable artificial intelligence in biometric applications by introducing a novel and promising direction for generating explanations in face verification.

1.2 Objectives

The main objectives of this work are:

- **Comparative evaluation of existing interpretability methods:**
  The second objective is to systematically evaluate existing interpretability methods adapted to face verification systems. This evaluation will highlight the strengths and limitations of current techniques, some of which will be transversal to various state-of-the-art methods, consequently providing a comparative analysis to help develop a solution that offers improvements over what currently exists.
- **Develop a new strategy to explain the decision-making process of face verification systems:**
  Based on the knowledge acquired in the first two objectives, the final objective is to propose and develop a new strategy to explain the decision-making process of face verification systems. This strategy should prioritize clarity and use visual explanations to highlight the most influential regions of an image pair in the verification process, where the answer can be positive or negative.

1.3 Document Organization

In order to reflect the work that has been done, this document is structured as follows:

1. The first Chapter – **Introduction** – Background to the work, motivation, summary description and objectives.
2. The second Chapter – **Related Work** – Describes the most important concepts within the scope of this project, state-of-the-art (SOTA) methods, relevant information gathered, as well as some important technologies.
3. The third Chapter – **Proposed Method** – Provides a detailed explanation of the proposed methodology. Contains a description of the core concepts, and a description of the proposed algorithms.
4. The fourth Chapter – **Experiments and Results** – Describes the experimental setup, evaluation metrics, qualitative analysis of the generated explanation maps, comparisons with state-of-the-art methods, and an ablation study that quantifies the contribution of key components.

5. The fifth Chapter – **Conclusions and Future Work** – Provides a summary of the contributions, highlights limitations, and suggests future exploration paths.

Chapter 2

Related Work

2.1 Introduction

In this chapter, we review prior work related to the key components of this dissertation, namely FV, interpretability, and realistic perturbation techniques. The goal is to provide the necessary background to understand the motivations behind the proposed method and to position it within the current state of the art.
We begin by reviewing fundamental concepts and recent advances in Face Verification. Then, we explore the field of Interpretability and Visual Explanations, both from a general perspective (Section 2.3) and in the specific context of face verification systems (Section 2.4). Special attention is given to model-agnostic methods, which are particularly relevant to this work.
Finally, we discuss the use of Face Image Inpainting (Section 2.5) as a tool to produce realistic and semantically coherent perturbations, which serve as an alternative to traditional binary masking strategies. Although inpainting is not traditionally associated with interpretability, the proposed approach in this dissertation demonstrates its potential to generate more faithful and visually plausible explanations by reconstructing facial regions instead of removing them. This final section provides the necessary technical foundation for understanding the reconstruction model adopted in our method.

2.2 Face verification

Face verification is a computer vision task that involves determining whether two images of faces belong to the same person. This task is commonly addressed using machine learning techniques, which involves extracting features from these images, such as the shape and texture of the face, and then using them to compare and verify the similarity between the images.
FV is one of the two main approaches to FR, the broader task of identifying or verifying a person's identity based on face images. The other approach is Face Identification (FI). While FV focuses on determining whether two face images belong to the same person (a 1:1 comparison), FI involves matching a given face image to a specific identity in a database(a 1:N comparison). Both tasks share approaches based on deep learning, particularly in feature learning and the optimization of the embedding space.
The methods discussed in this chapter leverage deep learning approaches that enable robust face feature extraction and learn highly discriminating features to effectively capture similarities and differences between facial images.
Face verification systems typically follow a common pipeline (Figure 2.1 that consists of several steps: First, given two face images as input, the system performs face detection and

alignment. Then, a feature extraction network processes each image to produce feature embeddings - high-dimensional vector representations that capture the distinctive characteristics of each face. These embeddings are designed to map faces into a feature space where images of the same person are close together, while images of different people are far apart.

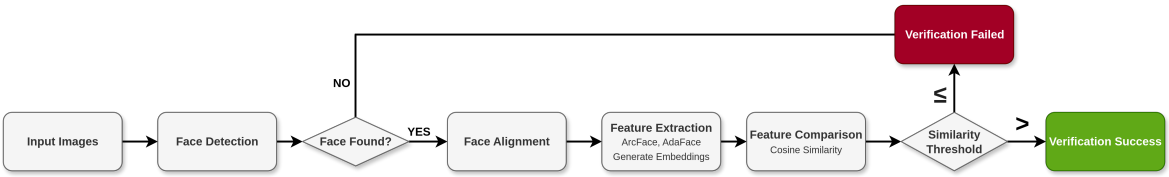


Figure 2.1: **Pipeline of a face verification system:** The process begins with input images, followed by face detection and alignment. If no face is detected, the verification immediately fails. Otherwise, facial embeddings are extracted using a FR model. The extracted features are compared using cosine similarity, and the verification decision is made based on whether the similarity score exceeds a predefined threshold.

The feature extraction network is trained on large datasets of face images using specialized loss functions that help enforce these desired embeddings properties. During inference, when comparing two faces, the system computes the similarity between their feature embeddings, typically using metrics like cosine similarity, cosine distance or Euclidean distance. The score is then compared against a threshold to make the final verification decision.

The following subsections present state-of-the-art face recognition models that have contributed to the learning of robust and discriminative feature embeddings for face verification.

2.2.1 ArcFace: Additive Angular Margin Loss for Deep Face Recognition

One of the most widely used methods in FV tasks is ArcFace [11]. Its approach based on angular margin aimed to directly solve one of the biggest challenges of FV: learning a feature space in which the distance between the embeddings of the faces reliably indicates identity similarity.

This method substantially improved the state of the art and was such a fundamental contribution that, despite being published in 2019, it is still widely used and recognized today.

This method introduces an additive angular margin loss function. This margin is applied in a hyperspherical manifold, where face embeddings are normalized and mapped onto a high-dimensional sphere. The angular margin is added to the angle between an embedding and its corresponding class center during training. This ensures that embeddings of the same identity are pulled closer together, while embeddings of different identities are pushed further apart. By introducing this margin, ArcFace enforces a more stringent separation between classes, leading to highly discriminative features.

While previous methods such as SphereFace [19] and CosFace [20] also exploited angular margins, ArcFace introduces a more geometrically interpretable approach, directly optimizing the geodesic distance between feature vectors, as depicted in Figure 2.2.

The loss function mentioned above can be formally expressed as:

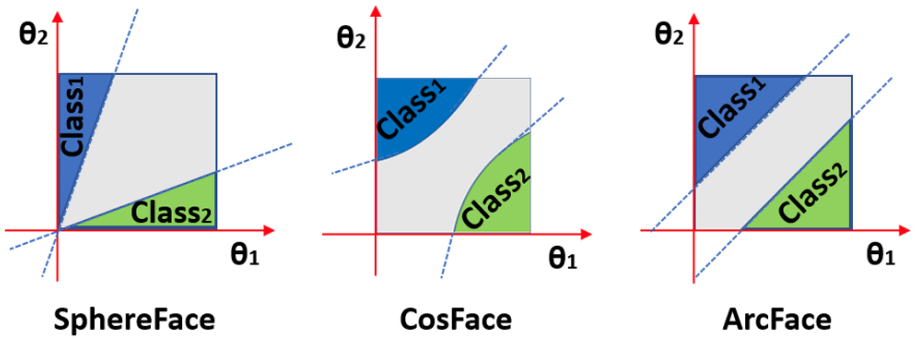


Figure 2.2: **Decision margins of different loss functions under binary classification case.** The dashed line represents the decision boundary, and the grey areas are the decision margins.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{yi}+m))}}{e^{s(\cos(\theta_{yi}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos(\theta_j)}}, \quad (2.1)$$

where θ_{yi} represents the angle between the feature vector and the corresponding weight vector, m is the additive angular margin and s is the scale of features. This ensures that the features that are learned have greater inter-class distances and fewer intra-class variations.
This method implements several important technical components that contribute to its effectiveness: (a) normalization of features that restricts facial features to a fixed radius hypersphere, (b) normalization of the weights in the last fully connected layer, and (c) an additive angular margin penalty that imposes similarity requirements for pairs of positive faces.
The architecture uses a ResNet backbone [21] for feature extraction, followed by the ArcFace loss layer for metric learning. During training, the network learns to project face images into a discriminative feature space where the angular margin between different identities is maximized, maintaining compact intra-class distributions.
Empirically, the ArcFace model has demonstrated superior performance on several benchmark datasets, including LFW, CFP-FP and AgeDB-30, which established state-of-the-art results at the time of publication (2019).

2.2.2 MagFace: A Universal Representation for Face Recognition and Quality Assessment

MagFace [12] introduced a new approach to face recognition and quality assessment by taking advantage of the magnitude of feature embeddings as an indicator of image quality. This methodology makes it possible to improve the accuracy of face recognition and improve the estimation of face image quality at the same time. Unlike previous methods that relied solely on the angular margin, this method integrates optimization that takes magnitude into account directly into the feature learning process.
The main contributions of this method were:

1. **Magnitude as a quality indicator:** This method introduces a novel approach by using the magnitude of feature embeddings as a direct indicator of face image quality.

High-quality images have greater magnitude, and the model uses this property to guide its optimization.

2. **Adaptive margin and regularization:** The method also introduces two auxiliary functions to the loss function, a magnitude sensitive angular margin and a regularizer. This mechanism dynamically adjusts the training loss based on the quality of the image, bringing high quality samples closer to their class centers while pushing low quality or ambiguous samples towards the origin, resulting in more robust embeddings.

2.2.3 AdaFace: Quality Adaptive Margin for Face Recognition

The AdaFace method [13] represented a significant advance in facial recognition, as it addressed the challenge of recognizing faces in low-quality images.
Unlike the previous margin based loss functions, AdaFace introduces a quality-adaptive margin that dynamically adjusts the emphasis placed on the training samples based on the quality of the image.
This approach is particularly relevant for real-world applications, where images of faces often vary significantly in quality due to factors such as lighting, pose and resolution. This margin is calculated as follows:

$$f(\theta_j, m)_{\text{AdaFace}} = \begin{cases} s \cdot \cos(\theta_j + g_{\text{angle}}) - g_{\text{add}}, & \text{if } j = y_i, \\ s \cdot \cos(\theta_j), & \text{if } j \neq y_i, \end{cases} \quad (2.2)$$

where g_{angle} represents the angular adjustment of the margin, and g_{add} denotes the additive margin adjustment. Both are calculated as functions of the standardized feature norm $\|\hat{z}_i\|$, which serves as an indicator of image quality:

$$g_{\text{angle}} = -m \cdot \|\hat{z}_i\|, \quad (2.3)$$

$$g_{\text{add}} = m \cdot \|\hat{z}_i\| + m, \quad (2.4)$$

$$\|\hat{z}_i\| = \left\lfloor \frac{\|z_i\| - \mu_z}{\sigma_z / h} \right\rceil_1^{-1}. \quad (2.5)$$

Here, $\|\hat{z}_i\|$ is the standardized norm of the embedding z_i, normalized using the batch-level average μ_z and standard deviation σ_z. These two statistics are updated over training steps using an exponential moving average:

$$\mu_z = \alpha\mu_z^{(k)} + (1-\alpha)\mu_z^{(k-1)}, \quad \sigma_z = \alpha\sigma_z^{(k)} + (1-\alpha)\sigma_z^{(k-1)}, \quad (2.6)$$

The key innovation lies in the terms g_{angle} and g_{add}, which introduce adaptive modifications. This mechanism ensures that for high-quality images($\|\hat{z}_i\|$ is large), the model applies a larger margin, encouraging tighter intra-class clustering and greater inter-class separation. This helps learn more discriminative features. For low-quality images($\|\hat{z}_i\|$ is small), the margin is reduced, preventing the model from overfitting to noisy or unidentifiable samples.

The adaptive margin function is complemented by a mechanism that adjusts the margin based on image quality, a gradient scaling approach that emphasizes hard samples in high-quality images while de-emphasizing difficult cases in low-quality images and finally a normalization strategy that maintains feature discriminability while adapting to sample difficulty.

This method therefore represents a step forward in the design of *loss* functions adapted to the challenges of facial recognition in diverse, unconstrained environments where image quality can be very low.

2.3 Interpretability & Visual Explanations

Visual explanations methods can be categorized into two main approaches: model-specific methods [3, 22] and model-agnostic methods [1, 2, 4–8, 10].

2.3.1 Model-specific Methods

Model specific strategies work with previous knowledge of the details of the specific structures of the machine learning or deep learning model which is applied. These techniques are used focusing only on a specific model architecture.

The advantage of this strategy is that it allows the developer to get a deeper understanding of the decision having the knowledge of the intrinsic workings of the model, allowing a better customization for the explainable model.

The downside of such techniques is the need for going through entire structures of each model the developer wants to use, what can compromise its performance, this strategy lacks also flexibility.

Some of the most used model specific approaches to deep learning models include strategies based on deconvolution which traverses the path of Convolutional Neural Network (CNN) in reverse order (from final class to original image, pointing out specific regions in the image which contribute to the decision).

Extensions of the deconvolution based approaches include guided backpropagation that will be described in 2.3.1.1 and 2.3.1.2 subsections.

