UNIVERSIDADE
BEIRA INTERIOR

Detection of Overflowing Waste and Litter near Trash Bins

Diogo José dos Santos Paulo

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

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

Covilhã, junho, 2025

Declaração de Integridade

Eu, Diogo José dos Santos Paulo, que abaixo assino, estudante com o número de inscrição M13364 de Engenharia Informática da Faculdade de Engenharias, 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 11/06/2025

Assinado por: **Diogo José dos Santos Paulo**
Num. de Identificação: 30680923

Acknowledgements

This dissertation is dedicated to everyone who has been involved throughout my journey. My sincerest and deepest appreciation goes to all who have been part of this journey.

Above all, I would like to express my gratitude to my supervisor, Professor João Neves, and my co-supervisor, Professor Hugo Proença, for their teaching, guidance, support, patience, and expertise throughout this work. I thank you for your guidance, particularly from Professor João Neves. Your ever-present support, more often the support of a genuinely trusted friend rather than just your position of supervisor, meant everything to me.

I would like to thank my friends and colleagues at Socialab for having built such an amazing working environment. Thank you especially to Ana Dias for the support and friendship throughout this work.

I am grateful to my girlfriend, Vitória Dinis. My inspiration and strength came from your patience and your boundless love, both through long hours and trying times. My deepest gratitude to you for your support, even when it had to come hardest, and for the sacrifices you made so that this work would become possible.

I am also grateful to my family. The constant support and loving guidance I received throughout my scholarly pursuits have been the cornerstone of my achievements. You have been there all along and have supported me since the start.

Finally, to everyone who has supported me in big and small ways, thank you. This accomplishment is as much yours as it is mine.

Resumo

Atualmente, a gestão eficaz de resíduos domésticos apresenta muitos desafios, incluindo situações de transbordo de resíduos em contentores e a acumulação de resíduos à volta dos mesmos. Esta dissertação oferece um sistema inteligente que deteta e segmenta de forma automática os resíduos em situações de transbordo, utilizando técnicas de visão computacional. Para isto, foi criado um conjunto de dados personalizado com mais de 7.200 imagens anotadas, capturadas em ambientes urbanos reais, por câmaras *fisheye* montadas em veículos em movimento. Modelos de segmentação de imagem de estado da arte, como o YOLOv11 e YOLACT, foram usados e as métricas utilizadas para avaliar estes foram a precisão, o recall e a área sob a curva precisão-recall. O método proposto para superar os métodos tradicionais e limitados ao uso de imagens RGB é uma abordagem que integra a estimativa de mapas de profundidade e normais a partir de uma única imagem RGB. A interação de informações geométricas com o modelo oferece uma fusão de informações que permite aumentar o desempenho do modelo, distinguindo o *foreground* do *background*. Este método chega a conseguir ganhos de até 47% em relação a métodos usados como referência, o que demonstra o seu potencial para aplicações em tempo real de monitorização de resíduos urbanos.

Palavras-chave

Gestão de Resíduos Urbanos, Visão Computacional, Segmentação de Objetos, Estimativa de Profundidade, Estimativa de Superfícies Normais

Resumo alargado

A gestão eficiente de resíduos urbanos constitui um dos principais desafios enfrentados pelas cidades, impactando diretamente a saúde pública, o meio ambiente e a qualidade de vida nas zonas urbanas. A crescente urbanização, aliada a práticas de deposição inadequadas, resulta frequentemente em situações de contentores a transbordar e acumulação de resíduos fora dos mesmos, designados como lixo parasita. A deteção automática destas ocorrências é de extrema importância para otimizar os sistemas de recolha e melhorar a gestão de resíduos urbanos.

Esta dissertação propõe um sistema inteligente baseado em técnicas de visão computacional para detetar e segmentar automaticamente áreas de transbordo de resíduos acumulados fora dos contentores. Para tal, é explorado um conjunto de modelos de segmentação de estado da arte, considerando cenários reais em contexto urbano.

Foi criado um conjunto de dados personalizado, com mais de 7.200 imagens anotadas, recolhidas com câmaras *fisheye* montadas em veículos em movimento, simulando uma recolha contínua em ambiente urbano. As imagens foram cuidadosamente selecionadas e anotadas para representar situações reais de transbordo e lixo parasita, capturadas sob diferentes condições de iluminação, perspetiva e zona urbana. Este *dataset* constitui uma valiosa contribuição científica e prática.

Modelos como o YOLOv11, SOLOv2, Mask R-CNN, YOLACT, Mask2Former e LISA foram treinados e avaliados com base em métricas como a precisão, recall, *mean Average Precision* (mAP) e *generalized Intersection over Union* (gIoU), de forma a comparar o seu desempenho e adequação para aplicações em tempo real.

Para superar as limitações observadas nas abordagens RGB tradicionais, esta dissertação propõe uma abordagem inovadora que combina a imagem original com mapas de profundidade e superfícies normais gerados a partir de uma única imagem RGB. Esta fusão de informação geométrica, obtida por modelos do tipo zero-shot como o Metric3Dv2, foi incorporada em modelos de estado da arte ajustados para aceitar *inputs* "multicanal", ou seja, com mais de três canais. Ao enriquecer o *input* com características espaciais, o sistema revelou-se mais robusto na diferenciação entre resíduos urbanos e elementos do *background* visualmente semelhantes, permitindo uma segmentação mais precisa, mesmo em condições adversas.

A abordagem proposta registou melhorias significativas, com ganhos de até 47% na precisão de segmentação em relação a métodos de referência, com uma redução de falsos positivos e falsos negativos. Complementar a isto, estudos de remoção confirmaram o impacto positivo da inclusão de informação geométrica no desempenho final do sistema, destacando as superfícies normais como o contributo mais relevante para a melhoria da segmentação.

Esta dissertação contribui com um novo conjunto de dados anotado, uma avaliação rigorosa de modelos de segmentação de estado da arte e uma arquitetura inovadora de deteção baseada em características geométricas. O trabalho demonstra a viabilidade de aplicar este tipo de sistemas em contextos urbanos reais, possibilitando a monitorização automática em veículos de recolha urbanos. O sistema desenvolvido tem o potencial de melhorar as estratégias de recolha de resíduos, contribuindo para cidades mais limpas, inteligentes e susten-

táveis.

Abstract

Urban waste management faces growing challenges due to population growth and improper disposal practices, which often lead to overflowing bins and the presence of parasitic waste in public areas. This dissertation proposes an intelligent system for the automatic detection and segmentation of overflowing waste using advanced computer vision techniques. A custom dataset comprising over 7,200 annotated images was collected using fisheye cameras mounted on moving vehicles in real urban environments. Several state-of-the-art segmentation models, such as YOLOv11, YOLACT, and others, were evaluated in terms of precision, recall, and mean average precision (area under the precision-recall curve). To overcome the limitations of traditional RGB-based methods, a novel approach is introduced that combines RGB images with estimated depth and normal surface maps generated from a single RGB image. The interaction of geometric information with the model provides a fusion of information that enhances the model's ability to distinguish between real waste and background clutter. The proposed system significantly improves segmentation accuracy, achieving up to 47% mAP gains over baseline methods. The results highlight the potential of this approach for real-time urban waste monitoring and contribute a novel dataset and methodology to the field.

Keywords

Bin Overflow Segmentation, Litter Segmentation, Instance Segmentation, Computer Vision, Depth Estimation, Surface Normal Estimation

Contents

List of Figures

List of Tables

Acronyms List

AP	Average Precision
CSTM	Canonical Space Transformation Module
FCN	Fully Connected Network
gIoU	generalized Intersection over Union
IoU	Intersection over Union
LLM	Large Language Model
mAP	mean Average Precision
MLLMs	Multimodal Large Language Models
NMS	Non-Maximum-Suppression
RoI	Region of Interest
RoIAlign	Region of Interest Align
RoIPool	Region of Interest Pooling
RPN	Region Proposal Network

Chapter 1

Introduction

1.1 Motivation and Scope

Around the world, urban areas are becoming increasingly concerned about the difficulties associated with trash management. For instance, properly managing and disposing of trash has a direct impact on environmental sustainability and public health. The problem becomes worse because of the increasing quantity of waste produced by cities due to the world's population expansion and fast urbanization. One of the effects of this is the growing number of littered urban areas and overflowing trash cans, which can negatively affect city residents' quality of life and worsen the environment.

Urban waste management is a complex task due to its dynamic nature. That is, trash accumulates at varying rates depending on the location, time, and population density, and traditional waste collection methods use fixed schedules, which often result in inefficiencies. For instance, some areas accumulate excessive waste before pickup, while others do not. These inefficiencies lead to wasted resources and a larger environmental footprint due to excessive fuel consumption in waste collection logistics.

This dissertation is conducted in collaboration with EVOX Technologies, a company devoted to developing technological solutions for the waste market. This dissertation aims to address important problems in contemporary urban waste management, such as litter detection and bin overflow. Waste management systems will benefit greatly from the capacity to precisely detect and segment locations where trash cans are overflowing and litter is piling up. A successful approach to identifying such occurrences would help cities spend resources more effectively by identifying high-density garbage zones and enabling more focused pickup schedules.

To achieve this, this work leverages advanced computer vision techniques to detect and segment overflowing bins and scattered litter in urban environments, ensuring cleaner cities, reducing unnecessary waste collection trips, and minimizing environmental impact.

1.2 Objectives

The main goal of this dissertation is to create a structured research plan supported by state-of-the-art methodologies with the final goal of designing and implementing an intelligent system capable of detecting and segmenting both bin overflow and parasitic waste, commonly referred to as litter, around waste collection bins in urban areas to monitor waste accumulation dynamically.

To fulfill this objective, this dissertation presents a comprehensive study of current state-of-the-art waste detection, segmentation, and waste area estimation techniques. This includes

a detailed analysis of current methods, from conventional computer vision algorithms to the most recent developments in deep learning methods, with the goal of creating a method that not only matches but also surpasses current approaches in terms of accuracy, speed, and reliability.
However, developing such a system may present some challenges. One primary concern is the variability in the appearance of bin overflow, which can take on different shapes, forms, and textures. This variability may lead to misclassification, where the system incorrectly detects overflow in areas without existing or fails to identify actual instances of waste accumulation.

1.3 Main Contribuitions

This dissertation contributes to advancing the field of automated waste monitoring, providing valuable tools for cities to improve waste collection operations, reduce litter, and promote environmental sustainability. The expected key contributions of this work are as follows:

- **A dissertation entitled "Detection of Overflowing Waste and Litter near Trash Bins", written in English;**
- **Novel Dataset** - An annotated dataset containing images of waste bins and litter in urban environments, along with detailed segmentation masks for both the litter and overflow areas. This dataset will be a significant asset for the wider scientific community, facilitating future investigations in garbage detection and urban management;
- **Research Paper** - A comprehensive research paper detailing the methodologies, experimental design, results, and findings of the study;
- **Novel Approach for Waste Management** - An innovative segmentation approach that has been optimized for the purpose of detecting and controlling overflowing waste bins in actual urban settings. It will provide a significant improvement over existing methods.

1.4 Work Planification

To achieve the objectives of this dissertation, the following tasks have been proposed in the work plan and are presented in the Gaant Chart 1.1.

1.4.1 T1 — Literature's Revision

Conduct a thorough review of existing segmentation models used for waste detection and other related computer vision applications. This will help identify the most promising techniques and prepare performance benchmarks.

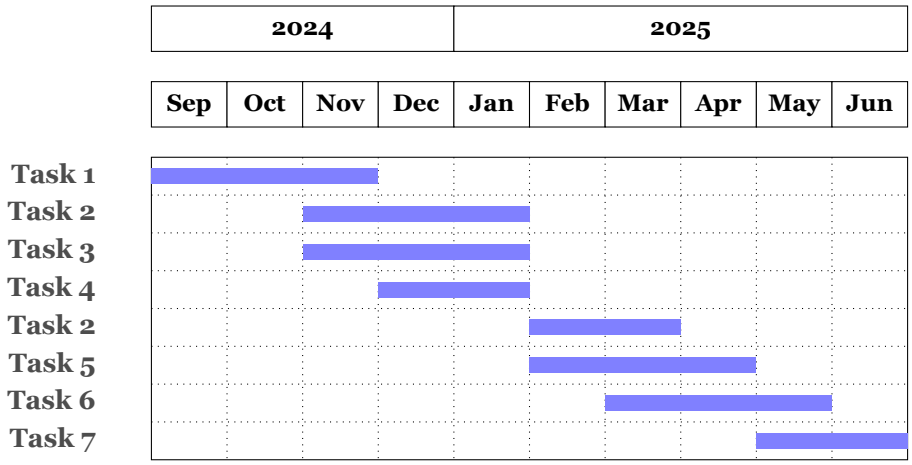


Figure 1.1: Gantt diagram regarding the development of the tasks.

1.4.2 T2 — Annotation of the Dataset

The dataset is annotated with precise segmentation masks that delineate both the litter and overflow regions. Additionally, the bounding boxes around each trash bin are defined and labeled. These annotations serve as the ground truth for model training and evaluation.

1.4.3 T3 — Evaluation of State-of-Art Models

Evaluate and analyze the current state-of-the-art segmentation models in terms of their applicability, accuracy, and speed in the task of waste detection and segmentation in urban environments. This includes reviewing their strengths, limitations, and performance in the collected dataset. Insights gained from this evaluation will inform the design of the proposed model.

1.4.4 T4 — Write the First Part of the Dissertation

The first part of this thesis will provide a comprehensive review of relevant literature, document the dataset annotation process, and present preliminary results from benchmarking existing segmentation models. This will include an analysis of initial findings and the identification of key performance gaps.

1.4.5 T5 — Implementation

Based on the findings from the literature review, a custom segmentation model will be developed or modified and trained to identify and segment overflowing bins and litter in images captured in urban environments. The focus will be on achieving high accuracy in detecting objects under varying conditions, such as different lighting, weather, and spatial distributions of waste.

1.4.6 T6 — Evaluation

The proposed model will be rigorously tested on a diverse dataset, which includes images

from various urban environments. The performance of the developed model will be compared to existing solutions to assess improvements in accuracy and robustness using performance metrics such as mAP and generalized Intersection over Union (gIoU).

1.4.7 T7 — Write the Second Part of the Dissertation

The second part of this dissertation will present a detailed analysis of the results obtained from the model evaluation. This will include a discussion on the implications of these findings, potential limitations of the current approach, and suggestions for future research directions to further enhance the model's performance and applicability.

