UNIVERSIDADE
BEIRA INTERIOR

Vision-Based Waste Detection for Industrial Sorting Lines

Sara Oliveira Inácio

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

junho 2025

Declaração de Integridade

Eu, Sara Oliveira Inácio, que abaixo assino, estudante com o número de inscrição M13503 do Mestrado em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

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

Universidade da Beira Interior, Covilhã, 11/06/2025

Assinado por: **Sara Oliveira Inácio**
Num. de Identificação: 30994073

Agradecimentos

Aos meus pais,
pelo apoio incondicional, pela oportunidade, por acreditarem sempre. Por serem o meu porto seguro.

Às minhas irmãs, Raquel e Marta,
por serem seres tão especiais e essenciais na minha vida, por me darem força inabalável, por estarem comigo em cada passo desta jornada. Por serem uma fonte inesgotável de motivação.

Aos meus amigos,
pela paciência, por toda a compreensão nos momentos mais frágeis, por terem sido pacientes com o meu tempo, às vezes, escasso e por continuarem a estar presentes apesar de tudo.

Aos que caminharam ao meu lado durante estes dez meses — Diogo, João e Bernardo,
por todas as vezes que me fizerem esquecer o medo, que me fizeram rir e foram uma boa companhia.

Ao meu orientador, Professor Doutor João Neves,
por todo o acompanhamento cuidadoso, pela segurança transmitida quando os desafios pareciam insuperáveis, e pelo conhecimento generosamente partilhado.

Ao meu coorientador, Professor Doutor Hugo Proença,
pela orientação clara, pela partilha de conhecimento e pela ajuda preciosa na definição dos objetivos que guiaram esta investigação.

Por fim,
a todas as pessoas que cruzaram o meu caminho, que me permitiram errar e aprender, que me moldaram com as suas palavras e exemplos.

Prefácio

“One day, in retrospect, the years of struggle will strike you as the most beautiful.”
— Sigmund Freud

Resumo

A crescente produção de resíduos, impulsionada pelo crescimento populacional e pelo consumo excessivo, tem dificultado a sua gestão e reciclagem. A separação manual dos resíduos é uma prática comum, mas apresenta riscos para o ser humano e, ultimamente, tem-se mostrado insuficiente para lidar com as quantidades produzidas. A presença de materiais valorizáveis nos resíduos encaminhados para aterro contribui para aumentar a poluição dos solos e dos recursos hídricos, impactando negativamente a saúde humana e os ecossistemas. Com o avanço das tecnologias de aprendizagem profunda e de visão computacional, surgem alternativas para automatizar e otimizar o processo de triagem destes resíduos. Assim, esta dissertação propõe uma abordagem baseada em visão computacional para a deteção de resíduos valorizáveis numa linha de triagem de resíduos sólidos urbanos (RSU). Com este objetivo, foi desenvolvido um *dataset*, recolhido numa linha de tratamento mecânico e biológico (TMB), com oito categorias de resíduos. Foram treinados e avaliados diversos modelos de deteção de objetos do estado de arte, como Faster R-CNN, RetinaNet, TridentNet e YOLO. Os resultados demonstraram um desempenho promissor, alcançando uma mAP de 59.7% no conjunto de teste. A investigação e o *dataset* desenvolvidos representam um contributo para a aplicação destas tecnologias no setor industrial, promovendo a segurança dos trabalhadores, a eficiência do processo de reciclagem e o desenvolvimento de soluções inovadoras para a valorização de resíduos.

Palavras-chave

Aprendizagem Profunda, Visão Computacional, Deteção de Objetos, Deteção de Resíduos, Triagem de Resíduos, Reciclagem