2.3.1.1 Gradient-weighted Class Activation Mapping (Grad-CAM): Visual Explanations from Deep Networks via Gradient-based Localization

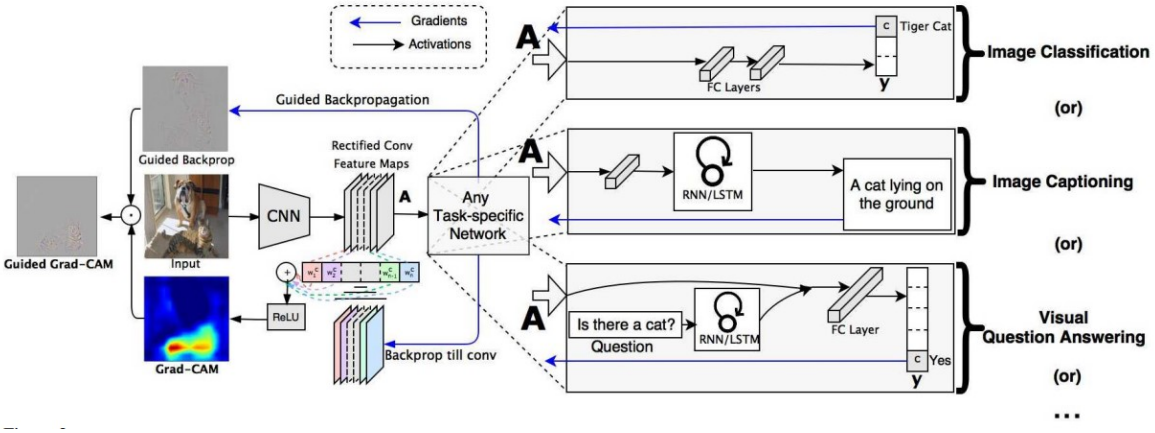
The paper "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" [3], proposed a technique named Grad-CAM, that produces visual explanations for decisions made by Convolutional Neural Networks (CNNs).
The key innovation of Grad-CAM lies in its ability to generate class-discriminative localization maps using the gradients of any target concept flowing into the final convolutional layer. The gradients of a specific class score are computed with respect to feature maps in the last convolutional layer, then they are globally average-pooled to obtain importance weights, which are finally used to create a weighted combination of forward activation maps. The output is a coarse localization map highlighting the regions most important for the model's decision.
The equation in 2.7 represents how the importance weights are calculated for each feature map, for a specific class.

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (2.7)$$

where w denotes the importance weights, κ indexes each feature map, and c refers to the specific class being analyzed. The variable Z represents the total number of pixels in the feature map, and A denotes the feature map activations.

The Grad-CAM visualization map can be calculated as the following equation 2.8 shows:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k w_k^c A^k\right). \quad (2.8)$$



visualization process.

Figure 2.3: **Overview of the Grad-CAM [3] visualization process.** Given an image and a class of interest, Grad-CAM produces a coarse localization map highlighting important regions in the image by computing gradients with respect to the final convolutional layer.

The Figure 2.3 demonstrates how the method generates class-discriminative visual explanations. Given an image and a class of interest, the process begins with a forward pass of an input image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. The signal is then backpropagated to the rectified convolutional feature maps of interest, which are combined to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model has to look to make the particular decision. Finally, the heatmap is pointwise multiplied with guided backpropagation to get guided Grad-CAM visualizations which are both high-resolution and concept-specific.

2.3.1.2 Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks

A year later, an improved version of [3] was published, the authors proposed Grad-CAM++ [22] to provide better visual explanations of CNN model predictions (when compared to Grad-CAM [3]), in terms of object localization as well as explaining occurrences of multiple objects of a class in a single image.
This was achieved because, while Grad-CAM uses a global average of gradients to compute the feature maps weights, this new version, Grad-CAM++ introduces a more complex weighting mechanism based on positive partial derivatives.
The original version struggles when an image contains multiple instances of the same class, the newest version overcomes this problem by assigning weights to each pixel of a feature map using positive partial derivatives as the following equation shows:

$$w_k^c = \sum_{i,j} \alpha_{ij}^k \, \text{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right), \quad (2.9)$$

where w_k^c is the importance weight for the k-th feature map and class c. The term Y^c denotes the class score for class c, while A_{ij}^k refers to the activation at spatial location (i, j) in the k-th feature map. The factor α_{ij}^k represents a pixel-specific weighting coefficient, and the ReLU function ensures that only positive gradients are considered in the importance calculation.

By using the 2.9 equation instead of the 2.8, the problem of identifying multiple occurrences of the same class in an image is solved, as well as the improper object localization problem.

2.3.2 Model agnostic Methods

Model agnostic interpretation methods are way more flexible than the model-specific ones. These methods can be applied to any models or algorithms without having to deeply understand how the model works and without having access to the intrinsic architecture or gradient information of the model.
In this type of methods, the strategy of obtaining explanations consists in doing perturbations to the input data and assess the performance of these perturbations with respect to the original data (without any perturbation or modification) performance.

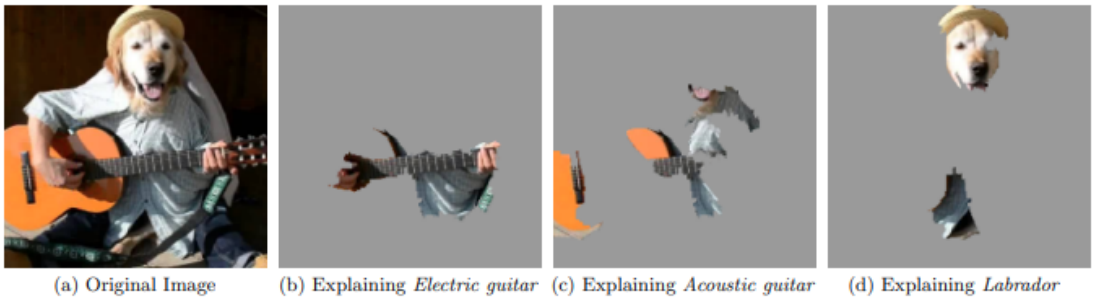
2.3.2.1 LIME: "Why Should I Trust You?" Explaining the Predictions of Any Classifier

LIME [1] was one of the first techniques in the category of model agnostic methods that explains the predictions of any classifier in an interpretable way.

The key idea behind this method is to be able to explain individual predictions by learning an interpretable model that is faithful to the classifier's behavior around a specific instance.

This method works by doing perturbations in the input data and observing how the predictions of the model change. The corresponding perturbations create a new dataset that is weighted by their proximity to the original. After that, an interpretable model is trained on this weighted dataset to approximate the decision boundary.

If the task of explaining an image classification prediction is required, LIME divide the image into contiguous patches of similar pixels (superpixels), create variations by turning some of them on and off. With that, it is able to learn which components most strongly influence the model's prediction. An example of an explanation made by LIME is present in Figure 2.4.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2.4: **Image classification prediction made by Google's Inception neural network explained with LIME [1].** The method perturbs input regions (superpixels) and fits a simple interpretable model to approximate the classifier's behavior locally, highlighting the regions that contribute most to the prediction.

2.3.2.2 RISE: Randomized Input Sampling for Explanation of Black-box Models

Another method that was designed to explain the outputs of black-box models, in this case, particularly for the task of image classification was RISE [2].

It generates importance maps to identify which regions of an image contributed the most to the model's decision-making process.

One of the biggest contributions of RISE was the Random Masking generation technique. This method creates random binary masks at a lower resolution and upsample them to the input image size using bilinear interpolation that randomly obscure parts of an image, these masks are applied to the input image to generate a series of perturbed images.

After that step, using Monte Carlo sampling, a numerous number of random masks are used to estimate the importance of each pixel, the more times a pixel appears in unmasked regions that result in high confidence scores, the higher its importance.

The final step is the importance map computation, this map is a weighted sum of the masks, where weights are derived from the model's confidence scores for the corresponding masked inputs. An overview of RISE pipeline is present in Figure 2.5.

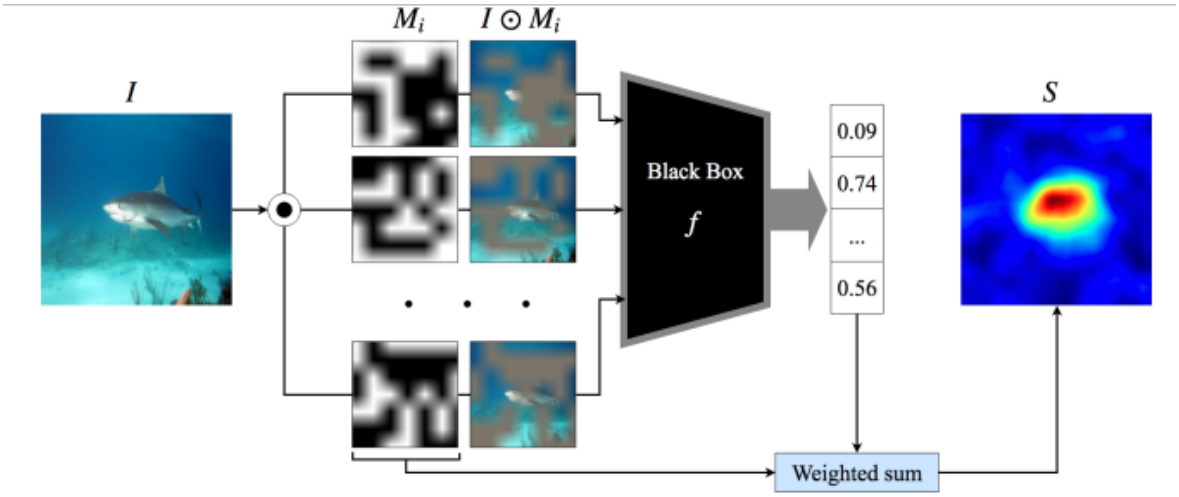


Figure 2.5: **RISE [2] pipeline:** The input image is firstly element-wise multiplied with the generated random masks M_i, these masks are then fed to the model. The target class score for each of the masked inputs determines the weights in the saliency map, which is a linear combination of the masks.

Besides this technique, RISE also introduced two metrics to evaluate the causal relevance of identified regions:

- Deletion Metric: This metric measures the drop in confidence as important pixels are progressively removed from an image.
- Insertion Metric: Measures the increase in confidence as important pixels are added back to a blurred image

2.4 Interpretability & Visual Explanations applied in Face Verification

Most of the proposed methods related to Interpretability in machine learning systems were applied mainly to classification tasks.
Recently, there is a growing interest in applying these methods to face matching tasks, with the objective of telling how important are the parts of a probe image in determining whether they match an enrolled image or not. Some of the SOTA techniques authors in this field have been inspired by [1] and [2] fundamentals to develop their own strategies.

2.4.1 AVG: On Black-Box Explanation for Face Verification

In [4], the authors present six different saliency maps to explain any FV method without intrinsic knowledge of the FR model.

The key idea behind the algorithms lies in how the similarity score of the image pair changes when the probe image is perturbed, as the Figure 2.6 shows. This similarity score s is usually the dot product of vectors $X_A$andX_B, that are respectively the embeddings of images **A** and **B**. If this score is $s = 1$, it means that **A** = **B**.

The matching score is defined as:

$$s = score(A, B). \quad (2.10)$$

In this context, if there is a true pair, meaning the two images are from the same person, the score should be greater than a predefined threshold. **B´** is defined as a perturbed version of **B**.

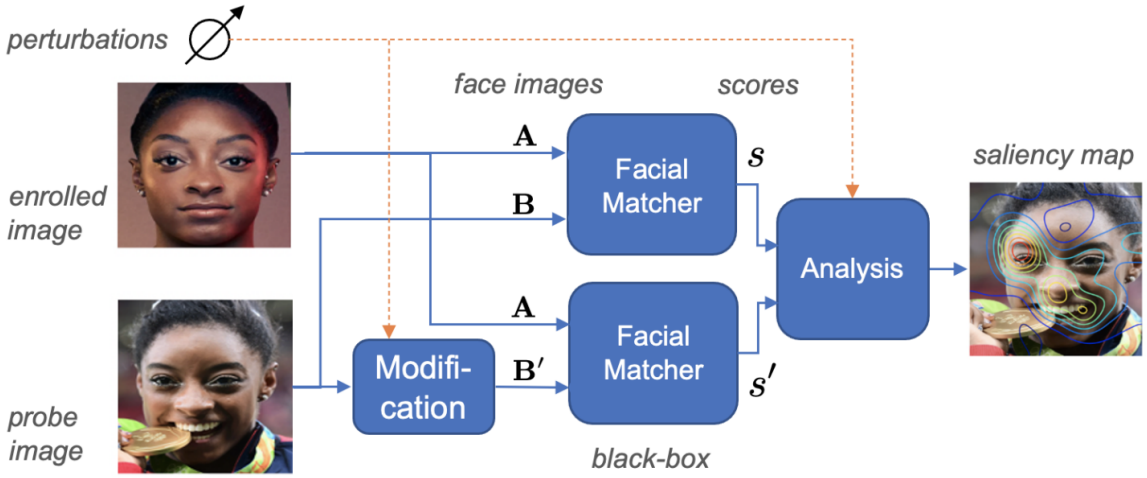


Figure 2.6: **Overview of the AVG [4] method for explaining face verification.** The method perturbs the probe image and monitors changes in similarity scores to generate saliency maps, allowing interpretation of which regions contribute most to the verification decision.

The 6 saliency map methods developed by the authors were:

- **Single Removal(So-):** This method removes circular regions (defined by a Gaussian mask centered at coordinates (i, j)) and calculates the saliency as the difference :

$$H_0^-(i, j) = \text{score}(A, B) - \text{score}(A, B'_{ij}). \quad (2.11)$$

The map that results from here is smoothed using a Gaussian convolution:

$$D = conv(H, G(\sigma)), \quad (2.12)$$

where σ is the width of the convolutional Gaussian kernel.

After that, the smoothed saliency map is normalized to a [0,1] range, using the min-max normalization:

$$S = \frac{D - D_{min}}{D_{max} - D_{min}}. \quad (2.13)$$

- **Greedy Removal(S1-):** Building on the previous method, this one iteratively removes the most important regions, the saliency is computed as:

$$H_1^-(i^*, j^*) = \text{score}(A, B_t) - \text{score}(A, B_{t-1}), \quad (2.14)$$

where (i^*, j^*) corresponds to the coordinates of the most relevant region removed in each iteration of the algorithm. This process begins with the original image $B_0 = B$ and continues until the change in score Δs is below a determined threshold or the predefined maximum number of iterations is reached. The resulting saliency map is smoothed with the same strategy as the previous method.