1.5 Thesis Organization

The present document is structured as follows:

- Chapter 1 - Introduction - this chapter presents the goals and related tasks for this dissertation, along with its scope and motivation;
- Chapter 2 - Related Work - this chapter reviews the fundamental concepts and techniques applied in state-of-the-art methods related to waste detection and segmentation. Moreover, it presents how current methods can be applied to determine the size of objects based on metrology computer vision techniques;
- Chapter 3 - Proposed Method - this chapter introduces our proposed method that seeks to enhance segmentation accuracy and robustness, particularly in the context of overflow detection.
- Chapter 4 - Dataset - this chapter describes the dataset used to train and evaluate the state-of-the-art models, as well as to conduct the experiments detailed in the next chapter.
- Chapter 5 - Experiments and Results - this chapter outlines the conducted experiments and an explanation of the performance evaluation process. A thorough discussion of the outcomes of these experiments is also included;
- Chapter 6 - Conclusions and Future Work - this chapter describes the main conclusions drawn from this dissertation, as well as the future work.

Chapter 2

Related Work

This chapter reviews the fundamental concepts and techniques applied in state-of-the-art methods related to waste detection and segmentation, including approaches for identifying waste bin overflow and parasitic waste on streets. Furthermore, the possibility of using computer vision approaches to estimate the size of objects using metrology is investigated.

2.1 Segmentation Models for Object Detection

Segmentation models for object detection are methods used in the field of computer vision that are responsible for detecting, localizing, and segmenting objects within images or videos. There are two main types of segmentation models: one-stage and two-stage models.
Unlike traditional object detection algorithms, which only provide bounding boxes, the segmentation models are capable of pixel-wise classification, which makes it possible to accurately locate and segment objects. These models are useful for tasks that require knowing or obtaining the object's shape and size, such as medical imaging, autonomous driving, and waste management systems, where the boundaries of objects (like trash or containers) must be accurately identified.

2.1.1 Two-Stage Models

Currently, there are several cutting-edge segmentation architectures, however one of the earliest to be developed was Mask R-CNN [1]. This model was proposed by Kaiming He *et al.* [1], and it is an extension of the Faster R-CNN [14] object detection framework since it adds a branch for predicting segmentation masks for each Region of Interest (RoI). There are two phases in Faster R-CNN [14]. The initial step, a Region Proposal Network (RPN), proposes suitable object bounding boxes. In the second stage, which is essentially Fast R-CNN [15], each candidate box's features are extracted using Region of Interest Pooling (RoIPool), and bounding-box regression and classification are carried out. RoIPool is an operation used in detection and segmentation-based applications to extract a tiny feature map from each RoI. In order to perform the actual scaling, the region proposal is divided into portions of equal size. The largest value in each segment is then copied to the output buffer. RoIPool is essentially max pooling on a box-based discrete grid.
For quicker inference, the features that are utilized by both stages can be shared. Mask R-CNN [1] uses the same two-step process, starting with the same RPN stage. In the second step, this model generates a binary mask for every RoI and forecasts the class and box offset.

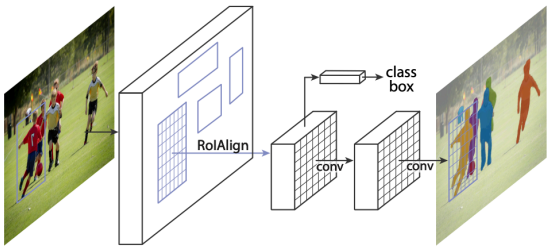


Figure 2.1: The Mask R-CNN [1] framework for instance segmentation.

This two-stage model, as depicted in Figure 2.1, first detects objects and then predicts masks, allowing it to handle both object detection and segmentation.
Kaiming He *et al.* [1], also introduce Region of Interest Align (RoIAlign) which improves the RoIPool operation by removing the quantization of region boundaries and bins, which aligns extracted features more accurately with the input. Instead of rounding coordinates, it uses continuous values and bilinear interpolation to sample features at precise locations within each bin. This approach eliminates misalignments introduced by RoIPool, significantly enhancing performance in pixel-accurate tasks like mask prediction.
The strength of Mask R-CNN's [1] is its ability to produce segmentation masks with high quality and its robustness in handling multiple objects within a scene. However, the model tends to be slower compared to some real-time alternatives, which makes it less suitable for applications that require low-latency processing, such as real-time image analysis in waste management systems.

2.1.2 One-Stage Models

2.1.2.1 YOLACT: You Only Look at Coefficients

Regarding one-stage models, Daniel Bolya *et al.* [2] proposed YOLACT to address the need for faster segmentation models. It uses a fully convolutional approach that completes instance segmentation in a single shot, making it faster and better suited for real-time applications. This methodology accomplishes this by first creating prototype masks, which are subsequently linearly merged with instance-specific coefficients to produce the final segmentation, as shown in Figure 2.2.
Prototype Generation. YOLACT's [2] backbone is responsible for extracting feature maps from the input image, which are then passed through a prototype head called a protonet. This prototype head generates k prototype masks, denoted by:

$$P = \{P_1, P2, \ldots, P_k\} \in \mathbb{R}^{H \times W \times k}, \quad (2.1)$$

where H and W represent the spatial dimensions of each prototype mask, and k represents the number of prototypes. The authors attached the protonet to a backbone feature layer and implemented it as a Fully Connected Network (FCN) with a final layer that includes k channels (one for each prototype). The difference between this formulation and ordinary semantic segmentation is that there is no explicit loss on the prototypes. Instead, all oversight

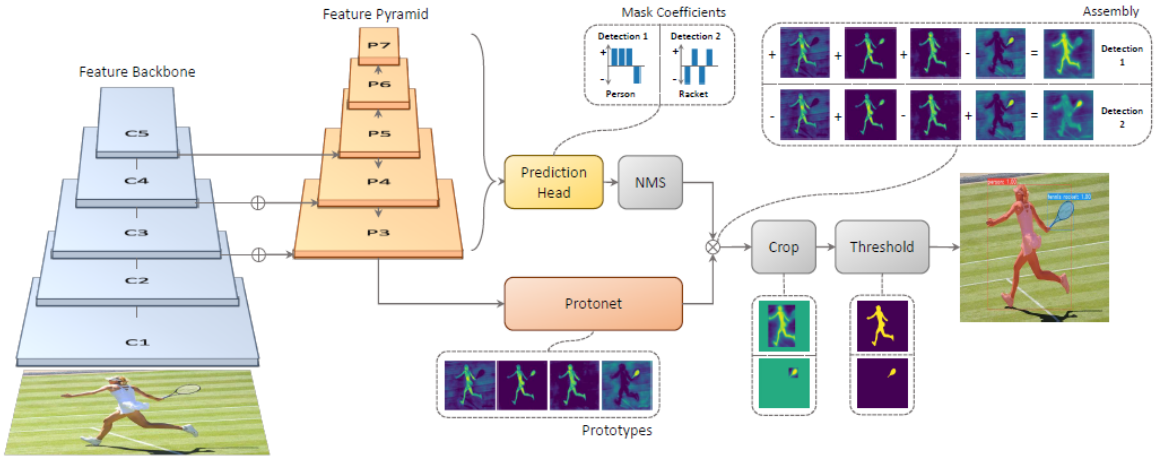


Figure 2.2: YOLACT [2] architecture.

for these prototypes is derived from the final mask loss during assembly.

Instance-Specific Mask Coefficients. The prediction heads of anchor-based object detectors feature two branches: one for predicting four bounding box regressors and the other for predicting c class confidence scores. The authors only add a third branch in parallel to estimate k mask coefficients, one for each prototype, in order to predict mask coefficients. As a result, they generate $4 + c + k$ coefficients per anchor rather than $4 + c$, as illustrated in Figure 2.3. According to the authors, the ability to remove prototypes from the finished mask is crucial for nonlinearity. Therefore, they apply $tanh$ to the k mask coefficients, resulting in more stable outputs than those with no nonlinearity. This decision is essential since neither mask could be constructed without the ability to subtract.
YOLACT's [2] detection branch also calculates mask coefficients $c \in \mathbb{R}^k$ for each detected instance. Each coefficient vector is unique to the instance and serves as a set of weights for combining the prototypes to produce the final mask.

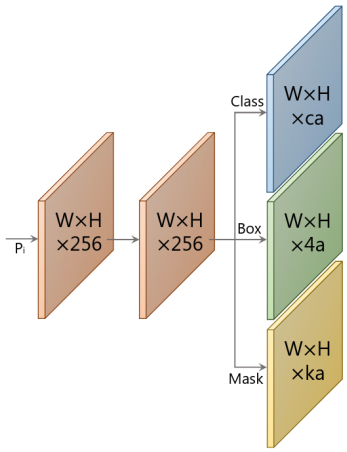


Figure 2.3: YOLACT's [2] head architecture.

Final Mask Formation. The final mask M for each instance is produced by linearly com-

bining the prototypes P using the instance's coefficients vectors C:

$$M_i = \sigma\left(P \cdot C^T\right), \quad (2.2)$$

where σ represents a sigmoid function applied element-wise to generate a binary mask.
Despite being far faster than Mask R-CNN [1], YOLACT [2] may have inferior mask quality, especially in complex scenarios with overlapping or occluded objects due to its "localization failure" problem. For instance, the network may not be able to localize each object in its own prototype if there are too many objects in one area of a scene. In these situations, for certain objects in the group, the network will produce an output that is more akin to a foreground mask than an instance segmentation. Nevertheless, YOLACT [2] may offer a reasonable mix between speed and accuracy for use cases like litter segmentation in real-time systems or bin overflow detection.

2.1.2.2 YOLO: You Only Look Once

YOLO models [16] are another real-time option due to their speed and real-time capabilities. At the moment, YOLOv11 [17] is the best and the latest model in its family. Efficiency is the YOLO [16] family's strongest point, which makes it ideal for embedded and mobile systems, such as waste monitoring systems that run on smartphones. Because of its good feature extraction and multi-scale processing, YOLOv11 [17], which was initially created for object detection, may also be used for segmentation tasks.

C3k2 block-enhanced backbone. YOLOv11 [17] innovates by using C3k2 which is a block-enhanced backbone. To enable feature extraction at multiple stages of the backbone, this model employs smaller 3×3 kernels, which enhance computational efficiency while preserving the model's ability to capture critical image features. The C3K2 block enhances information flow by dividing the feature map and applying a sequence of 3×3 convolutions, which are faster and more resource-efficient than larger kernels. By processing smaller feature map segments and merging them after several convolutions, the C3K2 block delivers improved feature representation with fewer parameters compared to the other blocks used in previous versions.

C2PSA: Cross Stage Partial with Spatial Attention. YOLOv11 [17] also introduces the C2PSA module-enhanced spatial attention, which enhances spatial attention by incorporating attention mechanisms that help the model prioritize important regions within an image, such as smaller or partially occluded objects. This is achieved by emphasizing spatial relevance in the feature maps.
The C2PSA block, as illustrated in Figure 2.4, utilizes two Partial Spatial Attention (PSA) modules that operate on separate branches of the feature map, which are then concatenated. This approach ensures that the model efficiently focuses on spatial information while maintaining a balance between computational cost and detection accuracy. By applying spatial attention to the extracted features, the C2PSA block improves the model's ability to selec-

tively focus on regions of interest, enabling YOLOv11 [17] to outperform previous versions in tasks requiring precise object detection.

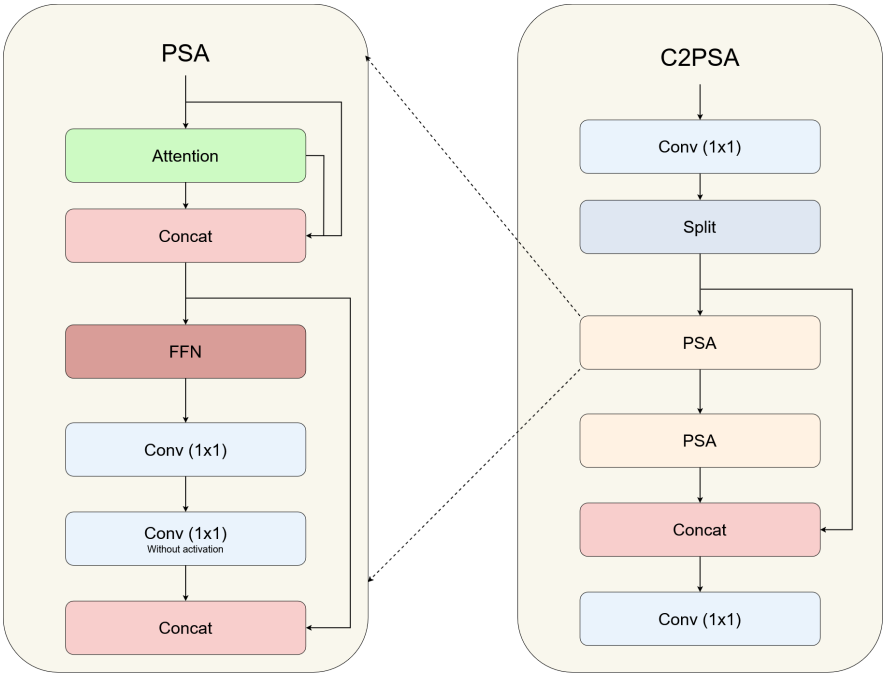


Figure 2.4: Cross Stage Partial with Spatial Attention (C2PSA) module.

YOLOv11 [17] is appropriate for applications that call for both object localization and fine-grained segmentation since the head uses these features to produce pixel-level segmentation masks.
Even though their segmentation accuracy is marginally lower than other models, i.e., Mask R-CNN [1], YOLO-based segmentation models may perform well in applications where computing efficiency is a top concern, such as identifying overflowing bins on moving vehicles.

2.1.2.3 SOLO: Segmenting Objects by Location

SOLO, introduced by Xinlong Wang *et al.*[3], marked a significant advancement in instance segmentation by eliminating the need for region proposals, a key component of traditional models like Mask R-CNN[1]. Instead, SOLO [3] directly segments objects by treating the problem as a pixel classification task, where each pixel is assigned to a specific instance. This approach greatly simplified the architecture while maintaining high-quality segmentation. A simplified version of the SOLO [3] architecture is depicted in Figure 2.5.