Resumo alargado

A crescente produção de resíduos, impulsionada pelo aumento populacional e pelo consumo excessivo, tem dificultado a sua gestão e reciclagem. A separação manual é uma prática comum nos centros de triagem, mas apresenta riscos para os trabalhadores e tem-se revelado insuficiente para lidar com as grandes quantidades produzidas.
Em Portugal, é comum a existência de duas linhas de triagem: uma de fluxo único, dedicada ao tratamento mecânico e biológico de resíduos sólidos urbanos, e outra de fluxo múltiplo, onde são separados plásticos e metais. Na linha de fluxo único, é frequente a presença de materiais valorizáveis que, na ausência de separação manual, acabam por ser encaminhados para aterro. Quando os plásticos são depositados em aterro, podem, ao degradarem-se, libertar substâncias tóxicas, contaminando o solo e os recursos hídricos, impactando negativamente o ambiente, na saúde humana e animal.
Neste contexto, esta dissertação propõe o desenvolvimento de uma solução baseada em visão computacional para a deteção automática de materiais valorizáveis, nas instalações de tratamento mecânico e biológico (TMB), onde é feita a triagem de resíduos sólidos urbanos (RSU). O trabalho incluiu a recolha e anotação de um *dataset*, captado numa linha de triagem industrial, com o intuito de treinar e testar modelos de deteção de objetos de forma realista.
O *dataset* desenvolvido foi recolhido numa linha de RSU e anotado manualmente com recurso à ferramenta CVAT. Este está dividido em conjuntos de treino, validação e teste, totalizando 5261 imagens, e contém oito categorias de materiais. Para avaliar a eficácia dos modelos do estado de arte no domínio da deteção de resíduos, foram treinados e testados diversos, entre os quais o Faster R-CNN, RetinaNet, TridentNet e YOLO. Os modelos demonstraram desempenhos promissores, revelando-se adequados à aplicação em cenários reais.
Inicialmente, foram realizados treinos e testes dos modelos utilizando as oito classes anotadas no *dataset*. Posteriormente, com o intuito de alcançar o principal objetivo desta dissertação – a deteção de plásticos nas linhas de triagem de RSU – algumas classes foram agrupadas e outras removidas, tendo sido então realizados novos treinos e testes com modelos de estado da arte. Adicionalmente, foram efetuados testes comparativos com um *dataset* de referência e realizada uma avaliação cruzada para testar a capacidade de generalização dos modelos em diferentes ambientes.
Assim, a investigação desenvolvida e o *dataset* são um contributo para o avanço da aplicação de técnicas de inteligência artificial na área da gestão de resíduos. Esta abordagem tem o potencial de aumentar a eficiência do processo de reciclagem, reduzir a exposição dos trabalhadores a ambientes potencialmente perigosos e promover soluções sustentáveis para a valorização de resíduos.

Abstract

The increasing production of waste, driven by population growth, has created challenges in managing and recycling materials effectively. Manual waste sorting is a common practice, but it has proven insufficient to handle the large quantities produced and poses health risks to workers involved in the process. The presence of valuable materials in waste streams directed to landfills contributes to soil and water pollution, negatively impacting human health and ecosystems. With advances in deep learning and computer vision technologies, new solutions have emerged to automate and optimize waste sorting processes. In this context, this dissertation proposes a computer vision-based approach for the automatic detection of valuable waste materials in a municipal solid waste (MSW) sorting line. To support this work, a dataset was developed, collected from a Mechanical-Biological Treatment (MBT) facility, comprising eight waste categories. Several state-of-the-art object detection models, including Faster R-CNN, RetinaNet, TridentNet and YOLO, were trained and evaluated. The results demonstrated promising performance, achieving a mAP of 59.7% on the test set. This research and the developed dataset represent a contribution to the application of these technologies in the industrial sector, enhancing worker safety, increasing the efficiency of the recycling process, and supporting the development of sustainable solutions for waste valorization.

Keywords

Deep Learning, Computer Vision, Object Detection, Waste Detection, Waste Sorting, Recycling

Contents

List of Figures

List of Tables

Acronyms

AP Average Precision

AUVs Autonomous Underwater Vehicles

CV Computer Vision

CNN Convolutional Neural Network

CBAM Convolutional Block Attention Module

CVAT Computer Vision Annotation Tool

CotNet Contextual Transformer Networks

C2PSA Cross Stage Partial with Spatial Attention

DL Deep Learning

DETR DEtection TRansformer

ECAL Liquid Food Carton Packaging

FFN Feed-Forward Network

FPN Feature Pyramid Networks

GINI Garbage in Photos

HDPE High-Density Polyethylene

IoU Intersection-over-Union

J-EDI JAMSTEC E-Library of Deep-sea Images

LWW Labeled Waste in the Wild

mAP mean Average Precision

MBT Mechanical-Biological Treatment

MSW Municipal Solid Waste

MRFs Material Recovery Facilities

NMS Non-maximum Suppression

PR Precision-Recall

PET Polyethylene Terephthalate

PANet Path Aggregation Network

RPN Region Proposal Network

RoI Region of Interesting

R-CNN Regions with Convolutional Neural Networks

SPP Spatial Pyramid Pooling

SSD Single Shot Multibox Detector

SPPF Spatial Pyramid Pooling-Fast

TACO Trash Annotations in Context

TRWD Taiwan Recycled Waste Dataset

YOLO You Only Look Once

Chapter 1

Introduction

1.1 Context and Motivation

The exponential global population increase and industrial development have boosted waste production, making it one of our most pressing environmental challenges. The inadequate management of generated waste negatively impacts human health, animals, water resources, and the environment [8]. As a result, finding innovative and more effective ways to manage and reduce waste is crucial for building a more sustainable future.