- **Single Aggregation(S0+):**

  In this method, there is a black image Z at the start, this image is progressively filled with circular regions from B. The saliency map is computed as:

$$H_0^+(i, j) = \text{score}(A, B'_{ij}) - \text{score}(A, Z). \quad (2.15)$$

  This, in a similar way of the previous methods, measures the contribution of each aggregated region centered at (i, j).

- **Greedy Aggregation(S1+):** This method iteratively aggregates the most important regions of B onto Z, recalculating the saliency map until a stopping criterion is met. The saliency is computed as:

$$H_1^+(i^*, j^*) = \text{score}(A, B_t) - \text{score}(A, B_{t-1}), \quad (2.16)$$

  where (i^*, j^*) corresponds to the coordinates of the most important region added in each iteration, as usual, the process continues until the change in score Δs becomes minimal or a predefined number of iterations is reached.

- **Sequential Removal/ Aggregation (SEQ):** This is a hybrid approach where greedy removal and greedy aggregation work alternately.It starts with greedy aggregation until the score is greater than a threshold, and then greed aggregation is used until the matching score is lower than another threshold. This sequence is repeated several times to find a reduced image B_r whose matching score is very similar to the original. The output saliency map is the corresponding map of the last iteration of the algorithm.

- **AVG**: This method combines the first four methods with the equation 2.17:

$$S_{avg} = \frac{S_0^- + S_1^- + S_0^+ + S_1^+}{4}. \quad (2.17)$$

  Based on the experiments done by the authors, the method that achieved the most stable results was AVG of the four methods mentioned before. They also present the results using contour visualization.

The authors also did approaches based on [1] and [2] to have more results to compare with their proposed methods in order to be able to assess the quality of their work.
Figure 2.7 shows an example of the saliency maps and contour visualization maps obtained with the proposed algorithms.

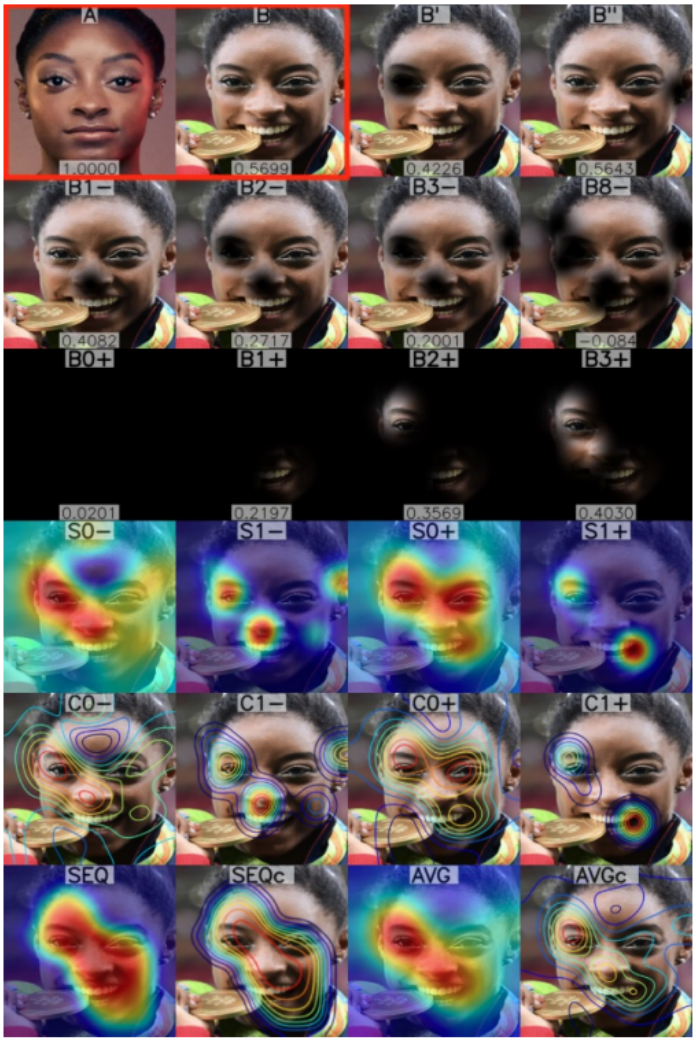


Figure 2.7: **Saliency maps for each algorithm proposed by [4] example.** Row-1: pair(A,B) and perturbed versions of B (B´, B´´). Row-2: Greedy removal algorithm results after iteration t=1,2,3,8, where the most relevant parts of the face are removed. Row-3: Greedy aggregation algorithm results after iteration t=0,1,2,3, where the most relevant parts of the face are aggregated. Row-4: saliency maps that result from SO-, S1-, S0+ and S1+. Row-5: contour visualization of the fourth row. Row-6: saliency maps for algorithms SEQ and AVG with the respective contour visualization SEQc and AVGc.

2.4.2 MinPlus: True Black-Box Explanation in Facial Analysis

MinPlus [10] represents an evolution from its predecessor [4], while AVG introduced the concept of combining removal and aggregation strategies for FV, MinPlus refines this approach to create more stable saliency maps that can be used to explain any facial analysis approach.
Instead of smoothing 2.12 and normalizing 2.13 individual saliency maps before averaging (as in AVG [4]), MinPlus first averages the raw saliency maps ($H_0^-, H_1^-, H_0^+, H_1^+$), then applies smoothing 2.12 and normalization 2.13 to the combined result.
Through this process, MinPlus maintains the original relevance of each saliency map. The normalization method ensures equal importance for each map without weighting, thus preserving their original values.
MinPlus has also been generalized to other facial analysis tasks, such as expression recognition and detection.

2.4.3 xFace: Explainable Model-Agnostic Similarity and Confidence in Face Verification

Addressing the problem with the lack of interpretability in FV systems, [5] proposed a method to enhance their explainability by introducing confidence scores as well as explanation maps, this approach uses model-agnostic techniques to visualize the face regions that are more important to predictions and also calculates a confidence score based on the distribution of feature distances.
An overview of the proposed pipeline is illustrated in Figure 2.8.

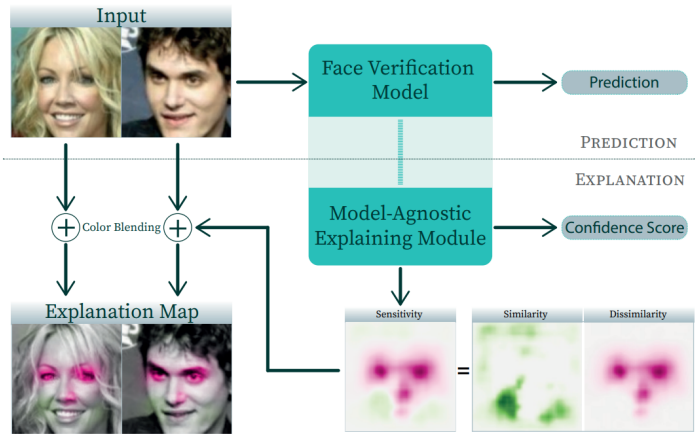


Figure 2.8: **Proposed method [5] pipeline:** It generates a similarity map and blends it with the input images into an x-map, a confidence score is also introduced to explain the decision further.

The proposed confidence score is used to provide a measure of certainty of FV predictions by incorporating the distribution of feature distances of genuine and imposter pairs.
The two embeddings are first obtained from a FR model, and the cosine distance (d) is calculated between the two vectors $(f_1 \text{and} f_2)$:

$$d(f_1, f_2) = 1 - \frac{f_1 * f_2}{||f_1||_2 * ||f_2||_2}. \quad (2.18)$$

Here, the term $\frac{f_1 * f_2}{||f_1||_2 * ||f_2||_2}$ represents the cosine similarity, which measures the angular similarity between two vectors and ranges from -1 (completely opposite) to 1 (completely identical). The cosine distance, defined as $d = 1 -$ cosine similarity, transforms this similarity measure into a distance metric where smaller values indicate greater similarity and larger values indicate grater dissimilarity. In this formulation, d is defined in the range $d \in [0, 2]$. If the two vectors are identical, their cosine similarity is 1, so the distance is $d = 1 - 1 = 0$. If the two vectors are orthogonal, their cosine similarity is 0, leading to $d = 1 - 0 = 1$. Finally, if the two vectors are opposite, their cosine similarity is -1, resulting in $d = 1 - (-1) = 2$.
A threshold (t) is obtained by applying 10-fold cross-validation on the test set, to differentiate between genuine and imposter pairs. An output from a FV with a distance d close to the threshold t reveals an uncertainty in the prediction. On the other hand, a large distance close to 2 or a small distance close to 0 indicates high confidence in the model's prediction.
To provide a more interpretable measure of certainty, the authors introduce a confidence score based on the distribution of cosine distances. First a histogram of cosine distances is created for genuine and imposter pairs, the ratio of genuine to imposter counts in each bin is modeled using a logistic sigmoid function 2.19:

$$c(d) = \frac{L}{1 + e^{-k \cdot (d - d_0)}} + b. \quad (2.19)$$

Here, $c(d)$ approximates the probability that a given distance d corresponds to a genuine pair when $d \leq t$, and an imposter pair when $d > t$. The parameters L,k,d_0, and b are obtained through curve fitting on the dataset's empirical distribution.
Finally, the confidence score (C) is defined as:

$$C = \begin{cases} c(d), & \text{if } d \leq t, \\ 1 - c(d), & \text{if } d > t. \end{cases} \quad (2.20)$$

This confidence score C ranges from 0.5 to 1 and is interpreted as the probability that the model's prediction is correct. Lower values indicate greater uncertainty, while values closer to 1 indicate high certainty in the classification.

Besides the confidence score proposed, the authors also proposed their x-maps, which consist of three different methods. Similarly to [2], [4] and [10] the proposed maps visualize the importance of different facial regions on the model's predictions by systematically occluding image areas, these occlusions use a sliding window approach with different path sizes (p) and strides (s).
In this case, the maps will contain the important face regions for both the images in the pair,

and not only one as the before methods did. So the input is a 2-tuple (I_1, I_2).

After applying the occlusions, a 2-tuple of occluded image sets (O_1, O_2) and a 2-tuple of mask sets (M_1, M_2) are retrieved.

Then, the facial features of every single occluded image O are extracted with a FV network, generating another 2-tuple of feature vector sets (F_1, F_2). To select the 2-tuple of pair-wise distances sets (D_1, D_2), three methods are employed:

- **Method 1:** The cosine distance is averaged across all occluded locations for one of the input images, while comparing it with the occluded(at a particular location) image of the other input image.

$$D_1 := \left\{ \frac{1}{N} \sum_{j=1}^{N} d(F_1^i, F_2^j) \mid \forall i \in [1, 2, \ldots, N] \right\}, \quad \textbf{(2.21)}$$

$$D_2 := \left\{ \frac{1}{N} \sum_{i=1}^{N} d(F_1^i, F_2^j) \mid \forall j \in [1, 2, \ldots, N] \right\}. \quad \textbf{(2.22)}$$

- **Method 2:** Compares the cosine distance between a non-occluded image and an occluded counterpart, with this, assessing how occluding specific regions in one image affects the similarity to the non-occluded corresponding image.

$$D_1 := \left\{ d(F_1^i, N(I_2)) \mid \forall i \in [1, 2, \ldots, N] \right\}, \quad \textbf{(2.23)}$$

$$D_2 := \left\{ d(N(I_1), F_2^j) \mid \forall j \in [1, 2, \ldots, N] \right\}. \quad \textbf{(2.24)}$$

- **Method 3:** This method measures the cosine distance between co-located occluded regions in both images.

$$D_1 = D_2 := \left\{ d(F_1^{(i)}, F_2^{(i)}) : \forall i \in [1, 2, ..., N] \right\}. \quad \textbf{(2.25)}$$

After this and independently of the method, a 2-tuple of distance sets is obtained (D_1, D_2), which is then compared with the original distance $d_{orig} = d(I_1, I_2)$ of both non-occluded input images. The difference between the cosine distance of occluded images (d_i) and the original distance (d_{orig})) is used to weight the occlusion masks. This results in similarity maps (S) that highlight the regions of the face most important to the prediction:

$$S = \sum_{i=1}^{N} \frac{(d_i - d_{orig}) \cdot M_i}{N}, \quad (2.26)$$

with $d_i \in D$ and $M_i \in M$. With this, the deviation caused by an occlusion at a certain location is visualized.
Finally, three similarity maps S are obtained for each pair of images (I_1, I_2).

2.4.4 FV-RISE: A RISE-Based Explainability Method for Genuine and Impostor Face Verification

FV-RISE [6] adapts the RISE [2] algorithm to provide comprehensive explanations for FV decisions.
The main contribution in this approach when compared to the previous strategies is that FV-RISE generates two types of heat maps to explain both acceptance and rejection decisions: Similarity Heat Maps (S-HMs) and Dissimilarity Heat Maps (D-HMs). Figure 2.9 contains an overview of FV-RISE.