Semantic Category. For each grid, SOLO [3] forecasts the C-dimensional output, where C is the number of classes, to show the semantic class probabilities. These probabilities change depending on the grid cell. The output space will be $S \times S \times C$ if the input image is split into $S \times S$ grids, as seen in Figure 2.5 (top). In this design, every cell in the $S \times S$ grid is assumed to correspond to a single instance and, hence, to a single semantic category. Each object instance's class probability is shown in the C-dimensional output during inference.

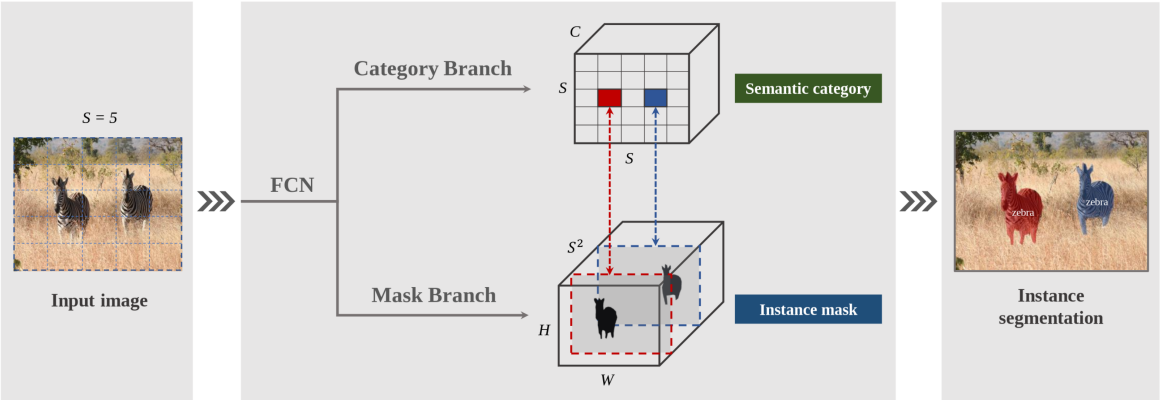


Figure 2.5: SOLO [3] architecture. Instance segmentation is reformulated by the authors as two subtasks: instance mask generation and category prediction. An input image is separated into $S \times S$ uniform grids. A grid with $S = 5$ is used to show this. A grid cell predicts the semantic category (top) and instance masks (bottom) if an object's center falls inside it. In order to simplify, the feature pyramid network (FPN) is not depicted in this figure.

Instance Mask. Every positive grid cell will produce its related instance mask concurrently with the semantic category prediction. If we split an input image I into $S \times S$ grids, the total number of predicted masks will be no more than S^2. Xinlong Wang *et al.*[3] explicitly encodes these masks in the third dimension (channel) of a 3D output tensor. In particular, the instance mask output will be $H_I \times W_I \times S^2$ dimension. The k^{th} channel will be responsible for segment instance at grid (i, j), where $k = i \cdot S + j$ (where i and j are zero-based). This is accomplished by establishing a one-to-one relationship between the class-agnostic mask and the semantic category (Figure 2.5). Inspired by the "CoordConv" [18] operator, Xinlong Wang *et al.*[3] directly feed normalized pixel coordinates to the networks in order to generate a model position sensitive because the segmentation masks are conditioned on the grid cells and need to be separated by distinct feature channels. A CoordConv [18] layer extends the standard convolutional layer by adding extra channels to the input representation. These additional channels encode hard-coded coordinates, typically including one channel for the x-coordinate and another for the y-coordinate. CoordConv [18] is particularly beneficial for tasks involving coordinate transformations, where traditional convolutional layers often struggle.
In particular, the authors produce a tensor with pixel coordinates that are normalized to [−1, 1] with the same spatial dimension as the input. The subsequent layers receive this tensor once it has been concatenated to the input features. Thus, if the original feature tensor is of size $H \times W \times D$, the size of the new tensor becomes $H \times W \times (D + 2)$, in which the last two channels are (x, y) pixel coordinates.

Forming Instance Segmentation. In SOLO [3], the category prediction and the corresponding mask are inherently associated by their reference grid cell, i.e., $k = i \cdot S + j$. Based on this, it is possible to directly form each grid's final instance segmentation result. The raw instance segmentation results are generated by gathering all grid results. Finally, Non-Maximum-Suppression (NMS) is used to obtain the final instance segmentation results. No other post-processing operations are needed.

2.1.2.4 SOLOv2

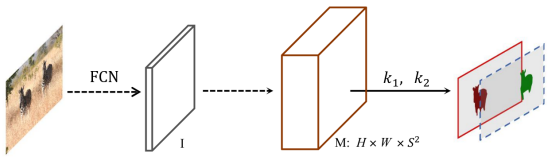
Building on the foundation of SOLO [3], SOLOv2 [4] introduces further innovations to improve both speed and accuracy.

SOLOv2 [4] addresses several critical limitations of SOLO[3], significantly improving its efficiency and accuracy in object detection and segmentation. SOLO [3] suffered from three major bottlenecks: inefficient mask representation, inaccurate mask predictions, and slow mask NMS. In SOLO [3], the mask representation was inefficient, as it required large, separate output tensors for each feature pyramid level, leading to high memory and computational costs. Additionally, the lack of shared learning across levels resulted in suboptimal training efficiency. This is improved by SOLOv2 [4], which lowers the processing overhead by offering a more compact and shared mask prediction method. It also tackles the issue of inaccurate mask predictions by leveraging higher-quality, multi-scale features and providing finer segmentation details without excessively increasing computational costs. Furthermore, SOLOv2 [4] optimizes the mask NMS process, making it faster and more efficient, thus reducing the overall runtime.

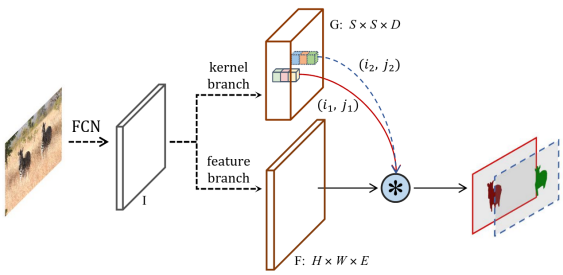
As illustrated in Figure 2.6, the instance masks of SOLOv2 [4] (subfigure b) are generated dynamically by convolving a location-specific kernel G with a feature map F, producing the final mask M. For each grid cell at (i, j), Xinlong Wang *et al.*[3] first obtain the mask kernel $G_{i,j} \in R^D$. Then $G_{i,j}$ is convolved with F to get the instance mask. There will be, at most, S^2 masks for each prediction level. Finally, it is used a custom matrix NMS proposed by the authors to get the final instance segmentation results. Thus, the final mask is calculated by:

$$M_{i,j} = G_{i,j} * F, \quad (2.3)$$

where $*$ represents the convolution operation.



(a) SOLO



(b) SOLOv2

Figure 2.6: SOLOv2 [4] in contrast with SOLO [3] architecture. These models divide the image in a grid $(S \times S)$. SOLOv2 [4] convolves the feature map with a specific kernel location.

Compared to Mask R-CNN [1] and YOLACT [2], SOLOv2 [4] provides a more structured way of segmenting objects by grouping pixels dynamically into different instances. This results in faster inference times while maintaining competitive mask quality. SOLOv2's [4] unique instance grouping method may make it highly effective for segmenting objects in cluttered environments, such as urban scenes with litter or overflowing trash bins.

2.1.3 Transformer-Based Models

Cheng *et al.* [5] introduced Mask2Former, a universal architecture for image segmentation. This model is designed to perform panoptic, instance, and semantic segmentation tasks using a single, unified framework. Its core innovation lies in employing a Transformer decoder equipped with a masked attention mechanism to predict a set of class labels and corresponding binary masks directly. The architecture of this model is depicted in Figure 2.7.

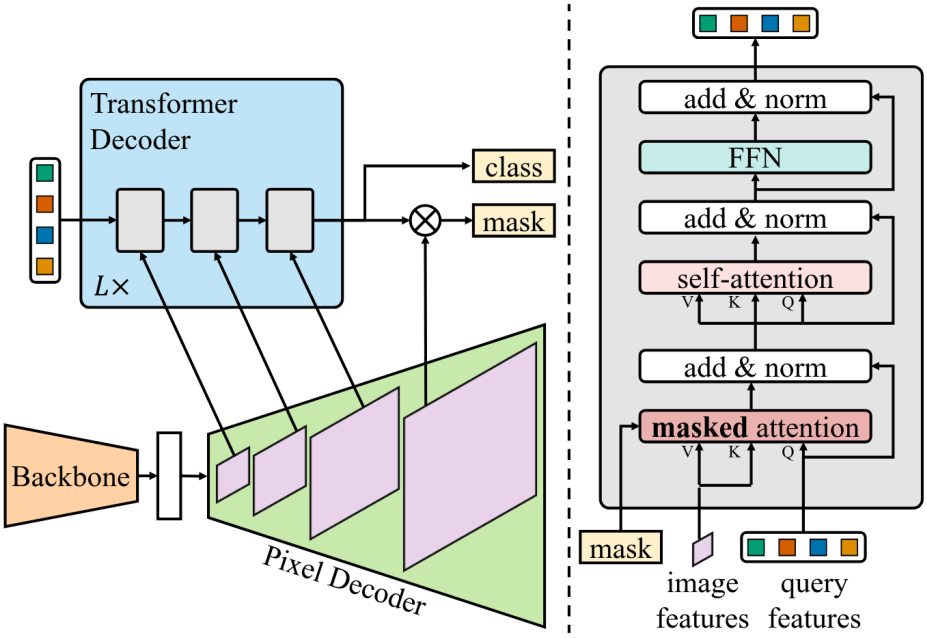


Figure 2.7: Overview of the Mask2Former [5] architecture, which includes a backbone for feature extraction, a pixel decoder for multi-scale feature generation, and a Transformer decoder with masked attention that predicts N (class, mask) pairs.

In Mask2Former [5], an input image is first processed by a backbone network (e.g., Swin Transformer [19]) to extract hierarchical features. These features are then fed into a pixel decoder module, which generates multi-scale feature maps. The core of the model is its transformer decoder, which takes these multi-scale features and a set of N learnable object queries as input. A key aspect of Mask2Former [5] is its utilization of masked attention within the Transformer decoder layers. Each object query attends only to the localized foreground region predicted for it by the preceding decoder layer, rather than the entire image. This enables more efficient and focused feature learning, resulting in improved mask quality. Finally, the transformer decoder outputs N pairs, each consisting of a class prediction and a corresponding binary segmentation mask.

2.1.4 Prompt Models

Kirillov *et al.* [6] proposed the ”Segment Anything Model” (SAM) with the goal of developing a universal segmentation model that can identify any object in an image without the need for task-specific training. By using large datasets to train on a diverse range of objects, SAM [6] is able to generalize effectively to previously undiscovered categories and be used over the zero-shot learning paradigm. It is particularly versatile, as it can divide ”anything” into segments based on spatial prompts, such as predefined points, or semantic inputs, such as textual descriptions. The architecture of this model is depicted in Figure 2.8.

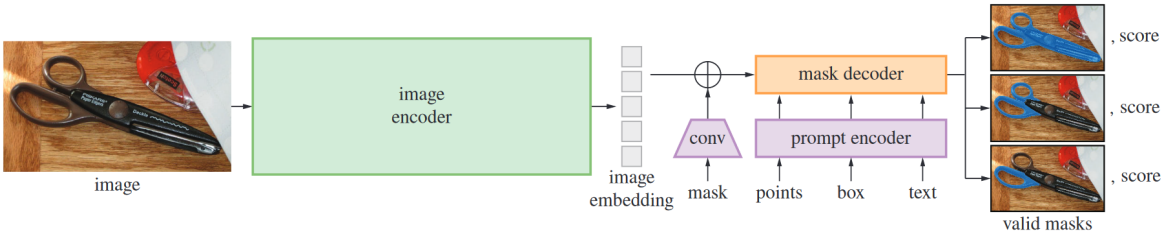


Figure 2.8: The (SAM) [6] has three components: the image encoder that extracts global features, the prompt encoder that processes user input (points, boxes, or text), and the mask decoder that combines both to generate the segmentation mask.

Additionally, SAM2 was proposed by Ravi *et al.* [7] aiming to naturally extend SAM [6] to the video domain by analyzing individual video frames and use a memory attention module to focus on the target object’s prior memories. Nevertheless, this model acts like SAM [6] when applied to images because the memory is empty. Its architecture is depicted in Figure 2.9.

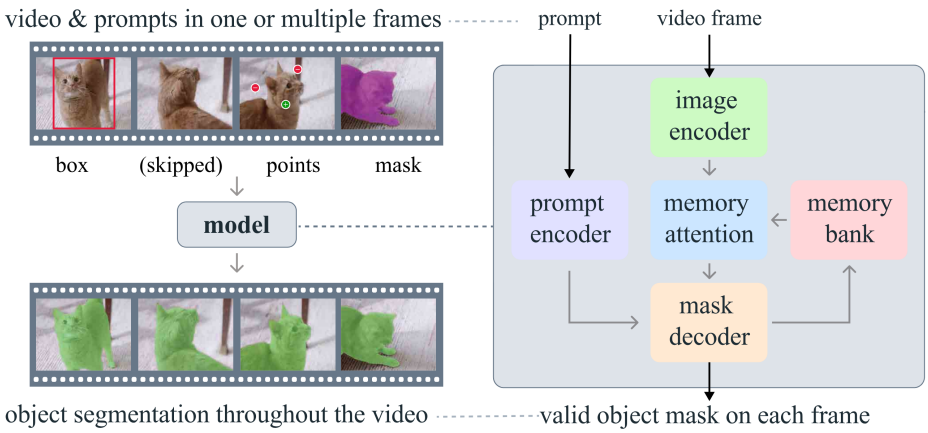


Figure 2.9: Segment Anything Model 2 architecture [7].

For tasks like bin overflow detection, where new types of objects or scenes may emerge, SAM [6] or even SAM2 [7] could offer the flexibility to adapt without needing extensive re-training. However, its general-purpose nature might lead to suboptimal performance when compared to task-specific models like SOLOv2 [4] in well-defined environments.

2.1.5 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) combine text with other types of data, like images, audio, or video, and are made to handle and produce content across multiple modalities.

One of the recent advances made in this field was the creation of LISA, a model that can process both text and images (input multimodal), proposed by Lai *et al.* [8]. Its architecture is presented in Figure 2.10.