One of the most significant challenges in this area is managing solid waste. According to a 2018 report by the World Bank [9], global waste production was estimated at 2.01 billion tons per year and is expected to increase to 3.40 billion tons by 2050 if current trends continue.

In Portugal, waste is sorted using two central systems. The first is a single-stream system, also known as Mechanical-Biological Treatment (MBT), which handles Municipal Solid Waste (MSW) from unsorted bins (typically grey or green). The second is a multi-stream system used for selectively collecting materials, such as plastics and metals. One issue with the single-stream MBT system is that it often contains valuable recyclable materials—things that could have been properly sorted but were thrown in the wrong bin. As [10] points out, in urban areas, MBT systems can sometimes recover recyclables at rates similar to those of selective collection systems, showing that improving MBT sorting could make a significant difference.

Deep Learning (DL) and Computer Vision (CV) advancements have emerged as opportunities to address the challenges of recycling and identifying valuable waste. These technologies have the potential to enhance efficiency, reduce the need for direct human intervention, and minimize health risks to workers involved in waste management. However, automating waste sorting presents intrinsic difficulties, such as variability in the types, sizes, and conditions of objects, which may be broken, deformed, or overlapped, as well as the absence of public datasets specific to this task. Considering this, creating a specialized dataset for training these models is essential.

This thesis aims to develop a new vision-based system for detecting valuable recyclable materials, specifically plastics, in MSW sorting lines. A key part of this work involves building a custom dataset designed explicitly for this purpose, as existing public datasets do not accurately reflect the real conditions found in MBT facilities.

This project was developed in collaboration with Evox Technologies, a company focused on innovation in waste management and recycling.

1.2 Objectives and Contributions

The primary objective of this dissertation is to develop a vision-based system for detecting valuable materials, specifically plastics, on sorting lines, with a focus on applications in MBT facilities. By doing this, the aim is to enhance recycling processes in industrial environments, minimize human intervention, and improve worker safety by mitigating exposure to potentially hazardous conditions.

One of the biggest challenges of this work is the lack of existing datasets that reflect real-world conditions in MSW sorting lines. Therefore, a key contribution of this project involves creating a new dataset collected from industrial environments. This dataset will be annotated and will be used to train and test state-of-the-art models, helping to identify limitations and leading to more accurate solutions for valuable waste.

In addition, this study intends to support improved waste management by highlighting how technology can address some of today's critical environmental challenges. The proposed solutions are designed to make recycling processes more efficient and sustainable while also improving worker safety by automating tasks that are typically performed manually and often involve hazardous working conditions.

1.3 Tasks and Timeline

The tasks delineated in the work plan have been designed to achieve the objectives of this dissertation and are illustrated in the Gantt Chart in 1.1.

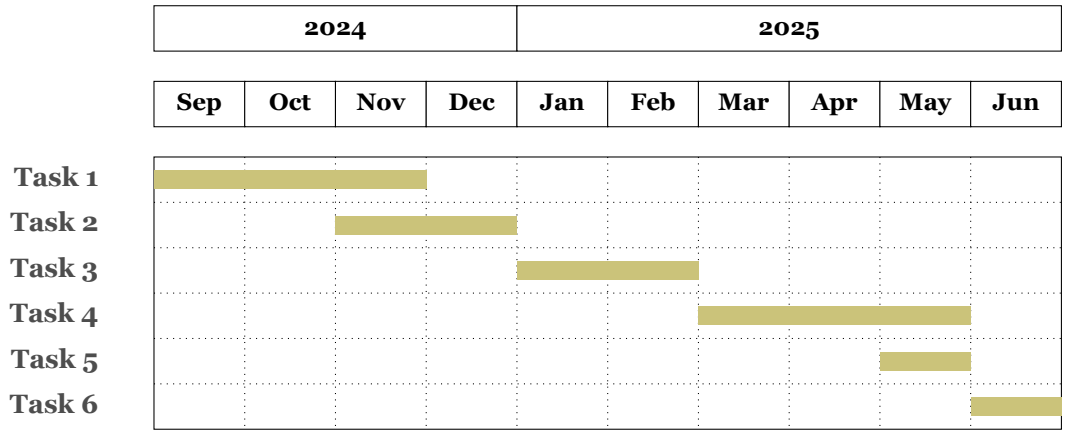


Figure 1.1: Gantt chart illustrating the development timeline and duration of the proposed tasks.