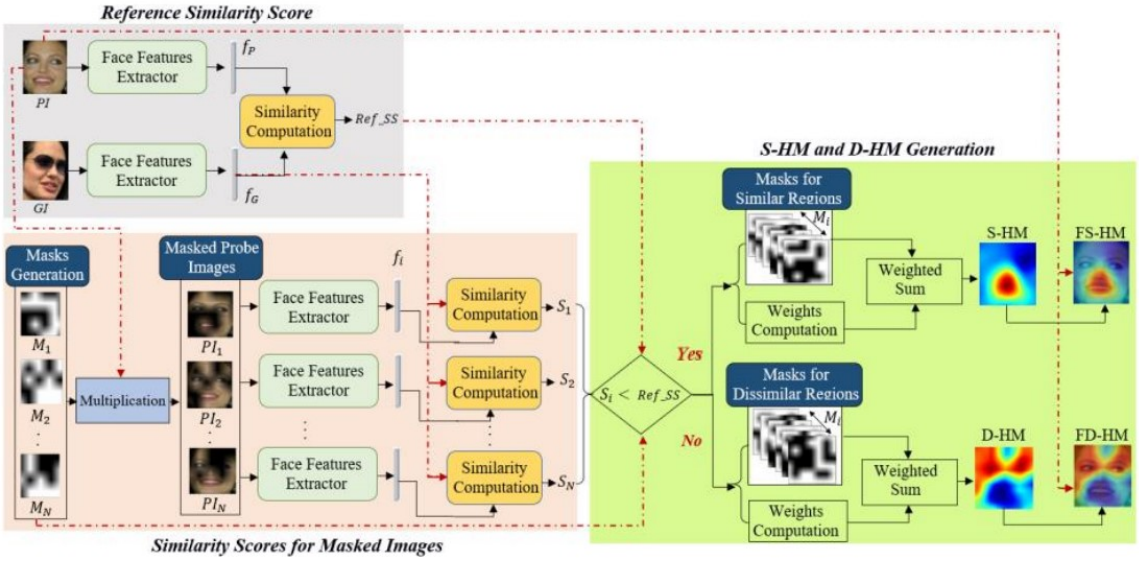


Figure 2.9: **Overview of the FV-RISE [6] method** The approach adapts the RISE algorithm to generate similarity and dissimilarity heat maps by comparing similarity scores of randomly masked images against a reference, offering interpretable visual explanations for both genuine and impostor cases.

First, the method computes a reference similarity score between the probe and gallery images using a FR model. Then, as usual in post-hoc agnostic methods, it generates a predefined number of random masks with values between 0 and 1, which are used to perturb random regions of the probe image.
For each probe image that was perturbed, the method computes a similarity score S_i with the gallery image.By comparing these scores to the reference score Ref_{SS}, the method determines whether the masked regions correspond to similar or dissimilar face areas:

If S_i is smaller than Ref_{SS}, the masked face region is considered relevant for an acceptance decision, corresponding to a similar face region in the pair, otherwise, it is considered non-important for an acceptance decision, corresponding to a dissimilar face region in the pair. So basically, the comparison results are used to categorize the masks and generate two different heat maps.
The final visualization is created by superimposing these heat maps over the original probe image's luminance channel.

2.4.5 CorrRISE: Towards Visual Saliency Explanations of Face Verification

Recently, [7] proposed CorrRISE, another model-agnostic method for explaining FV systems. The key innovation of this method lies in its ability to generate meaningful saliency maps for non-matching cases and its innovative approach to this process, since most of the methods focus only on the relevant parts between two images but not considering the irrelevant parts.
The algorithm consists of two main components: mask generation and correlation-based saliency map generation.These components can be seen in detail in Figure 2.10.

In the mask generation phase, CorrRISE creates multiple random masks containing small square patches, with random values between 0 and 1, in various locations of an image.

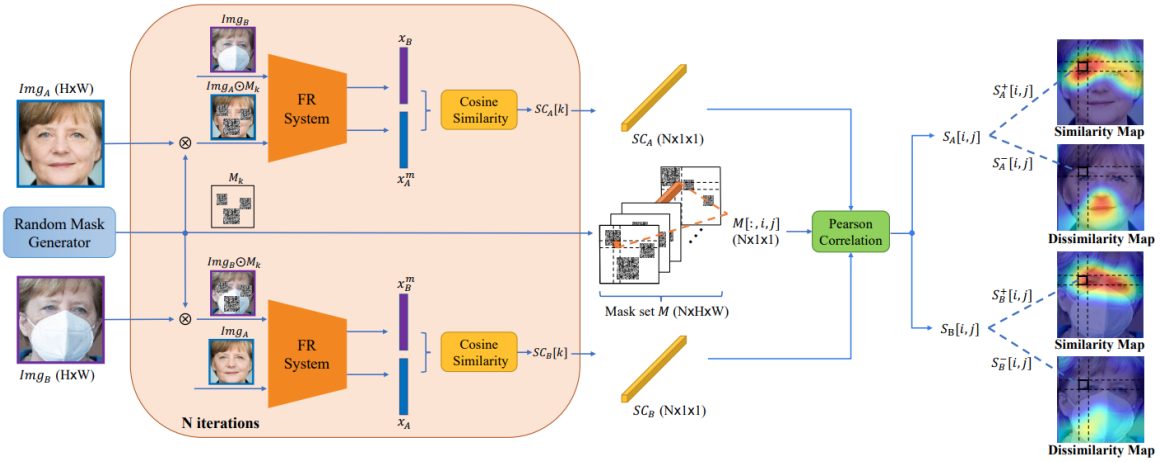


Figure 2.10: **Architecture of CorrRISE [7] method, given a face pair as input, the similarity and dissimilarity maps are calculated.** The middle block repeats N iterations using different random masks. Then, the obtained similarity scores and the mask set are fed to the Pearson correlation module to finally calculate the saliency maps.

The next step is the correlation-based saliency map generation, during this phase the generated masks are applied to the input images, and the resulting masked images are processed through a FR model to obtain similarity scores. Given N that is the number of the masks $M = \{M_i, i = 1, ..., N\}$ and SC_i that is the cosine similarity score obtained, after iterating all the N masks, the list of of scores $SC = \{SC_i, i = 1, ..., N\}$ that corresponds to the mask list is recorded. The final saliency map for an image is then obtained by performing pixel

wise Pearson correlation between SC and M. Finally, the position of positive correlation coefficients represents the regions that are similar, while the position of negative coefficients represents the dissimilar regions.

2.4.6 Bridging Human Concepts and Computer Vision for Explainable Face Verification

A novel approach to improve the interpretability of face verification systems was proposed by [8]. This method aims to explain FV models decision but aligning the technical output with the human cognitive processes.
This approach consists of three main phases: semantic extraction using human-defined facial regions, concept importance analysis and similarity mapping through strategic perturbation of facial features. An overview of the approach can be seen in Figure 2.11.

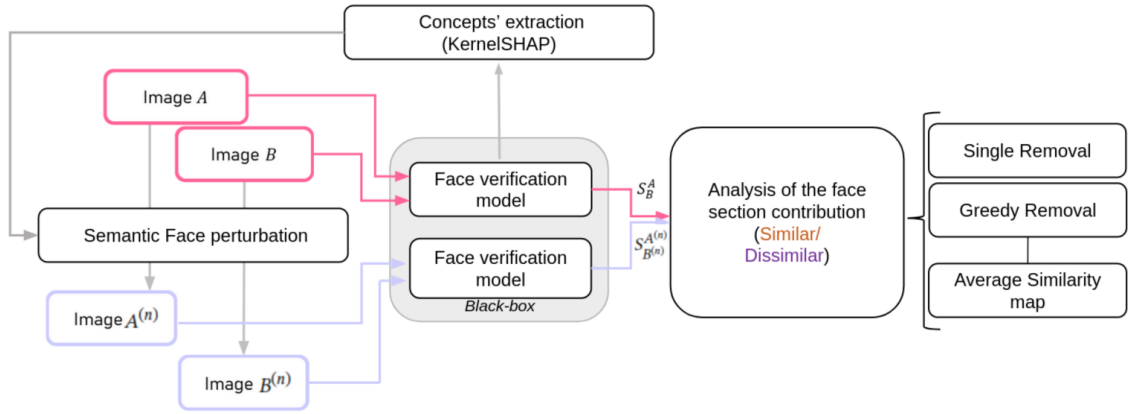


Figure 2.11: **Flowchart of the method proposed by [8] for explainable face verification.** The approach uses semantic facial regions defined by human concepts and applies SHAP-based analysis and strategic perturbations to generate interpretable similarity maps aligned with human perception.

The key innovation and the foundation of the method relies on MediaPipe, a framework that outputs an estimate of 478 3-dimensional face landmarks. Using these landmarks, the authors predefine 13 distinct semantic regions of the face that correspond to intuitive human-interpretable areas such as eyes, nose, and mouth. In contrast to previous approaches where the perturbations (masks) applied to the images were square patches or random circular binary masks, in this method the masks will be applied to the semantic regions defined before.
After defining the regions, there is a concept importance analysis: To determine which facial regions are most relevant for the FV process, the method employs KernelSHAP [23], combining LIME's [1] interpretable components with shapley values from game theory [24], this process involves extracting importance scores for each semantic region and analyzing 512 features per region. The method then obtains absolute Shapley Additive Explanations (SHAP) values to account for both positive and negative contributions. Subsequently. the scores are aggregated across multiple images, and finally, the regions are ranked using the borda [25] count voting technique.
Following this, the next phase consists in obtaining the similarity maps through perturbation,

in this case, the method introduces three types of similarity maps algorithms drawing inspiration from [4]. However, this approach has significant differences, as mentioned above, from previous research: The usage of semantically meaningful masks with a fixed shape.
The used algorithms consisted of three approaches. Single Removal (S0) applies mask individually to corresponding areas in both images, then calculates the contribution scores and finally defines the similarity map as the sum of the negative and positive contributions normalized. Greedy Removal (S1) employs an iterative approach where the most impactful facial regions are repeatedly removed, then the cumulative effects on the verifications scores are obtained. This iteration continues until reaching minimal score differences or until the maximum number of iterations is reached. Finally, the average of both approaches (S_{AVG}) combines S0 and S1 maps to provide a comprehensive view that incorporates both individual and collaborative feature contributions.

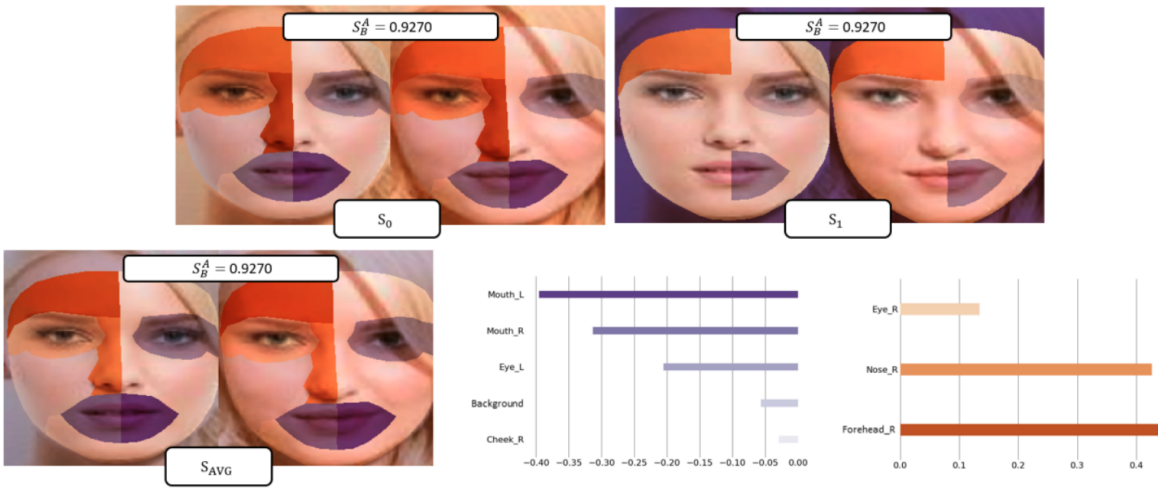


Figure 2.12: **Similarity maps proposed by [8].** S_0 is the output for the single removal algorithm, S_1 refers to the greedy removal one, and S_{AVG} is the average map obtained from S_0 and S_1. There is also a plot chart that considers the contribution values (C_n) for each section in the mask.

For the visualization, the authors use orange to indicate similar facial regions and purple to represent dissimilar areas. An example of the similarity maps obtained with this approach can be seen in Figure 2.12.

2.5 Face Image Inpainting

Face image inpainting is the task of generating plausible facial structures for missing pixels in a face image. This task requires not only filling in missing pixels, but also understanding and preserving the human face anatomical structure and unique characteristics.
In the scope of this work, inpainting is used not only for restoration but as a means of perturbing specific face image regions in a controlled and realistic way.
By modifying these regions with plausible alternatives, it becomes possible to analyze the importance of different face regions for the FV model. This approach contrasts with traditional methods that rely on binary masks only.
With this approach, perturbations that retain contextual and semantic coherence are allowed, thereby enabling a deeper and more interpretable analysis of model decisions.

In this work, we adopt the **RePaint** model [9], a Denoising Diffusion Probabilistic Models (DDPMs)-based approach that represents a significant advancement over traditional Generative Adversarial Network (GAN) or autoregressive methods.
The inpainting process begins with a noisy image and iteratively denoises it, conditioned on known image regions.
The stochastic nature of the DDPMs allows RePaint to produce realistic outputs that harmonize seamlessly with the surrounding content. The authors also introduce a novel resampling strategy that significantly improves the quality of generated content, rather than simply slowing down the diffusion process, RePaint implements a forward-backward resampling approach to improve the inpainting quality.
In a standard DDPM, an image is progressively denoised from a pure noise sample, iterating through time steps until a final image is generated. However, inpainting requires the model to integrate new content while preserving the known parts of the image. If this process follows a strict forward diffusion schedule, the generated content may not harmonize well with the given image, leading to boundary inconsistencies between the inpainted and known regions.
To address this, RePaint modifies the standard reverse diffusion process by occasionally taking steps forward in diffusion time before resuming the backward process. This approach allows the model to revisit earlier noise states, which helps in reintroducing coherence between the inpainted regions and the known parts of the image.
As illustrated in the Figure 2.13, RePaint's pipeline takes as input both the original image and a binary mask to produce the final inpainted result.