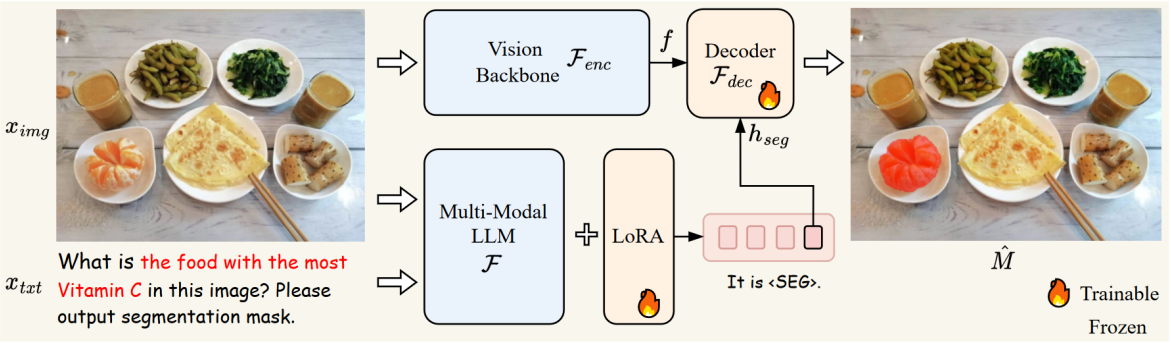


Figure 2.10: LISA [8] architecture.

Specifically, Lai *et al.* [8] start by expanding the original Large Language Model (LLM) vocabulary with a new token, i.e., $< SEG >$, which represents the request for the segmentation output. Given a text instruction x_{txt} along with the input image x_{img}, the multimodal LLM F is fed, which then outputs a text response $\hat{y}_{txt}$. It can be formulated as:

$$\hat{y}_{txt} = F(x_{img}, x_{txt}). \quad (2.4)$$

When the LLM intends to generate a binary segmentation mask, the output $\hat{y}_{txt}$ would include a $< SEG >$ token. Then, the last-layer embedding $\tilde{h}_{seg}$ corresponding to the $< SEG >$ token is extracted from the LLM. This embedding is then processed by a multi-layer perceptron (MLP) projection layer γ, which consists of two linear transformations with an intermediate ReLU activation and a dropout layer. The resulting transformed embedding, denoted as h_{seg}, is the final segmentation representation used in the subsequent mask generation process. Concurrently, the vision backbone F_{enc} extracts the dense visual features f from the visual input x_{img}. Finally, h_{seg} and f are fed to the decoder F_{dec} to produce the final segmentation mask $\hat{M}$. The process can be formulated as:

$$h_{seg} = \gamma(\tilde{h}_{seg}), \quad (2.5)$$
$$f = F_{enc}(x_{img}), \quad (2.6)$$
$$\hat{M} = F_{dec}(h_{seg}, f). \quad (2.7)$$

2.2 Overflow and Litter Segmentation

Kulkarni *et al.* [20] developed a system that uses a hybrid approach combining transfer learning with Faster R-CNN for waste object detection. Their system categorizes waste into classes like glass, plastic, paper, metal, trash, and cardboard. The main contribution is the use of hybrid transfer learning to speed up training on a specific dataset while leveraging pre-trained models for object detection. Based on the knapsack problem, they have generated collaged images containing multiple waste objects for training purposes. This is an interesting approach for applications where labeled datasets are limited and is a way to increase the size of the dataset. However, while the knapsack-based collage generation enhances training data, it might not entirely reflect the natural variations and occlusions encountered in real-world waste disposal environments and does not contribute significantly to our segmentation task.

Oğuz *et al.* [9] focused on a smart waste management solution, where they used deep learning models to estimate the fullness of garbage containers. Their key contribution is developing an automated monitoring system that can help optimize waste collection schedules based on real-time data. By predicting container fullness, cities can reduce unnecessary trips and improve resource allocation. However, the system's accuracy can be influenced by variations in container shape, placement, and lighting conditions, which may require further improvements in model generalization across diverse environments. Moreover, as depicted in Figure 2.11, the authors only classify whether there is overflow and don't detect or segment it on images.

Figure 2.11: Samples from the Oğuz *et al.* [9] dataset, depicted together with their class labels.

Balmik *et al.* [21] proposed a vision-based system that combines the Single Shot Detector (SSD) [22] with the MobileNetV2 [23] architecture to detect and classify litter in real-time. Their methodology involved training the model on a custom dataset. Initially, direct input into the model yielded a low mAP, but they enhanced performance through several prepro-

cessing and augmentation techniques. They applied methods such as histogram equalization and Gaussian blur filtering, which improved model accuracy. They also employed Canny edge detection as a feature extraction technique, significantly boosting the mAP to 84%. This preprocessing step sharpened the contours of litter, making it easier for the model to recognize objects against urban backgrounds. However, the authors acknowledged that their approach struggles with low-light images and the detection of small, partially obscured objects.

De Carolis *et al.* [24] focused on real-time detection of litter in continuous video streams using an improved YOLO [16] (You Only Look Once) architecture. Their system yields satisfactory results, presenting a mAP@50 of 59.57%. The authors collected a dataset using Google Images Download [25] and then presented an approach to exclude duplicate images by using the cosine similarity between the features of pairs of images. They extracted the features using the ORB (Oriented FAST and Rotated BRIEF) [26] algorithm. The real-time aspect of YOLO [16] makes it suitable for video analysis, allowing the system to process a stream without breaking it into individual frames manually. However, like other detection algorithms, YOLO TrashNet [24] faces limitations in complex urban settings, such as detecting litter that is partially occluded or obscured by other objects, leading to reduced performance in cluttered environments.

Proença *et al.* [10] introduced TACO (Trash Annotations in Context), one of the most comprehensive datasets for litter detection. TACO's key contribution is its wide range of annotated litter images taken in various urban and natural settings, making it an invaluable resource for training deep learning models. Some of the images of the dataset are depicted in Figure 2.12. Proença *et al.* [10] presented some results using Mask R-CNN [1]; however, despite the dataset's comprehensiveness, detection results on tiny objects (e.g., cigarettes) lead to poor performance and significantly affect the overall AP. Additionally, while TACO is highly useful for training, it does not directly address the performance of models in real-time detection scenarios, which is a critical aspect of practical urban litter detection systems.

Barra *et al.* [11] introduced LOTS (Litter On The Sand dataset for litter segmentation), a novel dataset for segmenting small litter particles on sandy beaches, addressing the challenge of distinguishing artificial litter from natural elements like shells, stones, and algae. This dataset comprises images collected from four beaches with varying sand textures and colors, captured in both controlled and outdoor environments to simulate real-world scenarios. Some of the images of the dataset are depicted in Figure 2.13. Each image is accompanied by its ground-truth mask corresponding to the litter while excluding natural elements, ensuring precise annotations for segmentation tasks.

While LOTS [11] contributes significantly by focusing on small-scale litter in beach environments, it has some limitations. The dataset is geographically restricted to four beaches in a specific region, which might limit its generalizability to other areas. Although images captured in controlled conditions ensure consistency, they might not fully replicate the complexities of other environments with varying lighting and occlusion. Despite these challenges,

Figure 2.12: Cropped annotated images from TACO [10] dataset.

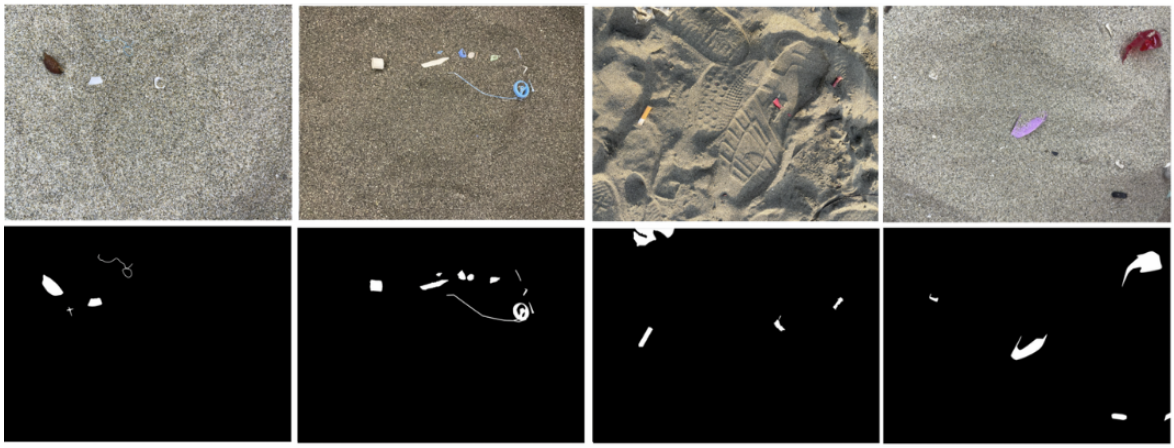


Figure 2.13: LOTS [11] dataset. This dataset comprises 3572 images with their corresponding segmentation masks.

LOTS [11] represents an important step toward developing more effective tools for environmental monitoring and beach litter management, providing a foundation for future advancements in litter detection and segmentation.

2.3 Metrology in Computer Vision

2.3.1 Single View Metrology in the Wild

Zhu *et al.* [12] created a single-view metrology technique that uses a single monocular image to estimate the absolute 3D heights of objects and camera parameters. Their method uses data-driven priors, which are statistical patterns learned from large datasets of commonly observed objects, such as cars and people, to calibrate camera geometry and integrates

weakly supervised constraints using deep learning. The method's limitations include its assumption of a single dominant ground plane, which might not be accurate in urban settings with multiple surfaces at varying heights, and its bias towards appearance, which reduces its effectiveness in complex scenes with occlusions or irregular ground profiles. For instance, in order to recover 3D parameters from 2D annotations in the images, Zhu *et al.* [12] leverage several assumptions to simplify the process, they assume the world is composed of a dominant ground plane on which all objects are situated and a camera that observes the scene. Their approach uses a perspective camera model, which is illustrated in Figure 2.14, and is parameterized by the camera angles (yaw ϕ, pitch θ and roll ψ), focal length f and camera height h_{cam} to the ground.

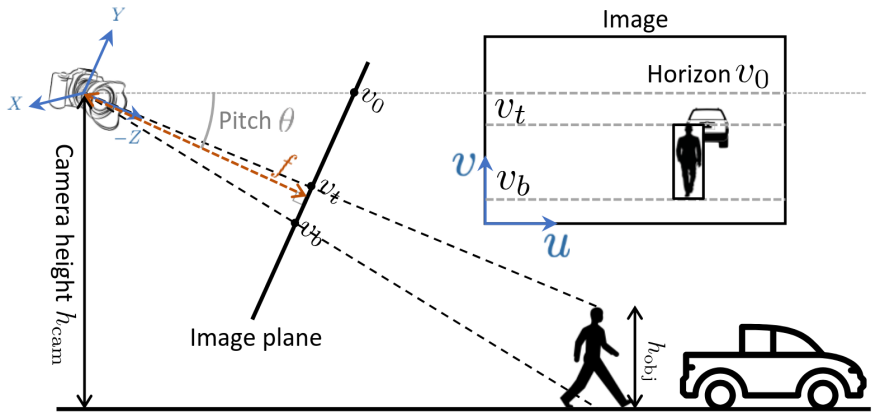


Figure 2.14: Measurements in image space (top) and the scene's camera model (bottom) [12].

Moreover, the camera calibration module estimates pitch θ and field of view h_θ from input picture I, whereas the object estimation modules estimate bounding boxes, object heights h_obj, and person points. An initial estimate h_{cam} is obtained by feeding the camera height estimation module with the estimated horizon v_0, bounding boxes, and object heights. After that, the reprojection module calculates the bounding box reprojection errors L_{v_t}, which are then input into the refinement network together with other variables to predict updates on the camera and object heights. The final estimate is the result of multiple refinement layers. The architecture of this method is depicted in Figure 2.15.

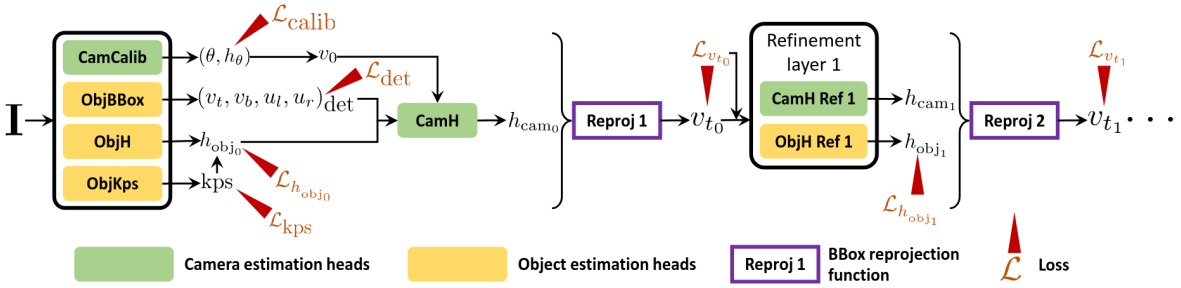


Figure 2.15: An outline of the Zhu *et al.* [12] approach.

2.3.2 Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image

Furthermore, a new method for metric 3D reconstruction from single images was presented by Yin *et al.* [13]. This method was specifically designed for zero-shot scenarios, in which the model generalizes without having been exposed to the target data beforehand. The authors presented a canonical camera space transformation module that recalibrates images into a shared reference space. The method achieves robust zero-shot generalization by integrating this transformation module with large-scale training data, surpassing earlier methods in standard benchmarks like NYU-Depth V2 [27] and KITTI [28]. In terms of mathematics, the canonical camera space transformation resolves the scaling and alignment problems that occur when images are taken with different cameras by transforming individual camera viewpoints into a single coordinate system. Then, without requiring particular camera calibration data, this method can be used immediately in zero-shot testing. The architecture of this method is presented in Figure 2.16.