1.3.1 Task 1 - Review of the literature

The project's first task is a comprehensive literature review and analysis of related works. This initial phase is dedicated to the study of waste detection. Different models were explored, including one-phase, two-phase, and transformer-based approaches. Next, models developed specifically for waste detection are explored to contribute to more efficient solid

waste management. In addition, public datasets in this field are analyzed to identify those that could be useful for developing this project.

1.3.2 Task 2 - Evaluation of the state-of-the-art models in ZeroWaste dataset

After analyzing state-of-the-art models and datasets, the models are evaluated on a unique dataset designed explicitly for detecting waste in industrial conditions. This phase involved adapting the model implementations to align with the task requirements. Performance metrics are established to evaluate the models, and the tested methods are compared to identify the best performance and the limitations of the dataset.

1.3.3 Task 4 - Collection and annotation of a new dataset

Posterior to the first experiments and the evaluation of state-of-the-art models on the current dataset, it was concluded that it is essential to develop a new dataset to train models that can perform the task satisfactorily. Thus, collecting and annotating a new dataset becomes a crucial step. With the new dataset, it is possible to train and evaluate models more effectively, ensuring that the conclusions obtained are valid and applicable to real-world scenarios.

1.3.4 Task 5 - Training and evaluation state-of-the-art models in new dataset

This task involves training and evaluating state-of-the-art models on the newly developed dataset to assess their performance in detecting plastics in MSW sorting lines. By applying these models to the updated dataset, it is possible to obtain more realistic insights into their capabilities, supporting the development of more effective solutions for waste sorting scenarios at MBT facilities.

1.3.5 Task 3 and Task 6 - Writing of the master's dissertation

Finally, the writing of the master's dissertation is conducted in two distinct stages. Task 6 represents the final phase of the research project, during which the research objectives, methodologies adopted, and results obtained will be synthesized into a dissertation.

1.4 Document Organization

The present document is structured into five chapters:

- Chapter 1 - **Introduction** - provides the context and motivation for undertaking this work, outlining the main objectives and potential contributions;
- Chapter 2 - **Related Work** - offers a comprehensive review of state-of-the-art object detection models. This chapter is divided into three subsections: two-stage models, one-stage models, and transformer-based models. Furthermore, the chapter explores

the state-of-the-art, specifically in the task of waste detection, and describes the available datasets used in the field;

- Chapter 3 - **SortWaste Dataset** - describes the dataset built and used in this project, including the collection and annotation process. In addition, some statistics from the dataset are presented;
- Chapter 4 - **Experiments and Results** - reports the experiments and results obtained, emphasizing the implementation, testing, and analysis of the results. It aims to provide a practical perspective on the development of the work;
- Chapter 5 - **Conclusion** - summarizes the work carried out, highlighting the main conclusions. It also presents the limitations identified and the possible impacts resulting from this work.

Chapter 2

Related Work

2.1 Introduction

This chapter provides a comprehensive overview of methods and techniques used in object and waste detection. Section 2.2 begins with an analysis of the main architectures used for object detection, which is divided into three sections. Section 2.3 presents the approaches and models developed for waste detection, focusing on the challenges of this task. Finally, in section 2.4, some datasets developed for waste detection will be presented.

2.2 Object Detection Models

Object detection is the process of identifying and locating objects of interest within an image or video. This task involves determining these objects' position and boundaries and classifying them into distinct categories. The most advanced methodologies in this domain can be divided into two primary types: one-stage and two-stage detectors.

2.2.1 Two-Stage Object Detection Models

Two-stage models take a more refined approach to object detection by dividing the process into two steps. In the first stage, they generate region proposals that may contain objects, and in the second stage, each proposed region is classified, and its boundaries are adjusted for greater precision. This two-step approach results in better performance, particularly in complex or crowded scenes, but tends to demand more computing power and time. As a result, two-stage models are better suited for tasks where getting the most precise results is more important than speed.

2.2.1.1 Faster R-CNN

The R-CNN [11] family revolutionized object detection by improving speed, accuracy, and efficiency. Regions with Convolutional Neural Networks (R-CNN) [11] used selective search for region proposals and deep networks for classification, achieving significant accuracy gains over traditional methods. However, its dependence on precomputed proposals and individual forward passes for each region made it computationally expensive and unsuitable for real-time use.
Fast R-CNN [12] optimized R-CNN [11] by computing feature maps once per image and using Region of Interesting (RoI) pooling for proposals, reducing computation time and enabling

end-to-end training. However, reliance on external region proposal methods like Selective Search remained a bottleneck.