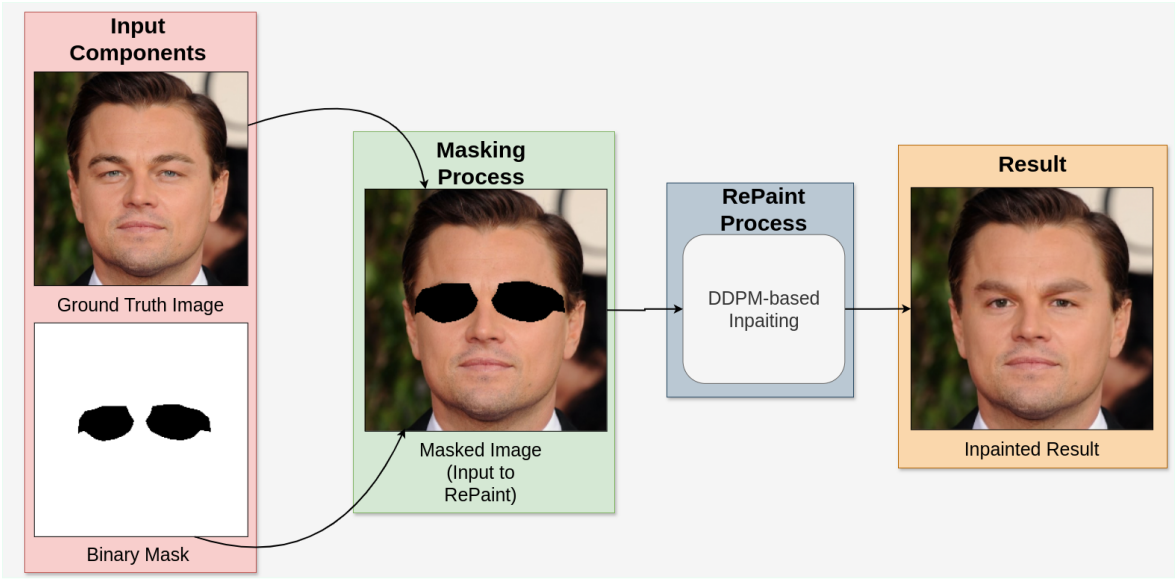


Figure 2.13: **Overview of RePaint's [9] inpainting pipeline:** Given an original image x, the mask m, the masked image $m \odot x$ serves as input to the DDPM-based inpainting process. RePaint leverages DDPMs to generate realistic completions by iteratively denoising while preserving the known regions through resampling.

In the context of this dissertation project, RePaint is employed to perturb specific regions of face images in a semantically meaningful manner, with the objective of generating realistic content but different from the original face image. The realism of the inpainted regions is cru-

cial to ensure that perturbed images remain visually coherent, plausible from the perspective of the FR model, guaranteeing that the inpainted samples remain closer to the training data distribution of the FR model. Unlike binary masks, realistic reconstructions reduce the introduction of artifacts that could distort the model's similarity score, leading to more faithful attribution of importance to the affected regions.

2.6 Conclusions

The research about the existing interpretability methods devised for FV provides a useful insight of the drawbacks of current approaches and what might be used as the foundation for a suggested approach.

The main challenge of the studied methods is to provide a clear and objective explanation of the regions that contribute the most to the FV model decision. This happens because the majority of the current perturbation based techniques use circular regions or square patches (instead of semantic face regions) to perturb the probe image or both in the pair, resulting in a subjective heatmap. In Figure 2.14, the results from [10] on a genuine pair are presented. It is possible to see that the LIME [1] and RISE [2] adaptations from the authors provide saliency maps that are not precise and is not very clear which region has the most impact in the model's decision. On the other hand, MinPlus [10] and AVG [4] offer superior explanations, since the saliency map is more precise and focuses on a particular area, being easier to understand. In Figure 2.15, the results on an impostor pair are presented, the conclusions drawn above are supported, where MinPlus [10] and AVG [4] achieve the best results. Since RISE [2] and LIME [1] contain randomness in the perturbations, the results can be unpredictable and less precise. Both MinPlus and AVG are only capable of generating similarity maps, so for impostor pair examples, the output is not very interpretable.

Figure 2.16 presents the obtained results for a genuine pair using [5]. Again, it is noticeable that the green region covers, in the 3 maps, almost every face region, so it is difficult to assess what is the region of the face that contributed the most.

The results for an impostor pair using [5] are shown in Figure 2.17. In this example, the first and second x-maps show a large area of similarities despite the pair containing two different persons. The last x-map is the one that behaves best for this example, where both face images have similar poses.

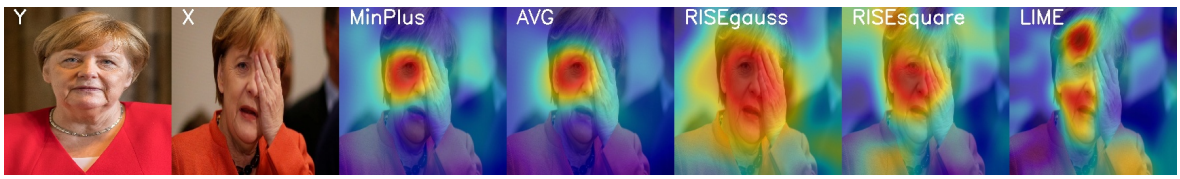


Figure 2.14: **Results obtained from [10] on a genuine pair example.** Y corresponds to the gallery image, X corresponds to the probe image, MinPlus and AVG are the algorithms proposed by [10] and [4] respectively. RISEgauss, RISEsquare and LIME are adaptations of the methods proposed by [2] and [1].

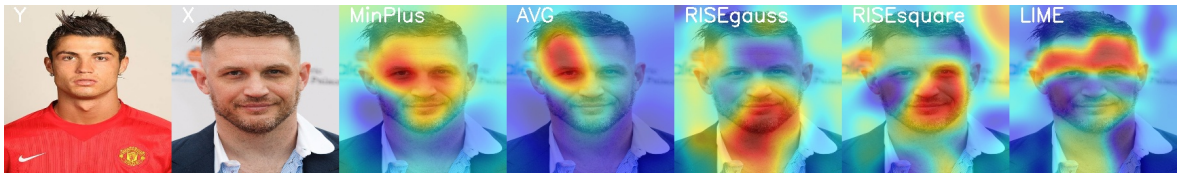


Figure 2.15: **Results obtained from [10] on a impostor pair example.** Y corresponds to the gallery image, X corresponds to the probe image, MinPlus and AVG are the algorithms proposed by [10] and [4] respectively. RISEgauss, RISEsquare and LIME are adaptations of the methods proposed by [2] and [1].

	Image 1	Image 2
Input Images		
X-Map 1		
X-Map 2		
X-Map 3		

Figure 2.16: **Comparison of the three proposed explanation maps algorithms proposed by [5] for a genuine pair**. Green represents similar face regions and purple represents dissimilar regions.

	Image 1	Image 2
Input Images		
X-Map 1		
X-Map 2		
X-Map 3		

Figure 2.17: **Comparison of the three proposed explanation maps algorithms proposed by [5] for an impostor pair.** Green represents similar face regions and purple represents dissimilar regions.

That's why using semantic face perturbation instead, seems to be the way to generate similarity maps that are more objective and more specific in terms of explanation.
There are also methods that generate similarity maps for both images in the pair instead of just for the probe one, this approach requires that both face images have a similar pose, otherwise, the generated similarity map will lose most of its meaning.
Using only binary masks for the perturbation purpose is also a potential issue, since these are unrealistic perturbations, and they are also not representative of real-world variations. These unnatural artifacts can disrupt the model's feature extraction processes, leading to changes in the similarity score that may not genuinely reflect the importance of the masked region.

Chapter 3

Proposed Method

3.1 Introduction

Based on the revision done and provided in Chapter 2 it was possible to assess what are the main things that can be improved in current explainability methods adapted to the FV task. It was also possible to determine what are the key ideas and methodologies that can be useful to develop a novel idea.
In this chapter, we present our proposed explanation framework designed to improve the interpretability of face verification decisions through semantically guided perturbations. While previous model-agnostic approaches often rely on binary masks that are not realistic, these perturbations differ significantly from the type of data used during the training of face recognition models. As a result, they can disrupt the model's feature extraction process, leading to changes in the similarity score that may not genuinely reflect the importance of the masked region. Our method aims to preserve facial realism by replacing masked regions with plausible content generated via face inpainting. Our method operates by systematically perturbing semantic face regions and reconstructing them with an inpainting model, the impact of each region on the model's similarity score is then measured, resulting in similarity maps that highlight the most influential areas in the verification decision. To generate the similarity maps, we propose two complementary strategies: **Single Inpaint (S0)**, which evaluates the individual contribution of each face region in isolation, and **Greedy Inpaint (S1)**, which captures the joint influence of multiple regions through an iterative process.

3.2 Methodology

This section describes the framework developed to generate interpretable visual explanations for face verification decisions. The approach is structured into two stages, beginning with the extraction of semantic face regions, followed by the application of realistic perturbations using an inpainting model. The following subsections provide a detailed explanation of each component of the pipeline (Figure 3.3).

3.2.1 Semantic Masks Extraction

To enable region-specific perturbations that are semantically meaningful, our method relies on the extraction of facial masks that segment the input face images into distinct facial regions. Inspired by the strategy introduced by [8], we adopt an approach that identifies key regions of the face, allowing the analysis of the impact of each region on the face verification process with more precision.

The extraction process begins with the application of a facial landmark detection model. We employ MediaPipe Face Mesh [26] that estimates 468 3D facial landmarks from RGB images in real time. These landmarks serve as the basis for defining the contours of several facial regions used in our method, including the forehead, eyes, eyebrows, nose, mouth, cheeks, and chin. A visual example of the landmark distribution is presented in Figure 3.1.

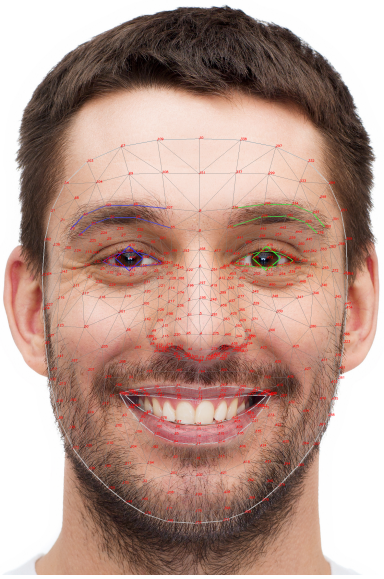


Figure 3.1: **Face landmark overlay generated by MediaPipe Face Mesh**, illustrating the 468 landmark points used as a reference to define polygonal face regions in the proposed method.

MediaPipe was selected because it provides high-resolution, dense, and consistent landmark estimates that are well suited for defining semantically meaningful face regions. No alternative landmark detection methods were considered, as MediaPipe fully satisfied the requirements of our approach in terms of accuracy and ease of integration.
However, some regions of interest, particularly hair and ears, are not represented in the facial landmark topology provided by MediaPipe, making it impossible to extract these masks using landmark estimation alone. To address this, we complement our approach with a semantic segmentation model. Specifically, we utilize a face parsing model [27] based on [28] that was trained on the CelebAMask-HQ dataset [29]. This solution was selected because it is widely used in face parsing tasks, easy to integrate, and provides accurate segmentation of multiple face regions, including the hair and ears, which are essential for out perturbation strategy but not available through landmark-based methods.
By combining both approaches, we construct a comprehensive set of nine facial masks for each image: forehead, eyebrows, eyes, nose, cheeks, mouth, chin, hair, and ears. The choice of these regions reflects a balance between being too specific (which could fragment the face into overly fine-grained parts) and too general (which could obscure localized effects). Moreover, these regions correspond to clearly identifiable and nameable parts of the face, which makes the analysis more intuitive and aligned with how humans typically describe facial features. They can also be reliably extracted using face parsing models, making them suitable for systematic perturbation and attribution. An illustration of these regions, along with their corresponding masks, is provided in Figure 3.2.

Figure 3.2: **Illustration of the nine semantic face regions used for guided perturbations.** Each color corresponds to a specific face region mask (e.g., Forehead (Purple), Eyes (Green), Nose (Orange), Mouth (Red), Cheeks (Pink), Chin (Cyan), Eyebrows (Yellow), Hair (Brown) and Ears (Gray)) obtained by combining the output of a face landmark detection model with a general-purpose semantic segmentation model.

Each mask defines an area of the face that will be individually or jointly perturbed using inpainting to assess its impact on the similarity score.

3.2.2 Similarity Map Generation Algorithms

The main goal of our framework is to generate interpretable similarity maps that highlight the contribution of different face regions to the output of a face verification model. To achieve this, we extend the Single Removal (S0) and Greedy Removal (S1) proposed by [4], where we adopt a perturbation-based strategy in which selected face regions are masked and then reconstructed using an inpainting model, ensuring that the perturbed images remain visually coherent and lie closer to the distribution the verification model was trained on. This approach reduces the risk of producing misleading or exaggerated similarity shifts due to unnatural artifacts.