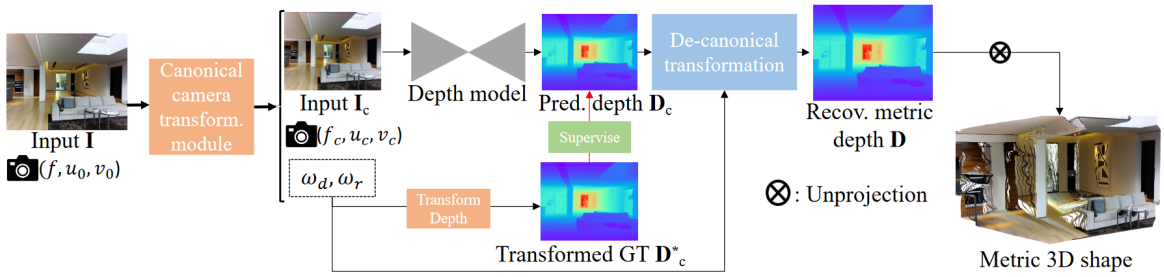


Figure 2.16: Pipeline of Yin *et al.* approach [13]. An input image I is transformed into a canonical space using CSTM. The transformed image I_c is fed into a standard depth estimation model to produce the predicted metric depth D_c in the canonical space. During training, D_c is supervised by a GT depth D_c, which is also transformed into the canonical space. In inference, after producing the metric depth D_c in the canonical space, we perform a de-canonical transformation to convert it back to the space of the original input I.

2.3.3 Metric3Dv2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation

More recently, Hu *et al.* [29] introduced Metric3Dv2, presenting it as an evolution and a complementary enhancement to the original Metric3D framework. While retaining the core strengths of Metric3D [13], particularly its zero-shot metric depth estimation capabilities through the CSTM, Metric3Dv2 [29] expands its scope to become a more versatile geometric foundation model by jointly predicting both metric depth and surface normals. This multi-task approach aims to provide a richer geometric understanding of the scene from a single image. The presented approach in Metric3Dv2 [29] is illustrated in Figure 2.17.
Similar to its predecessor, Metric3Dv2 [29] begins by transforming the input image I into a canonical camera space using the CSTM, resulting in I_c. This step is essential for handling diverse camera intrinsics and enabling zero-shot generalization, and remains a foundational element inherited from Metric3D [13]. However, a key difference and advancement in Metric3Dv2 [29] is the subsequent processing of I_c, which, instead of a dedicated depth estima-

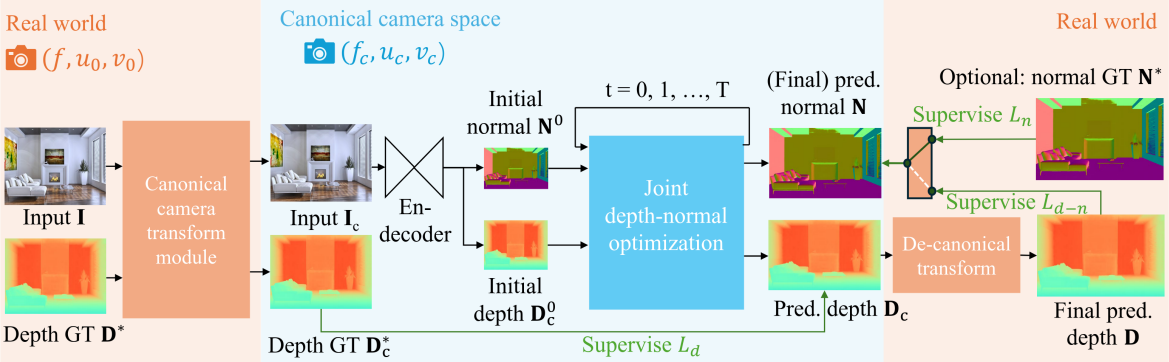


Figure 2.17: Pipeline. The authors initially use CSTM to convert an input image I, to the canonical space. The estimated metric depth D_c in the canonical space and metric-agnostic surface normal N are obtained by feeding the converted image I_c into a conventional depth-normal estimation model. During training, D_c is supervised by a ground truth depth D_c^*, which is also transformed into the canonical space. The authors apply a de-canonical transformation in inference to return the metric depth D_c to the space of the original input I after it has been produced in the canonical space. The canonical space transformation and de-canonical transformation are executed using camera intrinsics. The predicted normal N is supervised by depth-normal consistency via the recovered metric depth D as well as the ground truth normal N^*, if available.

tion model, I_c is fed into a joint depth-normal estimation module. This module is designed to simultaneously output two distinct geometric properties, the metric depth D_c, which is the predicted depth map in the canonical space, and the metric-agnostic surface normal N, which is the predicted surface normal map that captures the orientation of surfaces independently of their absolute scale.

The processing and supervision of the metric depth D_c follow the principles of Metric3D [13]. During training, D_c is supervised by a ground truth depth D_c^*, which is also transformed into the canonical space. For inference, the predicted D_c goes through a de-canonical transformation, utilizing camera intrinsics, to yield the final metric depth D in the coordinate system of the original input I. Metric3Dv2 [29] employs a dual supervision strategy for the normals, or uses a direct supervision with ground truth normals, which is when ground truth surface normals N^* are available for a training sample, and they are used to supervise the predicted normals N, directly providing a strong learning signal for accurate normal estimation or uses a depth-normal consistency supervision which introduces a depth-normal consistency loss. This loss leverages the recovered metric depth D (obtained after the de-canonical transformation of D_c) to enforce geometric consistency between the predicted normals N and the estimated depth map D. Specifically, surface normals can be analytically derived from a dense depth map. By comparing the predicted normals N with normals derived from the model's own depth output D, this consistency loss enables the model to learn robust normal estimation, even from datasets where explicit normal ground truth may be scarce or entirely unavailable.

Therefore, Metric3Dv2 [29] extends Metric3D [13] by not only adopting its successful canonical space transformation for robust metric depth but also by integrating surface normal estimation. This is achieved through a joint model architecture and a refined supervision scheme that combines direct normal supervision with an innovative depth-normal consistency mechanism. This makes Metric3Dv2 [29] a more comprehensive geometric foundation model,

capable of extracting 3D information from a single image in a zero-shot learning paradigm.

2.4 Conclusion

Despite significant advancements in waste detection and segmentation, existing methods fail to directly address the problem of segmenting overflowing waste in trash bins, especially under challenging conditions such as low-light environments and moving camera scenarios. Traditional approaches that classify bin fullness, like those proposed by Oğuz *et al.* [9], rely on global image classification rather than local image classification, making them unsuitable for identifying overflow scenarios. Similarly, detection-based approaches, such as YOLO-based methods, perform well in real-time applications but lack the precision to segment waste that spills beyond the container accurately.

Instance segmentation models like Mask R-CNN [1] may provide high-quality masks but suffer from computational inefficiency, making them less suitable for real-time applications. Faster alternatives such as YOLACT [2] offer a balance between speed and segmentation quality but struggle with localization failures when objects are occluded or densely packed. SOLOv2 [4], with its structured instance grouping approach, seems promising in handling cluttered scenes but may still require fine-tuning to adapt to the irregular shapes and textures of waste materials.

Recent advancements in prompt-based and multimodal models, such as SAM [6] and LISA [8], demonstrate zero-shot generalization capabilities. However, their dependence on general training datasets often lacks the task-specific refinement necessary for accurately segmenting the overflowing waste. While these methods offer flexibility to adapt without extensive training, their performance is inconsistent when dealing with unique cases like waste scenarios.

Regarding the estimation of the area of overflowing waste, state-of-the-art metrology techniques have demonstrated promising results in estimating object dimensions from single images. For instance, approaches like Metric3D [13] offer a robust zero-shot generalization, while methods based on perspective transformations and monocular depth estimation provide a foundation for measuring waste spillover in real-world settings. However, these techniques require careful calibration and validation to ensure reliability when applied to irregularly shaped and unstructured objects such as litter.

Given these challenges, this dissertation aims to surpass the existing gaps by developing a robust segmentation system customized explicitly to bin overflow detection and parasitic waste on streets.

Chapter 3

Proposed Method

3.1 Introduction

As discussed in the previous chapter, existing segmentation approaches for waste detection often struggle when applied to the specific challenge of identifying overflowing waste in real-world scenarios. Factors such as occlusions, low illumination, irregular object shapes, and cluttered urban scenes contribute to frequent misclassifications and reduced segmentation quality.
To mitigate these challenges, this chapter introduces a novel approach that seeks to enhance segmentation accuracy and robustness, particularly in the context of overflow detection. By incorporating additional features beyond color information, the proposed method aims to improve the model's spatial understanding and reduce false positives, particularly in cases where visual ambiguity or background noise may lead to misinterpretation.
This chapter details our approach, specifying how we contribute to a more precise segmentation of waste overflow.

3.2 Our Approach

The proposed method, illustrated in Figure 3.1, aims to address the limitations of the state-of-the-art, particularly the issues related to false positives (incorrectly classified overflowing waste) and false negatives (missed detections of actual overflow).

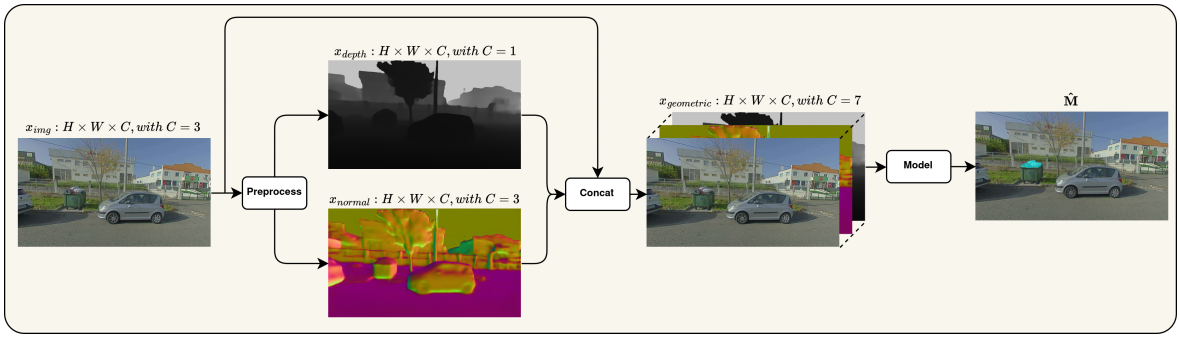


Figure 3.1: Overview of the proposed method. The input RGB image x_{img} is processed using a geometry estimation module, which produces both a depth map x_{depth} and a surface normal map x_{normal}. These are then concatenated with the original image to form an enriched input tensor $x_{geometric} \in \mathbb{R}^{H \times W \times C}$, where $C = 7$. This new representation is then fed to adapted segmentation models capable of handling multi-channel input, which output the predicted mask denoted as $\hat{M}$ for overflowing waste.

We propose to incorporate depth and surface normal information into the detection pipeline to mitigate the limitations of existing RGB-based segmentation approaches. Our method leverages a zero-shot learning framework, specifically the Metric3Dv2 model proposed by

Hu *et al.* [29], to estimate geometric properties from a single RGB image. As discussed in Chapter 2, this model enables the generation of both depth maps x_{depth} and surface normal maps x_{normal} without requiring intrinsic camera calibration or multi-view supervision.
Once these geometric representations are generated, we concatenate them with the original RGB image $x_{img} \in \mathbb{R}^{H \times W \times 3}$ to form an enriched multi-channel input:

$$x_{input} = concat(x_{img}, x_{normal}, x_{depth}) \in \mathbb{R}^{H \times W \times 7}. \quad (3.1)$$

This new input encapsulates both color and geometric features, enabling the model to leverage spatial structure and local surface orientation in addition to color and texture.
To accommodate this new input format, we modified state-of-the-art segmentation architectures to accept seven-channel data and learn effectively from this fused representation.
Depth information provides additional data that might enhance the model's ability to differentiate between objects based on their spatial positioning and three-dimensional structure. Traditional RGB-based approaches utilize only color and texture features. At the same time, depth data enables the system to estimate the relative distances of objects from the camera, helping to disambiguate overlapping elements and improve segmentation accuracy. In scenarios where an object might be mistaken for overflowing waste due to its texture or color similarity (e.g., background elements like walls, parked cars, or trees), depth information may help to verify whether the object is linked with the container or positioned in the background.
In addition to depth, surface normal information offers complementary geometric features that describe the local orientation of surfaces in the scene. This enables the model to better understand the shape of waste materials, particularly in environments where RGB and depth alone may be insufficient to distinguish overlapping or deformable objects.
We expect to reduce false positive cases using depth and surface normal information, enabling a more precise spatial understanding of overflowing objects. Additionally, depth-aware and geometry-aware detection may enhance the system's robustness against false negatives, even in cases of partial occlusion or complex backgrounds.

3.3 Conclusion

Through this chapter, we introduced a geometry-aware segmentation framework designed to enhance the detection of overflowing waste in complex urban environments. Recognizing the limitations of RGB-based methods, we proposed the integration of depth and surface normal information into the detection pipeline. These geometric representations were concatenated with the original RGB image to form a unified multi-channel input, thereby enriching the model's perception with spatial and structural features.
This fusion of modalities enables the model to better differentiate between actual waste overflow and visually similar background objects.
That said, the proposed method is expected to reduce both false positives and false negatives, improving the segmentation accuracy under real-world constraints. More generally,

this approach paves the way for the development of more context-aware and geometrically informed vision systems.

Chapter 4

Dataset

4.1 Introduction

This chapter describes the dataset used to train and evaluate the state-of-the-art models, as well as to conduct the experiments detailed in Chapter 5. The dataset is custom-acquired and processed for the purposes of this dissertation.

4.2 Dataset Description

The dataset employed in this work is collected in collaboration with Evox Technologies and comprises images of trash bins captured under various real-world conditions, including instances of overflowing waste and litter in the surrounding areas. Figure 4.1 illustrates some examples of the images included in the dataset.

Figure 4.1: Representative images from the dataset. Our dataset comprises a total of 7229 annotated images with overflowing and parasitic waste scenarios.

The images in this dataset are taken with a fisheye camera, which is a camera with a wide-angle lens that can capture a wider field of view than other cameras. This configuration facilitates the inclusion of surrounding litter and context in the scene, which is essential for the task of overflow detection.

4.2.1 Data processing

To create a more suitable input for the models and relevance to the task, the raw dataset is processed through a preprocessing pipeline. The original data consists of video recordings captured using two synchronized fisheye cameras mounted on a vehicle positioned on the left side and the other on the right. The total recording time is 2 hours and 7 minutes at 30 frames per second.
Subsequently, the videos are manually reviewed to isolate segments depicting overflowing waste. From this process, 8 minutes and 44 seconds of relevant videos are extracted from the right camera, and 3 minutes and 19 seconds from the left camera, totaling 12 minutes and 3 seconds of annotated overflow scenarios.
To reduce redundancy caused by high frame rates, the selected segments are downsampled to 10 frames per second, resulting in a final dataset comprising 7,230 images. These images are then used in both the training and evaluation phases of the experimental pipeline.