Faster-RCNN [1], proposed by Ren *et al.*, addressed this bottleneck by introducing the Region Proposal Network (RPN), a fully convolutional network that generates region proposals directly. The RPN integrates seamlessly with the detection network by sharing full-image convolutional features, enabling efficient, nearly cost-free region proposals. This fully convolutional network utilizes anchor boxes, pre-defined bounding boxes of different sizes and aspect ratios, to detect objects at various scales and shapes and predicts object boundaries and objectness scores at each location. These anchor boxes act as references for generating region proposals, enabling the RPN to handle objects of varying dimensions effectively.

Trained end-to-end, the RPN produces high-quality region proposals, which are subsequently used by Fast R-CNN [12] for detection. By sharing convolutional features, the RPN and Fast R-CNN [12] are unified into a single network, known as Faster R-CNN [1].

The architecture of Faster R-CNN [1], as illustrated in Figure 2.1, consists of multiple stages. The initial stage incorporates convolutional layers to extract feature maps, which capture critical spatial and semantic information from the input image. Afterward, the RPN is an attention mechanism that generates regional proposals. These proposals represent potential bounding boxes within which objects may be situated. The region proposals are subsequently aligned and pooled through RoI pooling, resulting in fixed-size feature representations. Ultimately, these pooled features are forwarded through a classifier, which predicts the object classes and refines the coordinates of the bounding boxes.

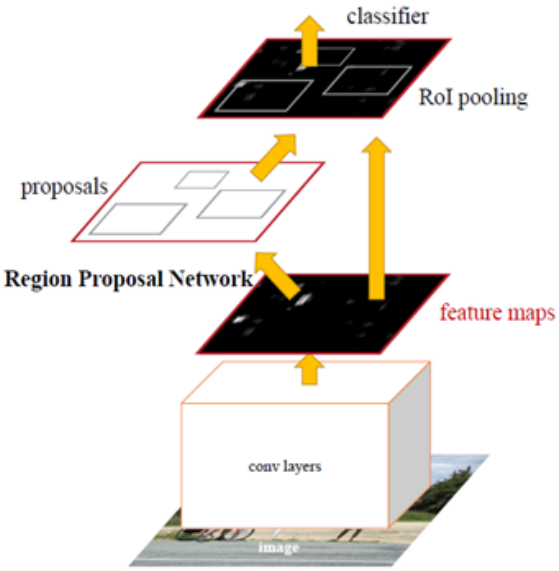


Figure 2.1: Faster-RCNN architecture for object detection [1].

Therefore, the Faster R-CNN [1] enhances both speed and efficiency while sustaining elevated detection accuracy. The incorporation of this layer, along with the addition of anchor boxes, improves the capability to identify objects of various sizes and shapes. As a result, it emerges as a versatile and robust model for object detection, proficient in adapting to many additional scenarios.

2.2.1.2 Mask R-CNN

Introduced by He *et al.*, Mask R-CNN [2] is a groundbreaking framework designed to tackle the challenges of instance segmentation by seamlessly integrating object detection and semantic segmentation.
Mask R-CNN [2] follows a two-stage architecture similar to Faster R-CNN [1]. In the second stage, Mask R-CNN[2] introduces a dedicated branch for mask prediction. As shown in Figure 2.2, this architecture incorporates three parallel heads that simultaneously predict the object class, refine the bounding box coordinates, and generate a binary mask for each RoI.

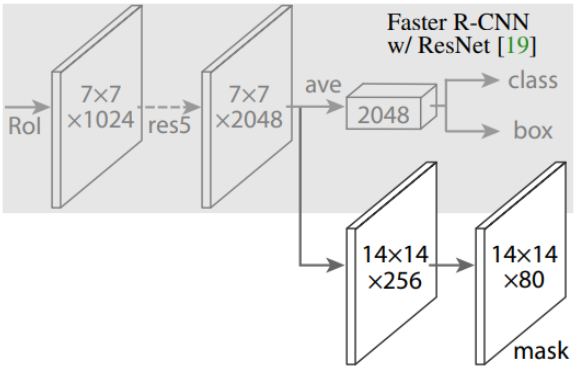


Figure 2.2: Mask R-CNN [2] architecture with three parallel heads: one for classification, one for bounding box regression, and one for mask prediction.