3.2.2.1 Single Inpaint (S0)

In this strategy, each semantic region is perturbed independently. Given a pair of face images (Img_A, Img_B) and their feature representations (x^A, x^B) extracted from the visual encoder, we calculate a reference similarity score:

$$S_{\text{ref}} = sim(x^A, x^B), \quad (3.1)$$

where sim represents the cosine similarity function applied to the embeddings extracted by the FR model. The algorithm begins by extracting $N = 9$ binary semantic masks for both images in the pair $\{Img_A, Img_B\}$, where each mask represents a different face region. The set of masks for image A and B are given by $M_A = \{M_{A,i} \mid i = 1, \ldots, N\}$ and $M_B = \{M_{B,i} \mid i = 1, \ldots, N\}$, respectively. For the i^{th} semantic region, the corresponding masks $M_{A,i}$ and $M_{B,i}$ are separately applied to Img_A and Img_B, generating masked versions where the targeted re-

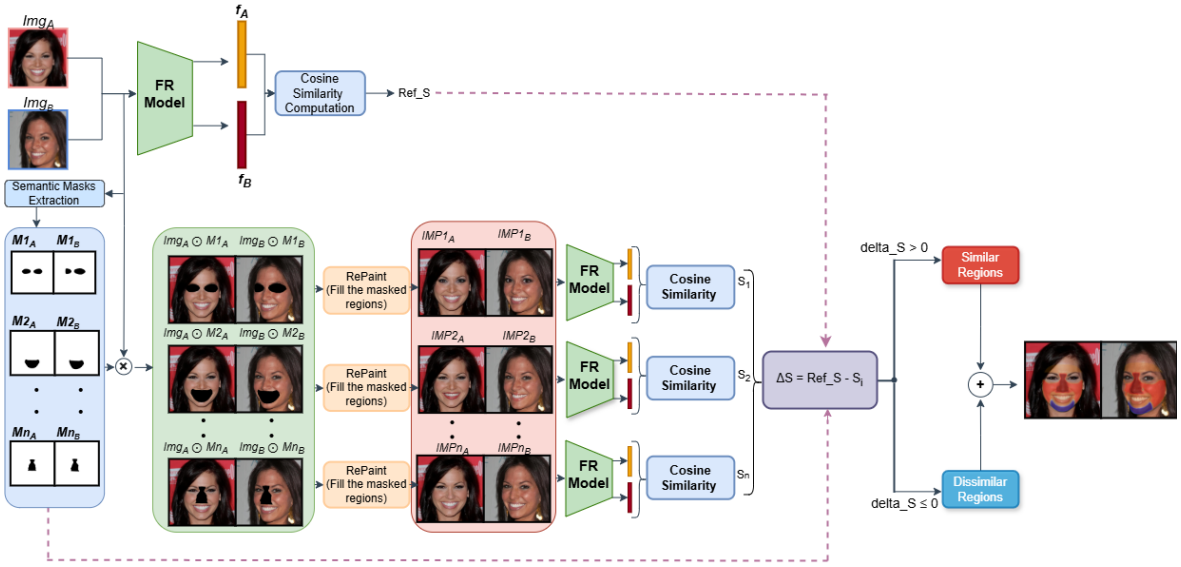


Figure 3.3: **Overview of the proposed method explanation pipeline.** Starting from an input face pair, semantic masks are extracted to segment the face into distinct regions. Each region is masked and then reconstructed to generate modified image pairs. A FR model computes a reference similarity from the original pair, and a set of region specific similarity scores from each modified pair. The contribution of each region to the verification decision is quantified as the difference between the reference and perturbed similarity scores. Regions whose perturbation leads to a significant decrease in similarity are considered similar, while those causing the score to increase are considered dissimilar.

gion is occluded. These masked images are processed through the inpainting model, which fills in the occluded areas with realistic but different from the original reconstructions. As the inpainting process is inherently stochastic, there is no guarantee that the reconstructed content in both images will be visually aligned. This randomness may introduce variability in the resulting similarity score, however, in practice, the perturbation strategy still reveals consistent trends across multiple samples and regions. The similarity score between the inpainted image pair is then calculated, and this step is visually illustrated in Figure 3.4. The score is obtained as:

$$S_i = sim(x_{in}^A, x_{in}^B), \quad (3.2)$$

where x_{in}^A and x_{in}^B are the embeddings extracted from the inpainted images. The contribution ΔS_i of the i^{th} region pair is determined by the change in similarity score resulting from its perturbation:

$$\Delta S_i = S_{\text{ref}} - S_i. \quad (3.3)$$

A positive ΔS_i (decrease in similarity after inpainting) suggests the original region i contributed positively to the match, meaning that it was probably a similar region. On the other hand, a negative ΔS_i (increase in similarity) suggests it was a source of dissimilarity. This procedure is systematically repeated across all N semantic regions, building a complete similarity map that reveals the contributions of each face region to the final verification score. Then, we calculate a normalized value $\Delta S_i'$ for each processed region i by dividing its raw con-

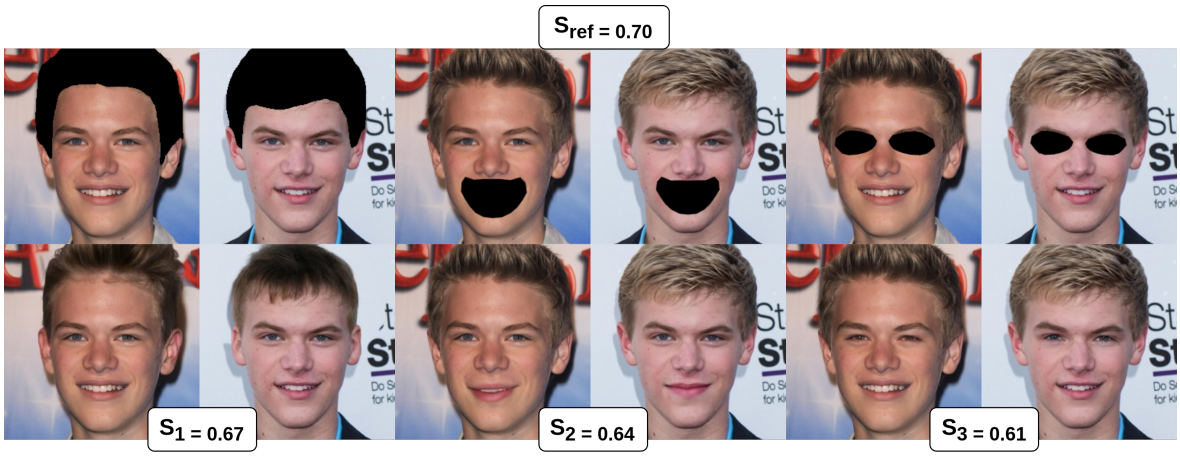


Figure 3.4: **Visual examples illustrating single-region inpainting (S0) perturbations.** Each pair of columns corresponds to a specific face region being perturbed: (1-2) hair, (3-4) mouth, and (5-6) eyes. The top row presents the masked versions of the images, where the selected region has been occluded using a semantic mask. The bottom row shows the corresponding inpainted reconstruction.

tribution ΔS_i by the maximum absolute contribution observed across all processed regions for the given image pair:

$$\Delta S_i' = \frac{\Delta S_i}{\max_j |\Delta S_j|}, \quad (3.4)$$

where the denominator corresponds to the maximum absolute value among all calculated contributions ΔS_j for the current image pair. Normalization by region area was considered during the design of the method but not applied. This decision was made to avoid disproportionately favoring very small regions, which could receive artificially high importance scores due to their size. Since the objective is to evaluate the overall impact of perturbing each semantically meaningful region, the variation in similarity is used directly, without adjusting for area. These normalized values are then used to construct pixel-level similarity maps, where each pixel within a region inherits the score of that region. Two similarity maps, one for each image in the pair, are produced to highlight the regions that most influence the verification score.

3.2.2.2 Greedy Inpaint (S1)

While the S0 strategy captures the independent effect of each region, it does not account for interactions between regions. To address this, we introduce the Greedy Inpaint strategy, which performs perturbations iteratively and cumulatively. To capture both similarity and dissimilarity effects, the algorithm involves two separate greedy searches: one minimizing similarity (Negative Run) and one maximizing it (Positive Run).

The algorithm begins with the unmodified face pair $\{Img_A, Img_B\}$ and their initial similarity score S_i (Eq. 3.2), two intermediate pixel-level contribution maps are initialized to zero: $H_{neg,A/B}$ for accumulating negative run contributions, and $H_{pos,A/B}$ for positive run contributions. In each iteration t of a negative or positive run, the algorithm evaluates the effect of

inpainting each remaining available region i onto the current cumulatively modified images. . It calculates the resulting similarity score $S_t(i)$ for each candidate, and the region i_t^* with highest/lowest $S_t(i)$ score in a positive/negative run is selected. The corresponding incremental change in score $\Delta S_t = S_t(i_t^*) - S_{t-1}$ determines the contribution: for the Negative Run, if the score decreased ($S_t < S_{t-1}$), the magnitude of the decrease ($S_{t-1} - S_t$) is added to $H_{neg,A/B}$ for the selected region i_t^*, for the Positive Run, if the score increased ($S_t > S_{t-1}$), the magnitude of the increase ($S_t - S_{t-1}$) is added to $H_{pos,A/B}$ for the selected region i_t^*. This process is visually illustrated in Figure 3.5.

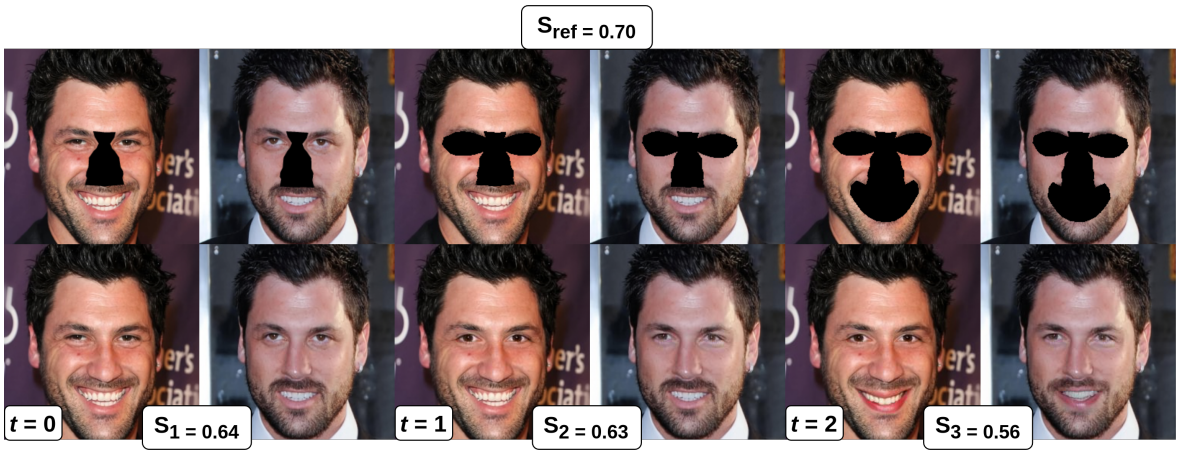


Figure 3.5: **Visualization of the greedy Inpaint (S1) strategy across successive iterations.** Each pair of columns corresponds to a specific iteration *t*. The top row displays the images after masking the selected region at that iteration, while the bottom row shows the corresponding inpainted reconstructions generated by the inpaint model.

These contributions are applied spatially across the mask $M_{A/B,i_t^*}$. The selected region is removed from the candidate set of regions, and the inpainted state is carried to the next iteration. This continues until the maximum number of iterations is reached or when the score difference is lower than a predefined threshold. This continues until the maximum number of iterations is reached or when the score difference is lower than a predefined threshold. After both runs complete, the final pre-normalized contribution map $H_{A/B}$ is calculated by combining the accumulated contributions. This map is then normalized using Equation 3.4. These maps reflect the joint importance of regions for the similarity score, considering the order and cumulative impact revealed by the greedy searches.

3.3 Conclusion

Bridging the gap between human understanding and face verification systems begins with explanations that are both faithful, precise, and intuitive. The method introduced in this chapter moves beyond simplistic occlusion techniques by leveraging semantic perturbations and inpainting-based reconstructions. By systematically analyzing the impact of each facial region, both in isolation and through joint interactions, this framework offers an interpretable approach to visual explanations.

Chapter 4

Experiments and Results

4.1 Introduction

This chapter presents the experimental setup and results used to evaluate the proposed explanation method for face verification. The objective is to assess the interpretability, precision, and reliability of the generated similarity maps, both through comparisons with existing explanation methods and through an analysis of the outputs produced by our approach in isolation. We begin by describing the data preparation and configuration steps involved in the experimental setup. We then present both qualitative and quantitative results, including a visual comparison and an ablation study. Finally, we discuss the strengths and limitations observed throughout the experiments, highlighting the interpretability gains provided by the proposed method.

4.2 Implementation Details

We conduct our experiments using the CelebAMask-HQ dataset [29]. This dataset was chosen because it provides high-resolution face images, which are essential for obtaining visually coherent inpainting results. Additionally, the dataset includes identity labels for each subject, which allows for the construction of both genuine (same identity) and impostor (different identities) image pairs. Specifically, we select a subset of image pairs (genuine and impostor) for qualitative visualization and analysis. For the ablation study all genuine pairs were considered. All the images are cropped and resized to 112x112 pixels, following the input size constraint imposed by the xFace [5] method. The similarity scores that guide our explanations are obtained using a pretrained state-of-the-art face recognition model [11]. For the inpainting process, we utilize the RePaint [9], chosen for its ability to generate diverse and semantically consistent face image completions. In addition to our proposed method, we compare against two state-of-the-art face verification explanation methods MinPlus [10], xFace [5]. We also include the adapted versions of LIME [1], and RISE [2], made publicly available by the original authors of [10]. Despite being promising, we decided to exclude CorrRise [7] from our comparisons, as no official implementations or publicly available code is provided by the authors. It is important to note that no models were trained as part of this work. All experiments were conducted using pretrained models, so, both the face verification model and the inpainting model were used solely for inference.
All experiments were performed on a desktop computer. The full system specifications are presented in Table 4.1.
In addition to reporting system specifications, it is important to quantify the computational effort of the main components of the proposed method. Table 4.2 presents the average ex-

Table 4.1: System Specifications

CPU	AMD Ryzen 5 1600 six-core processor × 12
GPU	NVIDIA GeForce RTX 2080 Ti
RAM	16GB DDR4

ecution times measured for each step in the pipeline, based on multiple runs. These values are intended to give an estimate of the time requirements associated with each component, particularly emphasizing the high cost of the inpainting component.