4.3 Dataset Analysis

To understand the characteristics of the annotated data, an analysis of object sizes and per-image object counts is conducted. Figure 4.2 provides the object area distribution and the number of objects of the annotated instances.

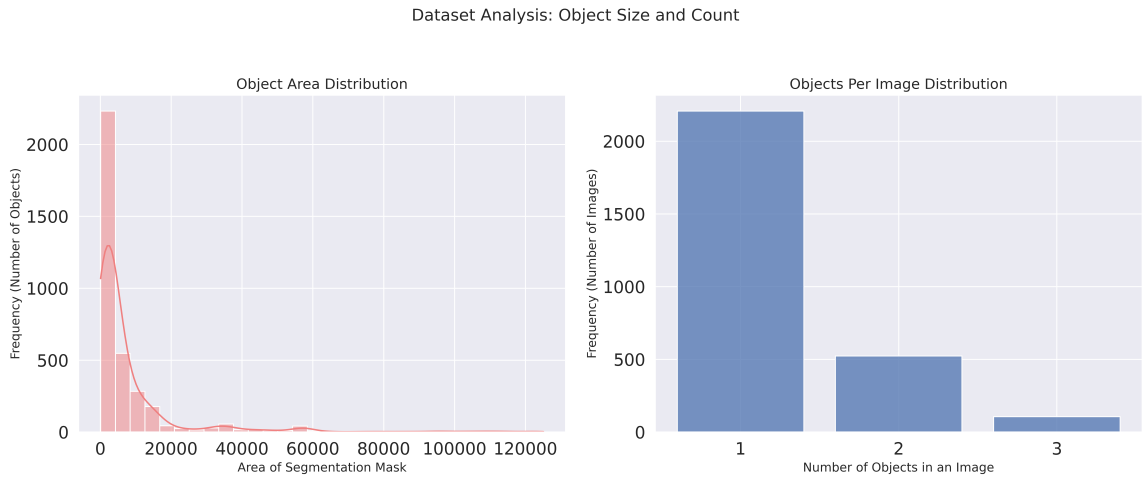


Figure 4.2: Statistical analysis of the annotated objects in the dataset, showing the distribution of object areas (in pixels) and the number of objects per image. The data highlights a predominance of small and localized waste regions, as well as a majority of images containing a single object.

On the left, the histogram illustrates the distribution of object areas (measured in pixels) across all segmentation masks. Given the original image's resolution of 1920×1080, the majority of annotated objects occupy relatively small areas. Most object masks are below 10.000 pixels in size, with a right-skewed distribution that reflects the predominance of small and localized waste regions, typically associated with litter near containers.
On the right, the histogram shows the number of objects per image. Most images contain a single annotated object, indicating isolated occurrences of overflowing waste. This distribution suggests that while some scenes capture multiple waste regions, the dataset is mainly

composed of single-instance images.
In summary, these statistics highlight the imbalanced nature of the dataset in terms of object size and object count per image, indicating that the majority of annotations correspond to small objects and that most images contain only one annotated region.

4.4 Conclusion

In summary, this chapter has detailed the acquisition and processing of a custom dataset. Although the original recordings were collected as continuous video streams, the final dataset used in this dissertation consists exclusively of images extracted from the video sequences. Following the preprocessing and filtering steps, the dataset contains a total of 7,230 images containing scenes with overflowing waste, serving as positive samples for the subsequent experiments.

Chapter 5

Experiments and Results

5.1 Introduction

In this chapter, we present the results of a series of experiments to evaluate the performance of state-of-the-art methods in comparison with our proposed method. These experiments provide valuable insights into the strengths and limitations of current approaches, serving as a baseline for assessing the improvements introduced by our method. We begin by providing a detailed explanation of the experimental setup, including the hardware configuration and the implementation details of the state-of-the-art methods under consideration. Then, to evaluate how well these approaches solve the objective problem, we compare their performance across several metrics. The results presented in this chapter offer both a reference point for current capabilities and a foundation for demonstrating the effectiveness of our proposed method.

5.2 Implementation Details

The selected state-of-the-art models were YOLOv11 [17], Mask R-CNN [1], SOLOv2 [4], YOLACT [2], LISA [8], and Mask2former [5]. All experiments were conducted on NVIDIA GeForce RTX 5070 with 12 GB of memory, and the framework utilized was PyTorch.

To ensure consistency across experiments, all models were trained using a uniform image size of 640×640. The batch size was set to 8 for all models, except for Mask2Former [5], which required a reduced batch size of 4 due to memory constraints.

For each model, hyperparameters were tuned to ensure convergence to an optimal state, maximizing the individual performance of each architecture.

5.3 Metrics

To evaluate the performance of the analyzed state-of-the-art methods, we use several metrics that thoroughly assess their ability to detect overflowing waste in trash bins and that are commonly used in segmentation and detection tasks. These metrics include precision, recall, mAP, precision-recall curves, and gIoU.

Regarding the precision-recall curves, these are created by plotting the precision against the recall at various thresholds. Precision is defined as the ratio of true positive detections to the total number of predicted positive detections (true positives + false positives), and recall is the ratio of true positive detections to the total number of actual positives (true positives +

false negatives). Formally, precision (P) and recall (R) are defined as:

$$P = \frac{TP}{TP + FP} \quad (5.1)$$

$$R = \frac{TP}{TP + FN}, \quad (5.2)$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. By adjusting the detection/confidence threshold, a trade-off between precision and recall is achieved, allowing us to visualize the curve. The area under the precision-recall curve (AUC-PR) is often used to summarize the performance, with higher values indicating better overall performance in distinguishing between true and false positives.

In terms of mAP, this metric can be interpreted as the area under the precision-recall curve. This metric is computed by averaging the Average Precision (AP) across different recall levels for all classes. The AP for a given class is the area under the precision-recall curve, which is calculated as:

$$\int_0^1 P(r)\, dr, \quad (5.3)$$

where $P(r)$ represents the precision at recall level r. The mAP is the average of APs for all classes considered in the task; if only one class is being predicted, then $mAP = AP$. This metric provides a comprehensive measure of a model's accuracy in detecting relevant objects across various thresholds and is widely used for evaluating detection and segmentation models. The measurement of inference time is straightforward. This is obtained directly as a mean of the time taken to infer the segmentation mask on each test set image. Inference time, typically measured in milliseconds or seconds, provides an important metric for evaluating the computational efficiency of a model. The inference time for each method is averaged over all test images to obtain a consistent performance metric.

Lastly, the gIoU is an extension of the traditional IoU, which is commonly used to evaluate object detection models. The IoU is defined as the area of overlap between the predicted segmentation mask and the ground truth mask, divided by the area of their union:

$$IoU = \frac{A \cap B}{A \cup B}, \quad (5.4)$$

where A and B represent the predicted and ground truth areas, respectively. However, IoU can be biased in cases of small or irregularly shaped objects. The gIoU modifies this by also considering the area outside of the union but within the smallest enclosing box that contains both the predicted and ground truth masks. The gIoU is calculated as:

$$gIoU = IoU - \frac{C - (A \cup B)}{C}, \quad (5.5)$$

where C is the smallest enclosing box containing both A and B, and C is its area. This adjustment helps to account for situations where objects may be partially detected but not fully enclosed, providing a more balanced evaluation of model performance.

5.4 Experiments

The experimental study is structured into four distinct phases:

- The first experiment serves as the baseline, where state-of-the-art models are trained on the original dataset composed of fisheye images;
- The second experiment also serves as the baseline, but the state-of-the-art models are trained on a corrected version of the dataset, that is, the original fisheye images are pre-processed to remove lens distortion, converting the pixels from the fisheye projection model to the pinhole model;
- The third experiment implements our proposed method. We seek to evaluate the performance of state-of-the-art methods when surface normals and depth maps are introduced as input to the models, providing geometric information of the scene and aiding in segmenting overflow;
- The last experiment consists of two ablation studies to assess the individual contribution of each geometric modality. The first evaluates performance using only surface normals (excluding depth), while the second uses only depth maps (excluding normals).

5.4.1 Baseline using Fisheye Projection Model

This experiment aims to evaluate the performance of state-of-the-art models when trained on the original fisheye images. The goal is to assess the limitations of current approaches under this projection and to identify which model performs best for the overflow segmentation task in this setting.
In Figure 5.1, the precision-recall curves for the analyzed state-of-the-art methods are depicted, and, as observed, YOLOv11 [17] demonstrates the best overall performance, achieving the highest mAP score of 0.49, with relatively stable precision across various recall levels. This highlights its robustness in maintaining accurate predictions even as recall increases. YOLACT [2] follows with an mAP of 0.41, showing reasonably consistent results but with some decline in precision at higher recall values. Mask R-CNN [1] achieves an mAP of 0.39, reflecting a balanced performance overall, though it suffers more noticeable precision drops as recall grows. In contrast, SOLOv2 [4] exhibits a lower mAP of 0.19, suggesting challenges in adapting to the fisheye images or segmenting complex scenes. Mask2Former [5] performs the worst among the evaluated methods, with an mAP of 0.18, indicating difficulties in both detection consistency and segmentation accuracy under these conditions. This result suggests that the model's architecture may require larger or more diverse training data due to its use of an attention mechanism.
As for quantitative results, Table 5.1 presents the mAP at IoU thresholds of 0.5 (mAP@0.5) and 0.5:0.95 (mAP@50:95), as well as the gIoU.
Among the evaluated methods, YOLOv11 [17] achieves the highest mAP@0.5 of 0.50 and

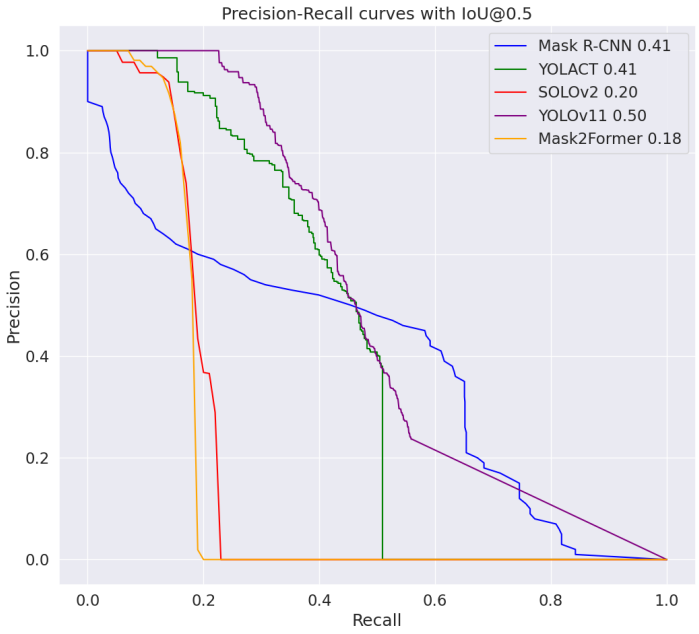


Figure 5.1: Comparison of the precision-recall curves with an IoU of 0.5 reporting the mean mAP@0.5 value for some state-of-the-art segmentation models using fisheye projection model.

mAP@50:95 of 0.30, confirming its robust detection capabilities across different IoU thresholds and its suitability for applications requiring consistent performance. YOLACT [2] follows closely, with an mAP@0.5 of 0.41 and the highest gIoU of 0.51, highlighting its strength in accurate segmentation despite slightly lower detection scores compared to YOLOv11 [17]. Mask R-CNN [1] obtains an mAP@0.5 of 0.41 and a gIoU of 0.43, reflecting a balanced performance but indicating some limitations relative to YOLACT [2]. In contrast, SOLOv2 [4] shows a lower mAP@0.5 of 0.20 and gIoU of 0.30, suggesting challenges in adapting its dynamic instance segmentation approach to the characteristics of the fisheye dataset, possibly due to the complexity of the task and insufficient training data. Mask2Former [5] exhibits the lowest performance among the tested methods, with an mAP@0.5 of 0.18 and gIoU of 0.12, indicating difficulties in segmentation accuracy under these conditions. This suggests that its transformer-based architecture may require larger or more diverse training data to fully exploit its potential in this context. Finally, LISA [8], a foundational model that leverages reasoning segmentation via LLM, presents an interesting case. Its performance, represented solely by gIoU, presents the lowest value (0.03), which can be explained by the generalist nature of LISA [8]. As it's pretrained on a broad range of tasks, allowing it to generalize well, it limits its performance, for example, in specific segmentation tasks. It is worth mentioning that the absence of mAP metrics for LISA [8] is due to the lack of confidence scores associated with its predicted masks.

5.4.2 Baseline using Pinhole Projection Model

This experiment aims to evaluate the performance of state-of-the-art models when trained

Table 5.1: Performance of the state-of-the-art methods using fisheye images.

Experiment 1 - Fisheye Images			
Model	mAP@0.5	mAP@0.5:0.95	gIoU
YOLOv11 [17]	**0,50±0,01**	**0,30±0,01**	-
Mask R-CNN [1]	0,41±0,02	0,26±0,01	0,43±0,01
SOLOv2 [4]	0,20±0,03	0,10±0,01	0,30±0,01
YOLACT [2]	0,41±0,02	0,22±0,01	**0,51±0,03**
LISA [8]	-	-	0.03±0,01
Mask2former [5]	0,18±0,02	0,11±0,01	0,12±0,01

on pinhole images. The goal is to assess the limitations of current approaches under this projection and to identify which model performs best for the overflow segmentation task in this setting.

Table 5.2 summarizes the quantitative results of the evaluated methods, while Figure 5.2 illustrates the corresponding precision-recall curves at an IoU threshold of 0.5.

Table 5.2: Performance of the state-of-the-art methods using pinhole images with distortion removal.

Experiment 2 - Pinhole Images			
Model	mAP@0.5	mAP@0.5:0.95	gIoU
YOLOv11 [17]	**0,45±0,02**	**0,30±0,02**	-
Mask R-CNN [1]	0,29±0,04	0,18±0,03	0,36±0,03
SOLOv2 [4]	0,14±0,01	0,03±0,01	0,26±0,01
YOLACT [2]	0,38±0,03	0,22±0,02	**0,51±0,02**
LISA [8]	-	-	0,10±0,00
Mask2former [5]	0,20±0,01	0,08±0,01	0,12±0,00

Among the models tested, YOLOv11 [17] achieves the highest mAP@0.5 of 0.45 and mAP@50:95 of 0.30, consistent with its superior performance under the fisheye projection. These results highlight YOLOv11's [17] robustness and generalization capabilities across distinct projection models, establishing it as the most reliable model for the overflow segmentation task in both fisheye and pinhole domains.