During the train, a multi-task loss function is employed for each RoI, $L = L_{cls} + L_{box} + L_{mask}$. L_{cls} represents the classification loss, which ensures accurate object class prediction. L_{box} corresponds to the bounding box regression loss, optimizing the localization of the object within the image. L_{mask} is the mask loss, which is a pixel-wise binary cross-entropy loss that guides the precise segmentation of the object. This multi-task loss function jointly optimizes three objectives: object classification, bounding box regression, and mask generation.
One key feature of this model, unlike its predecessor, Faster R-CNN [1], is the introduction of RoIAlign to replace the RoIPooL. RoIPool often suffers from quantization errors because it discretizes the coordinates of the proposed RoIs into fixed grid cells, leading to misalignment between the features extracted and the original image. This new layer, RoIAlign, resolves this issue by avoiding frequent quantization errors when aligning region-of-interest features. It ensures a more accurate match between the input image and the network's output, leading to better segmentation results.
The flexible framework supports various backbone architectures for feature extraction, including ResNet [13] and Feature Pyramid Networks (FPN). This adaptability allows Mask R-CNN[2] to achieve state-of-the-art results across diverse tasks. Notably, it has set benchmarks in the COCO challenges, for instance segmentation, object detection, and keypoint detection.
Mask R-CNN[2] is a significant step forward in object detection and instance segmentation.

It enhances Faster R-CNN[1] by adding the ability to precisely segment objects, offering better feature alignment, and using a multitask learning approach.

2.2.1.3 Trident Network (TridentNet)

In 2019, the Trident Network (TridentNet) [3], introduced by Li *et al.*, made a breakthrough in tackling the issue of scale variation in object detection. This challenge had previously limited the effectiveness of many models, including popular two-stage detectors like Faster R-CNN [1] and Mask R-CNN [2]. Although these models demonstrated notable success, they encountered challenges with objects of various sizes. To address this, they frequently depended on methods such as feature or image pyramids, which raised computational expenses and caused inconsistencies in feature representation across different scales.
TridentNet [3] enhances the capabilities of these two-stage models by implementing a parallel multi-branch architecture that can adapt to objects of varying scales without the requirement for external images or feature pyramids. This model incorporates trident blocks comprising multiple parallel branches that employ dilated convolutions with different rates across the branches. This design enables each branch to focus on a unique receptive field using the same transformation parameters. Thus, this consistency ensures that feature representation remains uniform across different scales. This approach helps overcome a limitation of feature pyramids in models like Faster R-CNN[1].

Figure 2.3: Proposed TridentNet architecture [3].

This architecture uses a scale-aware training method, offering a novel way to increase each branch's scale awareness and prevent the training of objects of extreme scales on mismatched branches. For each branch, i is defined a valid range $[l_i, u_i]$, where only object proposals that satisfy $l_i \leq \sqrt{wh} \leq u_i$ (with w and h representing the object's width and height) are selected during training. This ensures that each branch focuses on objects of appropriate scales, improving alignment between object size and the network's receptive field.
Rather than treating all objects the same size, the model uses different branches, as illustrated in Figure 2.3, to detect objects of specific sizes, which helps ensure the model can better match the objects with the areas it looks at, improving accuracy. Additionally, TridentNet [3] makes the model more efficient by sharing the same convolutional weights across all branches but with different dilation rates, reducing the number of parameters, which helps

prevent overfitting and allows the model to work well with objects of different sizes.
During inference, TridentNet[3] integrates results from all branches using Non-maximum Suppression (NMS) to generate final detections. This approach provides a notable advantage over previous models [2][1], which often require multi-scale testing or complex feature integration to achieve improved accuracy. Furthermore, TridentNet [3] ensures computational efficiency during evaluation, making it a convincing successor to traditional two-stage models for object detection tasks.

2.2.2 One-Stage Object Detection Models

One-stage models are an evolution of two-stage models designed to simplify the object detection process by locating and classifying objects in a single step over the image. These models allow faster processing, which makes them suitable for real-time, where quick detection is critical. Typically, the models are optimized speed, sometimes exhibiting a trade-off in accuracy, particularly in overlapping objects. Nonetheless, one-stage models offer an efficient solution for scenarios where rapid and reasonably accurate detection is essential.