Table 4.2: Average computational time (hours) required by each component of the proposed method.

Component	Avg. Time per Pair (hours)
Semantic Mask Extraction	0.0013
Inpainting (S0)	1.0666
Inpainting (S1)	19.0833
Similarity Map Generation	0.0018
Total Runtime (S0)	1.0694
Total Runtime (S1)	19.0083

4.3 Metrics

Evaluating the obtained explanation maps poses an inherent challenge due to the subjective nature of interpretability. The usefulness of an explanation is determined by how well it aligns with human understanding. Since interpretability is a human-centered concept, relying exclusively on quantitative metrics may overlook important nuances that are best captured by the human eye.
When comparing multiple explanation methods, it is often possible to qualitatively assess differences in visual properties such as precision (the extent to which an explanation map segments facial features with a high degree of specificity and accuracy), clarity (the extent to which an explanation map is easy to understand and interpret), and, for example, how often a certain method tends to fail in difficult cases. These qualitative impressions can provide valuable insights, particularly when the maps are analyzed side-by-side.
To complement these visual assessments with objective comparisons, we adopt quantitative metrics, in particular, we employ the Deletion and Insertion metrics, originally proposed by [7].

Deletion: This metric measures how rapidly the model's verification accuracy deteriorates when the most salient pixels are progressively removed from the input. Specifically, a percentage of the most important pixels is masked in successive steps, and the accuracy is calculated on the modified dataset at each step. A more accurate explanation map should lead to a steeper drop in accuracy. The final score is defined as the Area Under Curve (AUC), with lower values indicating better explanations.

Insertion: This metric assesses how effectively the model's performance recovers when the most salient pixels are gradually added back to an initially blank input. At each step, a greater percentage of the explanation map's most important pixels is restored, and the model's verification accuracy is evaluated. A better explanation map will quickly restore performance. The metric is calculated as the AUC, where higher values signify better explanations.

4.4 Qualitative Results

This section presents qualitative results that illustrate the behavior of the proposed explanation method. We begin by analyzing the similarity maps generated solely by our method in a variety of scenarios, including both genuine and impostor face pairs. Then, we perform a side-by-side comparison with existing state-of-the-art explanation methods to assess the performance of the generated maps.

4.4.1 Proposed Method Results

We first analyze the explanation maps produced by our method across several representative face pairs. Visual examples are presented using the two perturbation algorithms explored: Single Inpainting (S0) and Greedy Inpainting (S1).
These examples demonstrate that the explanation maps generated by the proposed method are reliable and highlight, with high precision and detail, the regions that the FV model perceives as similar and dissimilar between the pairs. Figure 4.1 illustrates representative outcomes for genuine pairs, Figure 4.2 provides examples for impostor pairs, while Figure 4.3 depicts matching pairs that were misclassified as non-matching by the FV model. As illustrated in Figure 4.1, the nose region consistently emerges as a key contributor to the verification decision in genuine face pairs. In many cases, perturbing this area leads to the most significant drop in similarity score, suggesting its central role in identity matching. Nevertheless, the regions highlighted as important vary across examples, with other facial areas such as the forehead and mouth occasionally appearing as relevant contributors.
A clear distinction also emerges between the S0 and S1 strategies, showing that some regions that appear less relevant in isolation gain prominence when considered jointly.

In the case of non-matching pairs (Figure 4.2), the results indicate that the eye and mouth regions most frequently serve as the primary sources of dissimilarity. These areas tend to increase the similarity score when perturbed, suggesting that the model relies heavily on discrepancies in these regions when determining that two faces belong to different identities. This behavior contrasts with the genuine case, where the nose consistently plays a central role in supporting the match. This distinction highlights how the facial features driving positive identity verification differ from those used to reject a match.
We further examine cases in which genuine face pairs are incorrectly classified as impostors

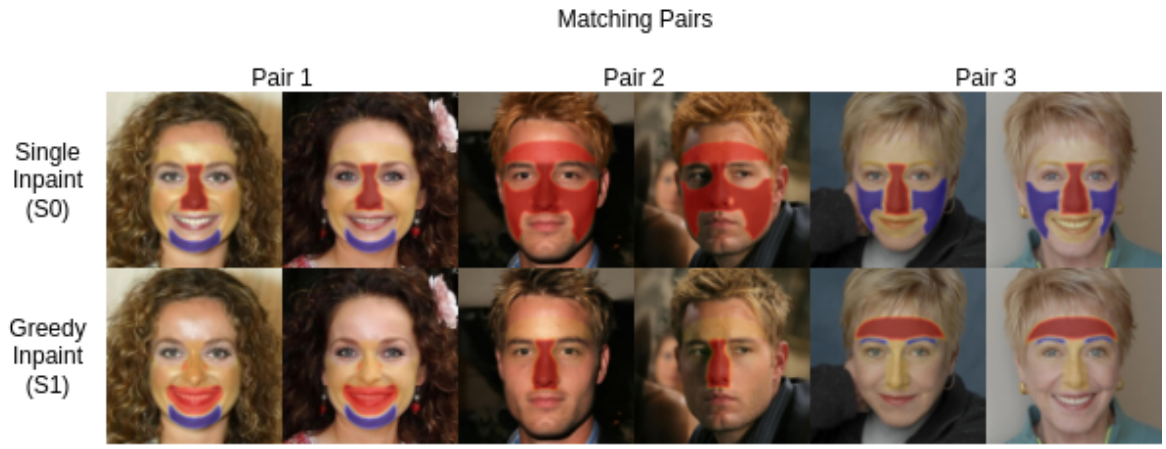


Figure 4.1: **Similarity maps generated by the proposed method for genuine face pairs.** The figure presents visual explanations produced by the Single Inpaint (S0) and Greedy Inpaint (S1) strategies for three genuine (matching) face pairs. The similarity value increases from blue to red.

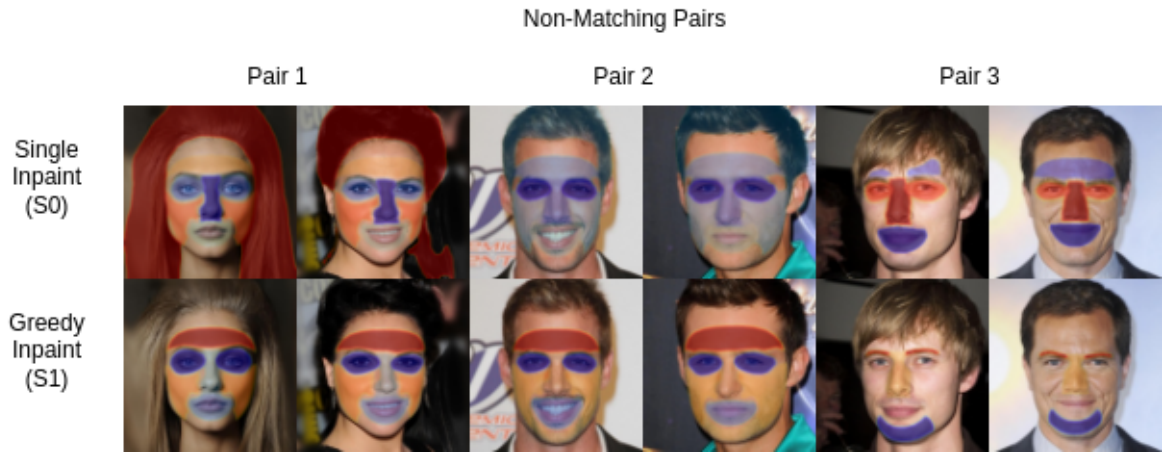


Figure 4.2: **Similarity maps generated by the proposed method for impostor face pairs.** The figure presents visual explanations produced by the Single Inpaint (S0) and Greedy Inpaint (S1) strategies for three non-matching face pairs. The similarity value increases from blue to red.

by the verification model. This analysis provides insight into which facial regions may have disproportionately influenced the model's erroneous decision. As illustrated in Figure 4.3, the similarity maps for these false negatives exhibit diverse patterns, with no single region emerging as a consistent cause across all examples.

In the first pair, the cheek region has the strongest negative influence. For the second pair, the nose (S0 algorithm) and the chin (S1 algorithm) are the most influential regions, accompanied by a smaller contribution from the eye area. In the third example, both S0 and S1 strategies highlight a mixture of similar and dissimilar regions, with the forehead consistently appearing as a dissimilar region.

4.4.2 Comparison of Explanation Maps

In this subsection, we present a comparative analysis between our proposed method, specifically the Single Inpaint (S0) strategy, and four existing explanation techniques applied to face verification. The comparison includes two face verification-specific methods, MinPlus [10]

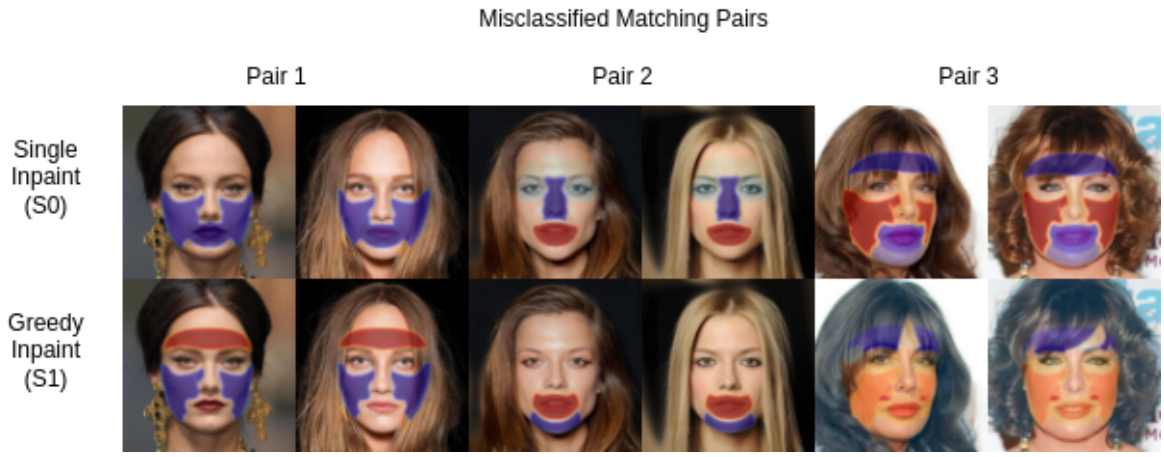


Figure 4.3: **Similarity maps for genuine face pairs incorrectly classified as impostors.** The maps were generated using the Single Inpaint (S0) and Greedy Inpaint (S1) strategies for three false negative examples. The visualization highlight which face regions may have influenced the model's misclassification. The similarity value increases from blue to red.

and xFace [5], as well as two general-purpose explainability approaches, LIME [1] and RISE [2], adapted for this task. The qualitative comparison is presented in Figure 4.4. The results demonstrate that our approach generates explanation maps that highlight the similar facial regions between image pairs with more precision. The method is also capable of reliably identifying dissimilar regions. For instance, in the first row, a genuine pair where the subject appears with glasses in one image and without in the other, the eye region is correctly marked as dissimilar.

In contrast, the outputs produced by LIME and RISE tend to be less interpretable, as their maps often highlight large areas across the face, making it difficult to draw clear conclusions about which features drive the model's decision. While MinPlus and xFace achieve better results and perform well on less complex examples (such as those in rows 3 and 4), they exhibit limitations when applied to more challenging cases (rows 1 and 2), where the localization of the face regions becomes less precise.

4.5 Quantitative Results

In this section, we present a quantitative evaluation of the proposed explanation framework. While the previous qualitative analysis provided insights into the interpretability and visual coherence of the generated similarity maps, a more objective assessment is necessary to measure the effectiveness and reliability of the explanations. To this end, we first conduct an ablation study that isolates the contribution of the inpainting component by comparing the full method against a simplified black-masking variant. This allows us to quantify the impact of realistic reconstruction on explanation quality using established perturbation-based metrics. Additionally, we perform a statistical analysis of region importance across all genuine pairs, computing the average contribution of each semantic region. This aggregation serves to validate the patterns observed in the qualitative results and provides further insight into which facial areas the verification model systematically relies upon when identifying matching identities.

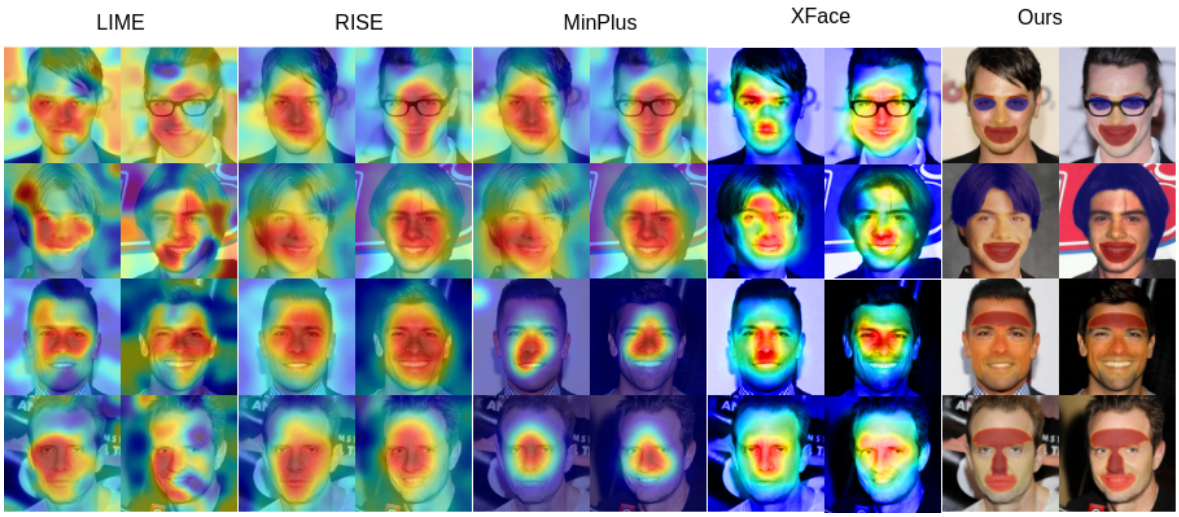


Figure 4.4: **Visual comparison of explanation maps generated by the proposed method and four existing explanation methods.** The figure compares the explanation maps produced by the proposed method (Single Inpaint(S0) algorithm) against LIME, RISE, MinPlus, and xFace. The proposed approach consistently highlights the most similar and dissimilar regions with high precision and interpretability. In contrast, LIME and RISE tend to produce less accurate maps, which are harder to interpret. While MinPlus and xFace generate more competitive results in simpler examples (rows 3 and 4), they struggle in more challenging cases (rows 1 and 2). The similarity value increases from blue to red.