YOLACT [2] also delivers a strong performance with an mAP@0.5 of 0.38 and the highest gIoU of 0.51. Mask R CNN [1] achieves an mAP@0.5 of 0.29 and a gIoU of 0.36, again showing degraded performance compared to the fisheye setting. SOLOv2 [4] achieves an mAP@0.5 of 0.14 and a gIoU of 0.26, performing similarly poorly as in the fisheye experiment. This suggests inherent challenges in its dynamic kernel generation mechanism when applied to the complex appearance of waste overflow scenarios. Mask2Former [5], with an mAP@0.5 of 0.20 and a gIoU of 0.12, outperforms the fisheye projection. Nevertheless, its performance is weak compared with the other models, which suggests that its transformer-based architecture may not be optimally suited for this specific segmentation task, especially

when training data is limited. Lastly, LISA [8], evaluated only in terms of gIoU (0.10), shows limited performance in the pinhole setting, similar to its results under fisheye projection.

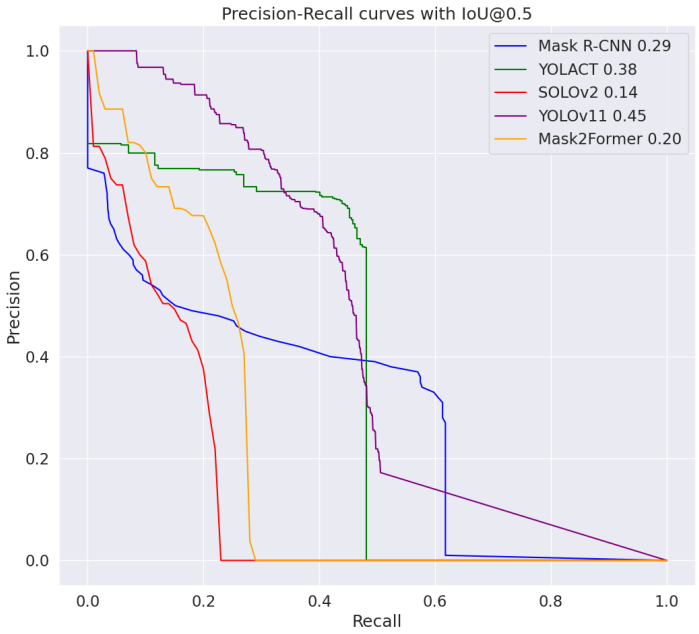


Figure 5.2: Comparison of the precision-recall curves with an IoU of 0.5 reporting the mean mAP@0.5 value for some state-of-the-art segmentation models using pinhole projection model.

The precision-recall curves in Figure 5.2 reinforce these findings, with YOLOv11 [17] showing high precision across a wide recall range. YOLACT [2] follows closely, maintaining high precision, up to moderate recall levels, whereas Mask2Former [5] and SOLOv2 [4] drop in precision at lower recall levels, indicating unstable predictions as recall increases.
Altough it might be expected that pinhole projections would benefit model performance due to their reduced distortion, the results suggest the opposite. This may be attributed to the reduced field of view in pinhole images, which limits scene context and reduces the visibility of some overflows. Typically, fisheye lenses are preferred for outdoor tasks, such as urban environments.

5.4.3 Proposed Method using Geometric Information

This experiment reflects our proposed method, which introduces additional information to the models' input.
The results shown in Table 5.3 and Figure 5.3 demonstrate that the integration of geometric information enhances segmentation performance for most evaluated models. Notably, YOLACT [2] and YOLOv11 [17] achieved the highest mAP@0.5 (both 0.52) and mAP@0.5:0.95 (both 0.31), with YOLACT [2] also obtaining the best gIoU score of 0.61.
The inclusion of geometric information appears to enhance segmentation performance, particularly in lightweight architectures, which often rely on strong inductive features and may

Table 5.3: Performance of the state-of-the-art methods for our approach.

Proposed Method			
Model	mAP@0.5	mAP@0.5:0.95	gIoU
YOLOv11 [17]	**0,52±0,01**	**0,31±0,01**	-
Mask R-CNN [1]	0,12±0,01	0,06±0,01	0,21±0,01
SOLOv2 [4]	0,07±0,01	0,03±0,00	0,19±0,00
YOLACT [2]	**0,52±0,02**	**0,31±0,02**	**0,61±0,02**
Mask2former [5]	0,29±0,02	0,13±0,01	0,14±0,00

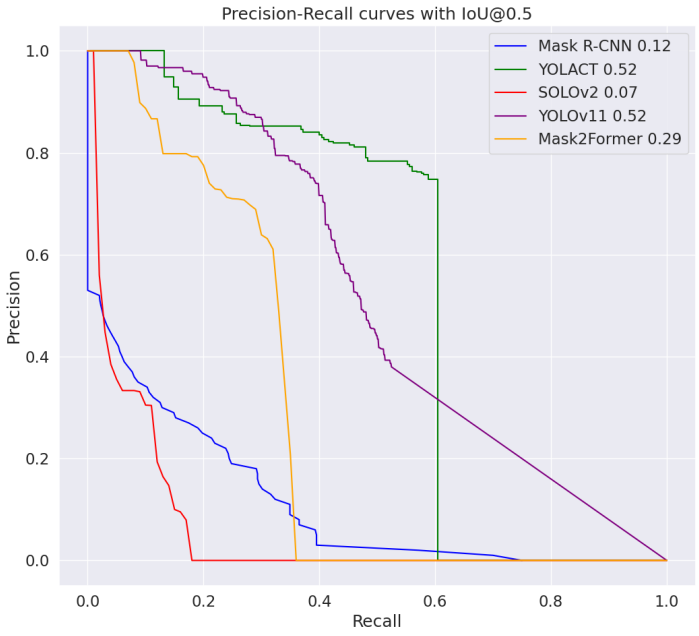


Figure 5.3: Comparison of the precision-recall curves with an IoU of 0.5 reporting the mean mAP@0.5 value for some state-of-the-art segmentation models for our proposed method.

lack the capacity to infer such relationships implicitly.

On the one hand, Mask2Former [5] shows a substantial performance improvement, increasing from 0.18 mAP@0.5 in the fisheye-only setup to 0.29 using our method. This improvement may be attributed to the model's attention-based architecture, which likely leveraged the added geometric channels to establish relationships between the channels. This result reinforces the robustness of our method, which demonstrates that even complex models with a risk of overfitting can benefit when geometric information is properly integrated.

On the other hand, models such as Mask R-CNN [1] and SOLOv2 [4] underperformed in this setup. While Mask R-CNN [1] achieved moderate scores (mAP@0.5 = 0.12, gIoU = 0.21), SOLOv2 [4] failed to detect most of the instances (mAP@0.5 = 0.07, gIoU = 0.19). A likely explanation lies in their architectural complexity and reliance on either fixed region proposals or grid-based instance grouping. These methods are more prone to overfitting under limited data scenarios, and the additional geometric channel may have exacerbated

this sensitivity rather than mitigated it. This limitation highlights a key challenge, while our method improves performance across diverse architectures, it is especially effective for efficient models, whereas very complex models may require larger datasets to fully leverage the enriched input representation.
In summary, our proposed method demonstrates consistent gains across most segmentation architectures and lightweight models.

5.4.4 Ablation Studies

We conduct ablation studies to analyze how surface normals and depth maps affect the performance of our method. These studies allow us to isolate the influence of each geometric feature and assess its impact on segmentation performance across different model architectures.

5.4.4.1 Impact of Depth Information

To analyze the impact of depth information in our method, we decided to remove this component from the input of the model. The results of this decision are presented in Table 5.4.

Table 5.4: Performance of the state-of-the-art methods for the first ablation study.

Ablation Study 1 - Impact of Depth Information			
Model	mAP@0.5	mAP@0.5:0.95	gIoU
YOLOv11 [17]	0,54±0,01	**0,32±0,01**	-
Mask R-CNN [1]	0,28±0,01	0,12±0,01	0,33±0,01
SOLOv2 [4]	0,15±0,02	0,05±0,01	0,08±0,01
YOLACT [2]	**0,56±0,01**	0,29±0,01	**0,61±0,01**
Mask2former [5]	0,23±0,03	0,12±0,02	0,08±0,01

It is observed that YOLACT [2] improves slightly, achieving the best mAP@0.5 (0.56) and gIoU (0.61). YOLOv11 [17] also demonstrates a slight improvement in mAP@0.5:0.95 (from 0.31 to 0.32) and in mAP@0.5 (from 0.52 to 0.54). These results suggest that depth may not be essential and may even harm lightweight models.
However, the removal of depth significantly affects complex architectures like Mask2Former [5], whose performance drops from 0.28 to 0.23 in mAP@0.5 and from 0.13 to 0.12 in mAP@0.5:0.95. This reveals that transformer-based architectures depend on explicit spatial cues, where the absence of depth harms their ability to capture spatial relationships between objects and the background.
In summary, this ablation study indicates that while depth information is not strictly necessary for lightweight models, it remains a valuable source for more complex architectures that use the attention mechanism.

5.4.4.2 Impact of Surface Normals Information

To analyze the impact of surface normal information in our method, we decided to remove this component from the input of the model. The results of this decision are presented in

Table 5.5.

Table 5.5: Performance of the state-of-the-art methods for the second ablation study.

Ablation Study 2 - Impact of Surface Normals Information			
Model	mAP@0.5	mAP@0.5:0.95	gIoU
YOLOv11 [17]	**0,53±0,01**	**0,32±0,01**	-
Mask R-CNN [1]	0,12±0,01	0,05±0,00	0,20±0,01
SOLOv2 [4]	0,07±0,01	0,03±0,00	0,14±0,01
YOLACT [2]	0,50±0,02	0,30±0,02	**0,72±0,02**
Mask2former [5]	0,14±0,02	0,08±0,01	0,16±0,01

In this case, the majority of models showcase performance drops, confirming the relevance of surface normals as a geometric cue.
Regarding Mask2Former [5], the mAP@0.5 drops from 0.28 (with complete geometric information) to 0.14 without surface normals, reinforcing the idea that the attention mechanism in this model benefits significantly from this information. SOLOv2's [4] performance remains similar with an mAP@0.5 of 0.07, which highlights the importance of surface normal features to disambiguate object boundaries in challenging scenes.
It is also observed that YOLOv11 [17] and YOLACT [2] maintain their performance, and YOLACT [2] even achieves a gIoU of 0.72, the highest across all settings. This suggests that while surface normals are useful, certain architectures may internally learn effective spatial priors that compensate for their absence.

5.4.5 Overall Results

In summary, Table 5.6 presents the performance across the five instance segmentation models under the three different projection settings: fisheye, pinhole, and our proposed method. Overall, the proposed method outperforms the fisheye and pinhole projections in terms of mAP@0.5 and mAP@0.5:0.95, particularly for lightweight models.
YOLACT [2] achieved the highest overall performance, reaching an mAP@0.5 of 0.53 and mAP@0.5:0.95 of 0.31 using our method, outperforming the fisheye (with an mAP@0.5 of 0.41) and pinhole (with an mAP@0.5 of 0.36) projections. Similar improvements were observed in YOLOv11 [17] and Mask2Former [5], with YOLOv11 [17] reaching an mAP@0.5 of 0.52 (vs. 0.49 in fisheye and 0.46 in pinhole), and Mask2Former [5] improving to 0.28 (from 0.18 in fisheye and 0.20 in pinhole). These results indicate that our method enhances segmentation performance in wide-angle images, likely due to better spatial representation and mitigation of projection-induced artifacts. Mask R-CNN [1] and SOLOv2 [4] did not benefit from the proposed method, even showing degraded performance (with an mAP@0.5 of 0.11 and 0.06, respectively). This suggests that complex models may require more data to benefit from additional information.
To further support these findings, Figure 5.4 presents the precision-recall curves for the

Table 5.6: Performance of the overall results.

Overall Results			
Models	Experiments	mAP@0.5	mAP@0.5:0.95
YOLOv11 [17]	Fisheye Projection	0,50±0,01	0,30±0,01
	Pinhole Projection	0,45±0,02	0,30±0,02
	Proposed Method	**0,52±0,01**	**0,31±0,01**
Mask R-CNN [1]	Fisheye Projection	**0,41±0,02**	**0,26±0,01**
	Pinhole Projection	0,29±0,04	0,18±0,03
	Proposed Method	0,12±0,01	0,06±0,01
SOLOv2 [4]	Fisheye Projection	**0,20±0,03**	**0,10±0,01**
	Pinhole Projection	0,14±0,01	0,03±0,01
	Proposed Method	0,07±0,01	0,03±0,00
YOLACT [2]	Fisheye Projection	0,41±0,02	0,22±0,01
	Pinhole Projection	0,38±0,03	0,22±0,02
	Proposed Method	**0,52±0,02**	**0,31±0,02**
Mask2former [5]	Fisheye Projection	0,18±0,02	0,11±0,01
	Pinhole Projection	0,20±0,01	0,08±0,01
	Proposed Method	**0,29±0,02**	**0,13±0,01**

YOLACT [2] model across all projection types, highlighting the superior precision-recall trade-off achieved with our proposed method. We can observe that the curve corresponding to the proposed method maintains higher precision across a broader recall range, reflecting fewer false positives and more confident detections.

Beyond the quantitative metrics, qualitative examples in Figure 5.5 showcase failure cases (false positives and false negatives) under the fisheye projection that were successfully corrected using our method. These visual comparisons illustrate how our approach improves segmentation accuracy and reduces false positives and false negatives, which are common challenges in waste overflow scenarios due to the highly heterogeneous shapes and appearances of overflowing materials.

In summary, the proposed method achieves improvements across various segmentation models, with the most gains observed in lightweight architectures such as YOLACT [2] and YOLOv11 [17].

5.5 Conclusion

This chapter presented a comprehensive experimental evaluation of our proposed method for overflowing waste segmentation, benchmarking it against several state-of-the-art methods. The work carried out was well structured, beginning with baseline performance assessments on original fisheye and distortion-corrected pinhole image projections, followed by an evaluation of our approach, which integrates geometric information as auxiliary input.