2.2.2.1 Focal Loss for Dense Object Detection

Lin *et al.* introduce Focal Loss[4], a novel loss function designed to address the extreme foreground-background class imbalance that hampers the performance of one-stage object detectors. Unlike traditional two-stage models, which rely on a proposal-driven approach to select candidate locations for classification, one-stage detectors evaluate dense grids of potential object locations. This results in many easy negatives dominating the loss during training, leading to inefficient learning and suboptimal accuracy.
Focal Loss [4] modifies the standard cross-entropy loss by introducing a scaling factor that down-weights well-classified examples, thereby emphasizing harder, misclassified examples. The loss function is defined as:

$$FL(p_t) = -(1 - p_t)^{\gamma} log(p_t),$$

where p_t represents the predicted probability for the correct class (p_t = p if y = 1, and p_t = 1-p if y = 0) and γ is a tunable focusing parameter that adjusts the rate at which well-classified examples are down-weighted.
The factor $(1-p_t)^{\gamma}$ is the key component differentiating this loss from standard cross-entropy loss. It down-weights the loss contribution of well-classified examples (where p_t is close to 1), making the loss function focus more on harder, misclassified examples (where p_t is small). This effectively mitigates the impact of class imbalance without relying on heuristic sampling methods, such as hard example mining, which can be computationally expensive and prone to overfitting. The γ controls how much the easy examples are down-weighted; higher values of γ emphasize hard examples, whereas lower values result in behavior closer to standard cross-entropy loss.
The efficacy of Focal Loss[4] was demonstrated with RetinaNet, a simple one-stage detector.

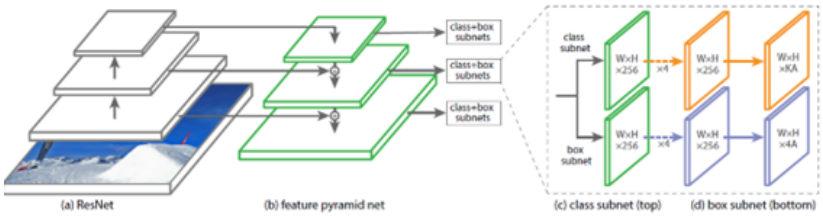


Figure 2.4: RetinaNet architecture [4].

RetinaNet, illustrated in Figure 2.4, is composed of a backbone and two task-specific subnetworks. The architecture employs a FPN backbone built on top of a feedforward ResNet [13] architecture to generate a convolutional feature pyramid. To this backbone, RetinaNet attaches two subnetworks: one for classifying anchor boxes and another for regressing from anchor boxes to ground-truth objects.
When combined with Focal Loss[4], RetinaNet achieved state-of-the-art accuracy on the COCO benchmark, surpassing two-stage detectors while maintaining competitive inference speeds. So, this loss makes it easier to deal with class imbalance. While two-stage models usually tackle this problem with a complex series of region proposals and biased sampling in the second stage, Focal Loss [4] simplifies things by being built into the one-stage training process, making the model more straightforward.

2.2.2.2 YOLO: You Only Look Once

Since its launch, the You Only Look Once (YOLO) family of object detection models has evolved steadily. With each new version, improvements make the models faster, more accurate, and more adaptable. YOLO simplified object detection into a single-step process, paving the way for real-time detection.
Starting with the foundational YOLOv1 [14], which is the first in the YOLO family, subsequent versions have progressively introduced architectural and methodological advancements. These include multi-scale detection, novel activation functions, anchor-free detection mechanisms, and reparameterization techniques, as YOLOv3 [15] at YOLOv10 [16].

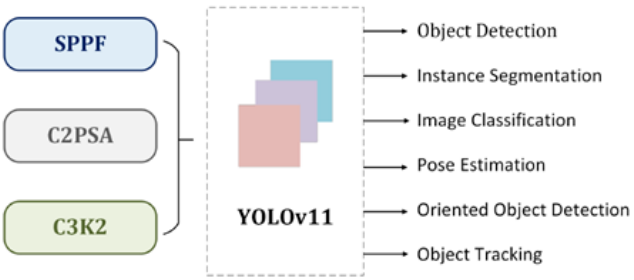


Figure 2.5: Architectural modules in YOLOv11 [5].

The latest iteration, YOLOv11 [5], introduced in 2024, represents a major advancement in object detection. While maintaining the core structure of the backbone, neck, and head, YOLOv11 [5] introduces key innovations, as shown in Figure 2.5, the C3k2 block for efficient

feature extraction, the Spatial Pyramid Pooling-Fast (SPPF) block, and Cross Stage Partial with Spatial Attention (C2PSA) mechanism. These enhancements make the model more efficient and accurate in object detection. YOLOv11 [5] has been improved to do more than just object detection; it can also handle tasks like instance segmentation, pose estimation, and recognizing objects from different angles. This versatility makes YOLOv11[5] an extremely adaptable solution for various applications.
The YOLO series has consistently progressed, leaving a significant mark on computer vision. Each iteration extends the boundaries of real-time object detection, with YOLOv11 [5] propelling this progression. It merges exceptional accuracy, speed, and adaptability, solidifying its status as a foundational element for intricate visual recognition challenges.