4.5.1 Ablation Study

To assess the specific impact of the inpainting mechanism within our explanation framework, we perform an ablation study comparing two variants of the method. The first is the complete version, which incorporates inpainting-based perturbations to reconstruct masked facial regions. The second is a simplified baseline in which the inpainting stage is omitted and the regions are instead occluded using direct black masking. This comparison isolates the contribution of the inpainting step to the overall accuracy of the similarity maps.

We did not include quantitative comparisons with existing state-of-the-art explanation methods, as these approaches differ significantly in their formulation and output representation, making direct metric-based comparisons difficult and potentially misleading.

The evaluation employs the Deletion and Insertion metrics, which are designed to quantify how well explanation maps capture the most informative regions.

Table 4.3: Ablation study evaluating the similarity maps generated using black-mask perturbations and inpainting-based perturbations on the CelebAMask-HQ dataset, evaluated using the Deletion (↓) and Insertion (↑) metrics (%). Lower Deletion (↓) scores indicate more effective perturbations, while higher Insertion (↑) scores reflect better preservation of important regions.

Method	Deletion (%)	Insertion (%)
Black Mask	30.04	25.45
Inpainting	**21.99**	**45.90**

Table 4.3 reports the Deletion and Insertion scores for both perturbation strategies, highlighting the improved performance achieved through inpainting-based explanations.

4.5.2 Region Importance Distribution in Genuine Pairs

To complement the qualitative results shown earlier, we performed a quantitative analysis of region importance across the entire set of genuine pairs. Specifically, we computed the average contribution of each facial region, using the Single Inpaint (S0) strategy for explanation map generation. The results, presented in Figure 4.5, show a consistent pattern in which the nose emerges as the most influential region, followed by the forehead and the eyes. This finding aligns with the observations from the individual qualitative examples illustrated at Figure 4.1 and supports the claim that certain facial features are systematically more important for positive identity verification.

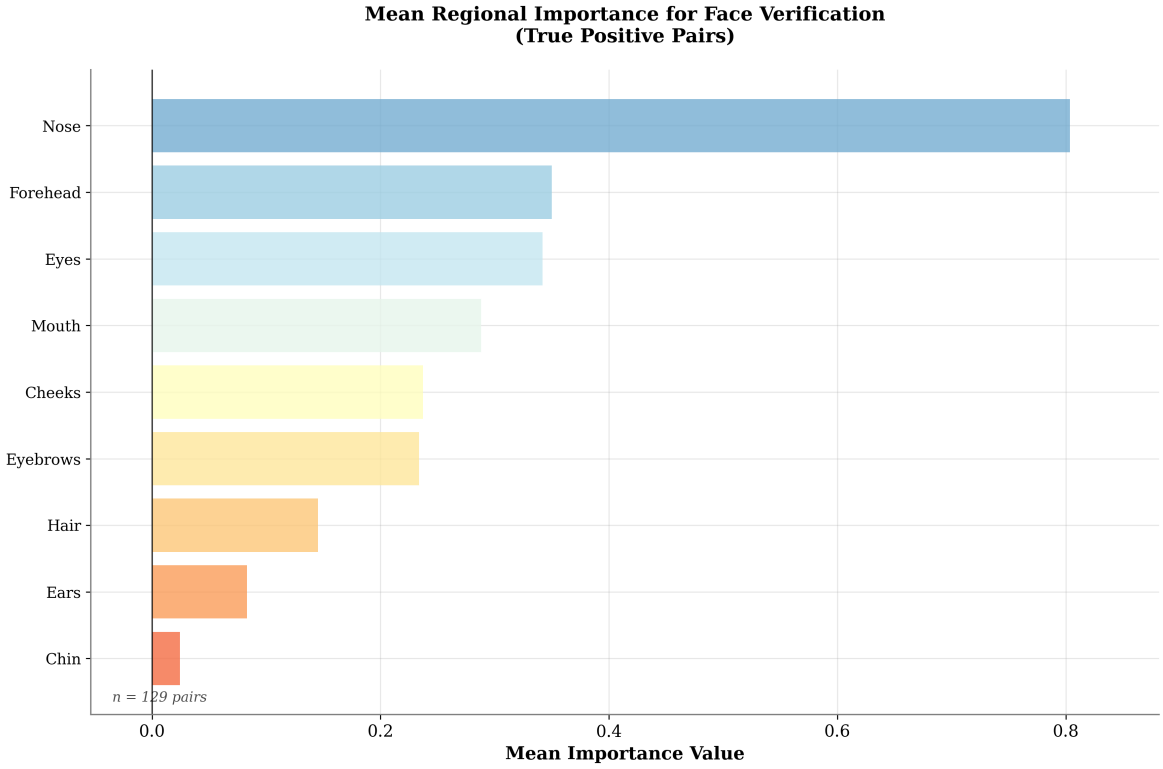


Figure 4.5: **Average importance of semantic facial regions across genuine face pairs.** The plot presents the mean contribution of each facial region to the verification decision, computed using the Single Inpaint (S0) strategy. The importance of each region is calculated based on the normalized similarity drop, using the normalization function 3.4. Higher values indicate regions that, when perturbed, caused a greater decrease in similarity score, and are therefore considered more relevant for confirming identity matches. The results reflect a consistent trend in which the nose emerge as the most influential region.

4.5.3 Method Limitations

The main limitation of the proposed framework lies in the computational overhead introduced by the inpainting step. Although the RePaint [9] model produces high-quality and realistic reconstructions of the masked face regions, it is associated with significantly long inference times. This added computational cost impacts the overall efficiency of the method and makes large-scale evaluations more time-consuming. As a result, the current implementation may not be well-suited for real-time applications.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Understanding how FV systems make decisions has become more and more important as these models are integrated into various critical applications. Despite having high accuracy, the lack of transparency and interpretability in these systems remains a significant challenge. Recent research have proposed the usage of different visual saliency maps to explain the decisions of the models, however, despite some advances and interesting approaches, they are frequently difficult to understand or prone to ambiguity.

To address this challenge, in this dissertation, we proposed a model-agnostic explanation framework that combines semantic facial segmentation with inpainting based perturbations to generate visual explanations in the form of similarity maps. This strategy enables a more transparent understanding of how face verification models process and compare facial identities.

This work builds on [10] and [8] to introduce two complementary strategies: Single Inpaint (S0), which independently assesses the contribution of each semantic region, and Greedy Inpaint (S1), which captures the cumulative effect of region combinations. These strategies are grounded in the hypothesis that realistic, semantically guided perturbations produce more faithful, interpretable and accurate explanations than those based on binary masking.

The effectiveness of the proposed approach was evaluated trough qualitative and quantitative experiments. Qualitative analysis demonstrated that the generated similarity maps are more precise, interpretable and semantically aligned than those produced by existing explanation methods. Quantitative validation was performed via an ablation study employing the Insertion and Deletion metrics, where results confirmed that the combination of semantic guidance with inpainting leads to explanations that are both accurate and faithful to the model's internal behavior.

While the method contributes meaningfully to the field of explainable face verification, it is not without limitations. The use of inpainting introduces significant computational overhead, limiting the practicality of the framework for real-time applications. Additionally, the method's reliance on external tools for face segmentation, landmark extraction and inpainting introduces a certain degree of dependency on the quality of the extracted masks and the quality of the reconstruction. This could affect robustness under extreme poses, occlusions, or image quality degradation.

Nonetheless, the contributions of this work represent an advancement toward more transparent and accountable biometric systems. By bridging semantic masks with inpainting, the proposed method opens new directions for face verification interpretability that balance precision with reliability.

Finally, this work reinforces the importance of designing explanation techniques that respect the semantic structure of the input and the logic of the model. As face recognition technologies continue to proliferate, methods like the one proposed in this work present a foundation for building more interpretable and trustworthy AI systems.

5.2 Future Work

While the results of this work demonstrate the effectiveness and interpretability of the proposed method, several directions remain open for future exploration and improvement.
One of the most immediate areas for extension involves addressing the computational demands of the inpainting process. The use of diffusion-based models such as Repaint [9], while effective in producing high-quality reconstructions, introduces significant inference time. The framework would greatly benefit from the integration of faster inpainting models that also maintain strong visual quality, enabling more efficient similarity map generation. However, during the development of this work, we were not aware of any such models offering a favorable balance between speed and quality with publicly available implementations.
The current framework treats inpainting as an entirely generative and unconstrained process. Future research could explore controllable inpainting or face editing models that allow for guided reconstruction of masked regions. The ability to direct how a region is filled could provide deeper insights into model behavior.
Another promising direction involves combining the proposed perturbation-based strategy with Vision Language Models (VLMs) to support natural-language explanations. In particular, VLMs could be used to assess and rank the similarity maps produced by different explanation methods, including the one proposed in this work, across qualitative criteria such as precision, clarity, and interpretability. By providing the VLM with pairs of similarity maps (e.g., from our method and from state-of-the-art techniques like xFace [5] or MinPlus [10]), along with descriptive prompts, it would be possible to obtain feedback and ratings (e.g., on a 1–5 scale with justification) for each explanation.
Finally, formal user studies could be conducted to assess the interpretability of the generated explanations. While the work relied on qualitative evaluations, engaging with people that are not familiarized with the domain could provide an unbiased insight into how explanation maps are perceived and understood. The goal would be to collect human judgments on the quality of similarity maps produced by the proposed method compared to state-of-the-art approaches. Participants would be presented with visual explanations generated by different methods for the same input pairs and asked to choose which method they prefer for each predefined criteria. Rather than providing numerical scores, participants would make pairwise selections for each criterion.

Bibliography

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. ix, xv, xvi, 9, 12, 13, 16, 22, 25, 26, 33, 37

[2] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *British Machine Vision Conference*, 2018. ix, xv, xvi, 9, 12, 13, 16, 18, 20, 25, 26, 33, 37

[3] 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 International Conference on Computer Vision*, 2017, pp. 618–626. ix, xv, 9, 10, 11

[4] D. Mery and B. Morris, "On black-box explanation for face verification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1826–1835. xv, xvi, 9, 13, 14, 16, 17, 18, 23, 25, 26, 29

[5] M. Knoche, T. Teepe, S. Hörmann, and G. Rigoll, "Explainable model-agnostic similarity and confidence in face verification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 711–718. xv, xvi, 9, 17, 25, 26, 33, 37, 42

[6] N. Bousnina, J. Ascenso, P. L. Correia, and F. Pereira, "A rise-based explainability method for genuine and impostor face verification," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2023, pp. 1–6. xvi, 9, 20

[7] Y. Lu, Z. Xu, and T. Ebrahimi, "Towards visual saliency explanations of face verification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4726–4735. xvi, 9, 21, 33, 34

[8] M. Doh, C. M. Rodrigues, N. Boutry, L. Najman, M. Mancas, and H. Bersini, "Bridging human concepts and computer vision for explainable face verification," *arXiv preprint arXiv:2403.08789*, 2024. xvi, 9, 22, 23, 27, 41

[9] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 11 461–11 471. xvi, 24, 33, 40, 42

[10] D. Mery, "True black-box explanation in facial analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 1596–1605. xvi, 9, 17, 18, 25, 26, 33, 36, 41, 42

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699. 1, 6, 33

[12] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 896–16 905. 1, 7

[13] M. Kim, A. Tran, and T. Hassner, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4321–4330. 1, 8

[14] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018. 1

[15] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015. 1

[16] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (xai) approaches for transparency and accountability in financial decision-making," *Available at SSRN 4640316*, 2023. 1

[17] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC medical informatics and decision making*, vol. 20, no. 1, p. 310, 2020. 1

[18] S. Lim, "Judicial decision-making and explainable artificial intelligence: a reckoning from first principles," *Singapore Academy of Law Journal*, vol. 33, pp. 280–314, 2021. 1

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2726–2735. 6

[20] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274. 6

[21] 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 (CVPR)*, 2016, pp. 770–778. 7

[22] 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. 9, 11

[23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NeurIPS, 2017, pp. 4768–4777. 22

[24] J. Castro, D. Gómez, and J. Tejada, “Polynomial calculation of the Shapley value based on sampling,” *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009. 22

[25] J.-C. de Borda, “Mémoire sur les élections au scrutin,” *Histoire de l’Académie Royale des Sciences*, vol. 102, pp. 657–665, 1781. 22

[26] C. Lugaresi, T. Sorensen, D. Hafner, G. Clemente, A. Markley, H. Prafullchandra, and S. C. Basu, “Mediapipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019. 28

[27] V. Yakhyokhuja, “face-parsing,” https://github.com/yakhyo/face-parsing, 2024, gitHub repository. 28

[28] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 28

[29] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 28, 33