The initial baseline experiments revealed that fisheye images, despite their distortions, pro-

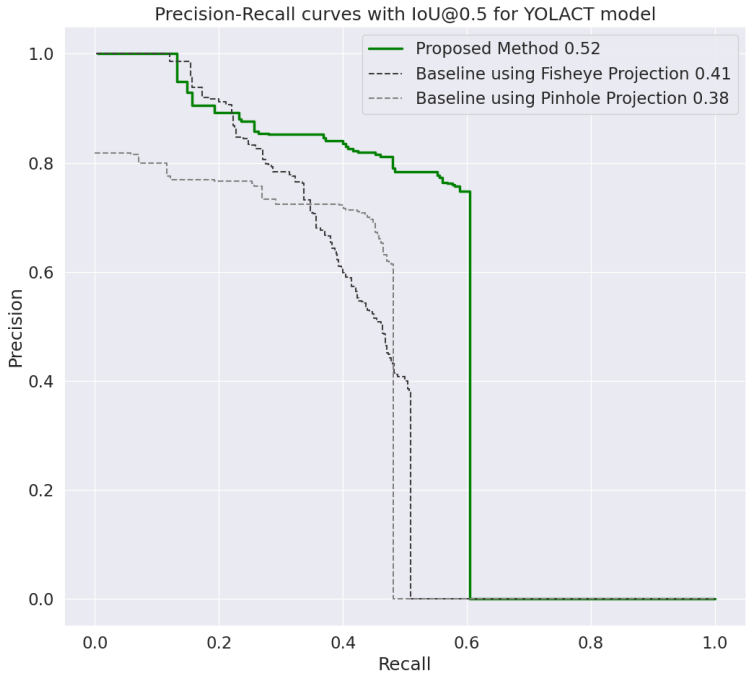


Figure 5.4: Comparison of the precision-recall curves at an IoU threshold of 0.5 reporting the mean mAP@0.5 value for the YOLACT [2] model. Our approach improves segmentation accuracy by approximately 47.22% compared to the pinhole method, and by 29.72% compared to the fisheye method.

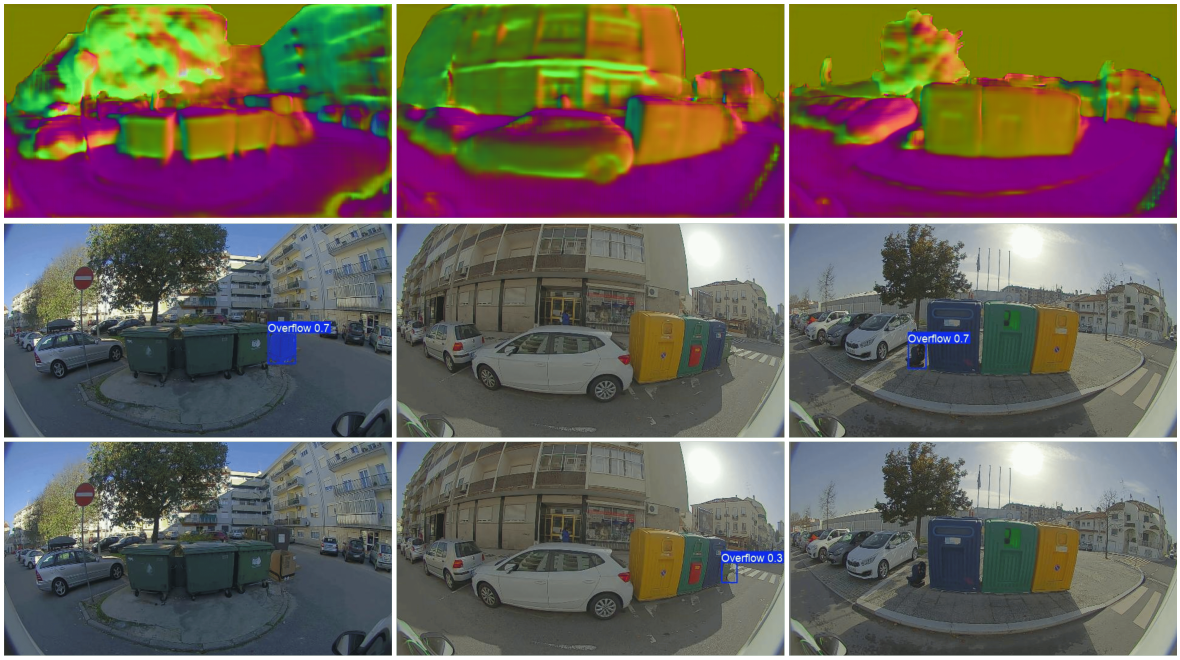


Figure 5.5: Qualitative comparison of segmentation results using our proposed method. The first row displays the estimated surface normals used in our approach. The second row shows the predictions produced by YOLACT [2] when using our method. The third row presents the predictions obtained using YOLACT [2] on the baseline fisheye images. Our method significantly improves segmentation accuracy by increasing confidence scores and reducing false positives and false negatives.

vided a richer contextual field of view, resulting in better performance than the pinhole projection for several models. This counterintuitive finding highlights the significance of the field of view for comprehensive scene understanding in this particular task. Across both baseline scenarios, YOLOv11 [17] and YOLACT [2] demonstrated strong performance.
The introduction of our proposed method, which uses both surface normals and depth maps, demonstrated an improvement in segmentation accuracy for several models. On the one hand, YOLACT [2] and YOLOv11 [17] achieved the highest mAP@0.5 scores of 0.52, with our method, representing substantial gains over their baseline performances. Mask2Former [5] also exhibited significant improvement, suggesting that its attention-based mechanisms can effectively utilize the additional geometric channels. These results validate our hypothesis that explicit geometric information aids models, particularly lightweight architectures.
On the other hand, more complex architectures, such as Mask R-CNN [1] and SOLOv2 [4], did not benefit and suffered from performance degradation. This may be due to an increased susceptibility to overfitting with the additional input dimensionality on the available dataset, or a need for architectural modifications to fuse these new modalities more effectively.
The subsequent ablation studies provided insights into the individual contributions of depth and surface normal information. The exclusion of depth information (retaining normals) resulted in tiny performance gains for YOLACT [2] and YOLOv11 [17], suggesting that for these lightweight models, surface normals might be a more important geometric cue, or that depth, in its current form, might introduce noise or redundancy. However, the removal of surface normals (retaining depth) generally resulted in performance degradation across most models, particularly for Mask2Former [5], highlighting the critical role of surface orientation information in accurate object detection and segmentation.
In summary, the experimental findings detailed in this chapter demonstrate the advantages of integrating geometric information for the task of overflowing waste segmentation. Our proposed method achieved superior performance, particularly for efficient architectures such as YOLACT [2] and YOLOv11 [17], enhancing both detection accuracy and segmentation quality, as evidenced by quantitative metrics and qualitative examples that corrected previous false positives and negatives. While the integration of geometric data poses challenges for some very complex models within the constraints of the current dataset, the overall results affirm the value of our approach.

Chapter 6

Conclusions and Future Work

6.1 Main Conclusions

This dissertation studied the challenge of segmenting overflowing waste in real-world urban environments, an important step toward intelligent and automated waste monitoring systems. Traditional computer vision approaches, which are often based solely on RGB data, struggle under conditions of visual ambiguity and diverse illumination limitations that significantly harm their performance in practical settings. In response to the limitations of existing methods, this dissertation proposed a novel geometry-aware segmentation framework that enriches visual input with geometric features derived from monocular fisheye images.

The main hypothesis of this work was that supplementing RGB data with estimated depth and surface normal maps, generated using the zero-shot capabilities of the Metric3Dv2 [29] model, could provide instance segmentation architectures with a better understanding of the scene. As a result, state-of-the-art models would improve their ability to identify overflowing waste, distinguish it from visually similar elements (such as bin contents or shadows), and mitigate false negatives.

A significant contribution of this dissertation was the creation of a specialized dataset of 7,230 annotated images of overflowing waste and parasitic litter, captured using a vehicle-mounted fisheye camera. This dataset served not only as the empirical foundation for benchmarking but also as a valuable resource for the research community in environmental perception and innovative city applications.

Extensive experiments were conducted across multiple state-of-the-art instance segmentation models, including YOLACT [2], YOLOv11 [17], Mask R-CNN [1], SOLOv2 [4], LISA [8], and Mask2Former [5], under three input conditions: raw fisheye, distortion-corrected pinhole, and RGB fused with geometric features, and the results demonstrated the superiority of our geometry-augmented method, particularly for lightweight models such as YOLACT [2] and YOLOv11 [17]. These models achieved mAP@0.5 scores of 0.52, representing significant improvements over their RGB-only baselines. Mask2Former [5], which is based on the attention mechanism, also benefited from the additional geometric context, highlighting the capacity of transformer architectures to leverage spatial features.

Ablation studies further confirmed that surface normal maps played a more critical role than depth alone in enhancing segmentation accuracy. While some models maintained high performance with just normals and RGB, removing surface normals consistently led to greater performance degradation than omitting depth.

Additionally, it was observed that fisheye images yielded better results compared to their pinhole-corrected version, suggesting that the fisheye lens's wider field of view can compensate for its geometric distortions for this specific dataset.

In conclusion, this research shows that augmenting RGB images with estimated geometric information can improve instance segmentation performance for complex urban perception tasks. The proposed method, combined with the new dataset and thorough analysis, lays the groundwork for more context-aware vision systems for waste monitoring. Beyond the specific case of detecting overflowing waste, this approach offers an efficient pathway toward smarter environmental monitoring in urban spaces, paving the way for sustainable and intelligent cities.

6.2 Future Work

Building upon the findings and contributions of this dissertation, several promising directions for future research can be identified to further advance the integration of geometric cues into waste overflow segmentation.

While Metric3Dv2 [29] demonstrated good performance in zero-shot settings, future work could explore the fine-tuning of this or similar monocular depth and normal estimation models using a dataset specifically adapted to urban waste scenarios. Such domain adaptation is likely to enhance the quality and task relevance of the estimated geometric cues, which remain an open problem for accurately modeling complex overflow structures.

In terms of architectural improvements, more advanced fusion mechanisms within segmentation models could be explored. Techniques such as attention-based cross-modal fusion strategies may enable the model to leverage RGB and geometric channels based on the scene's content. This could be helpful for complex architectures like Mask R-CNN [1] or Mask2Former [5].

Additionally, for future work, it would be beneficial to collect additional images in a wider variety of environmental conditions, including varying weather, lighting, seasons, bin types, and overflow compositions. This would not only increase model robustness but also enable the training of more expressive models. Complementarily, enriching annotations with attributes such as overflow material type, degradation level, or estimated volume would allow for more granular analysis and support downstream applications in waste analytics.

A key objective of this research is its integration into real-world waste monitoring systems. Future work should include deploying the proposed approach in operational settings, such as on municipal collection vehicles, to evaluate its performance under realistic constraints. This includes evaluating model latency and behavior under challenging lighting or occlusion scenarios.

Furthermore, segmentation outputs could be extended toward estimating the volume of overflowing waste. By combining predicted masks with geometric cues (e.g., depth or surface orientation), it would be possible to approximate 3D volume, a crucial metric for waste management services. This direction is based on the metrology concepts introduced in this dissertation and aligns with practical needs in urban sanitation.

Finally, for systems processing video streams, incorporating temporal consistency mechanisms (e.g., tracking or video object segmentation) could help stabilize predictions over time. Utilizing this type of information from previous frames may enhance segmentation coherence and minimize flickering or inconsistent detections.

Detection of Overflowing Waste and Litter near Trash Bins

By concentrating on these key areas, the work presented in this dissertation can be expanded to develop intelligent and context-aware systems for environmental monitoring, supporting more efficient urban waste management, and contributing to sustainability goals.

Bibliography

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870 xv, 5, 6, 8, 9, 12, 16, 21, 31, 33, 34, 35, 37, 38, 39, 40, 42, 43, 44

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," 2019. [Online]. Available: https://arxiv.org/abs/1904.02689 xv, xvi, xvii, 6, 7, 8, 12, 21, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43

[3] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 649–665. xv, 9, 10, 11

[4] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," 2020. [Online]. Available: https://arxiv.org/abs/2003.10152 xv, 11, 12, 13, 21, 31, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43

[5] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299. xv, 12, 31, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643 xv, 13, 21

[7] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. xv, 13

[8] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589. xv, 14, 21, 31, 34, 35, 36, 43

[9] A. Oğuz and Ö. F. Ertuğrul, "Determining the fullness of garbage containers by deep learning," *Expert Systems with Applications*, vol. 217, p. 119544, 2023. xv, 15, 21

[10] P. F. Proença and P. Simões, "Taco: Trash annotations in context for litter detection," *arXiv preprint arXiv:2003.06975*, 2020. xv, 16, 17

[11] P. Barra, A. A. Citarella, G. Orefice, M. Castrillón-Santana, and A. Ciaramella, "Lots: Litter on the sand dataset for litter segmentation," in *18th International Conference on Machine Vision and Applications (MVA)*. IEEE, 2023, pp. 1–4. xv, 16, 17

[12] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, "Single view metrology in the wild," in *European Conference on Computer Vision*. Springer, 2020, pp. 316–333. xv, 17, 18

[13] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, and C. Shen, "Metric3d: Towards zero-shot metric 3d prediction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9043–9053. xv, 19, 20, 21

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016. 5

[15] R. Girshick, "Fast r-cnn," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. USA: IEEE Computer Society, 2015, p. 1440–1448. [Online]. Available: https://doi.org/10.1109/ICCV.2015.169 5

[16] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 8, 16

[17] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024. 8, 9, 31, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43

[18] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *Advances in neural information processing systems*, vol. 31, 2018. 10

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 12

[20] H. N. Kulkarni and N. K. S. Raman, "Waste object detection and classification," *CS230 Stanford*, 2019. 15

[21] A. Balmik, S. Barik, M. Jha, and A. Nandy, "A vision-based litter detection and classification using ssd mobilenetv2," in *2023 10th International conference on signal processing and integrated networks (SPIN)*. IEEE, 2023, pp. 180–185. 15

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37. 15

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 15

[24] B. De Carolis, F. Ladogana, and N. Macchiarulo, "Yolo trashnet: Garbage detection in video streams," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 2020, pp. 1–7. 16

[25] H. Vasa, *Google images download*, 2024, accessed: 2024-10-23. [Online]. Available: https://google-images-download.readthedocs.io/en/latest/index.html 16

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571. 16

[27] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer, 2012, pp. 746–760. 19

[28] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013. 19

[29] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 19, 20, 24, 43, 44