2.2.3 Transformer-Based Models

Transformer-based models have brought a paradigm shift in the field of object detection. Unlike traditional Convolutional Neural Network (CNN)-based architectures, which rely on region proposals and post-processing steps such as non-maximum suppression, transformer-based approaches leverage the self-attention mechanism to capture global dependencies across the entire image.

2.2.3.1 DETR: End-to-End Object Detection with Transformers

In ”End-to-End Object Detection with Transformers” [6], Carion *et al.* present DEtection TRansformer (DETR), a groundbreaking approach to object detection that redefines the task as a direct set prediction problem. Traditional object detection methods rely on intermediate steps such as anchor generation, region proposal refinement, and non-maximum suppression to identify bounding boxes and classify objects. These steps embed task-specific heuristics and require meticulous tuning, resulting in complex and less generalizable pipelines. DETR [6] departs from this paradigm by introducing a streamlined, end-to-end architecture that eliminates hand-crafted components.

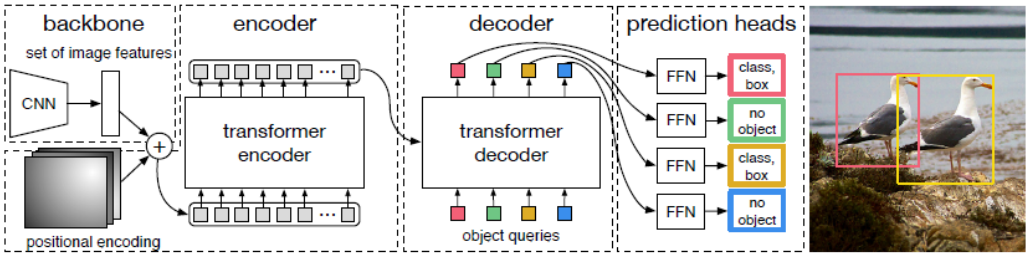


Figure 2.6: DETR architecture [6].

The DETR[6] architecture, depicted in Figure 2.6, combines a CNN backbone, a transformer encoder-decoder, and Feed-Forward Networks (FFNs) to predict object classes and bounding boxes directly. The backbone is a convolutional neural network that extracts feature maps from the input image. These feature maps are combined with positional encodings to retain

spatial information. The transformer encoder uses self-attention to model global dependencies in the feature maps, creating a representation that captures both local and global context. The transformer decoder employs a fixed set of learnable object queries to interact with the encoder output via cross-attention, focusing on regions of interest for object detection. The prediction FFNs generate the final outputs for each object query: the class branch predicts the object class or ”no object”, while the box branch predicts the bounding box coordinates. One of DETR’s key innovations is its bipartite matching loss, which ensures a one-to-one correspondence between predicted objects and ground truth annotations. This is achieved using the Hungarian algorithm, which computes the optimal assignment by considering both class probabilities and bounding box similarity. By enforcing permutation invariance in predictions, DETR [6] eliminates the need for post-processing techniques like NMS.
DETR [6] achieves results comparable to Faster R-CNN [1]. Thanks to its ability to capture global context using self-attention, this model excels at detecting large objects but struggles with smaller ones. Nevertheless, with the complex transformer-based components, the model demands more training time and greater computational power. Despite these, DETR [6] marks a significant advancement in object detection.

2.3 Waste Detection and Management

Waste detection and management have become critical areas of research, driven by the need to address growing environmental concerns and optimize resource recovery. Recent advancements have focused on innovative technologies facilitating efficient waste monitoring and sorting.

2.3.1 Urban Waste Detection

In [17], the paper discusses a novel real-time object detection system designed to identify and report abandoned waste in urban and suburban areas. The system employs the YOLOv3 architecture and a custom dataset created using the Google Images Download tool. The results indicate that the proposed approach performs well in real-time waste detection. The model can also recognize garbage bins and verify whether the areas they are located in need cleaning. Similarly, in India, [18] demonstrates the capability to perform real-time detection and classification of waste in public areas and unauthorized dumping sites by utilizing the YOLOv8 architecture, which is trained on the TrashNet dataset. The system enables rapid alerts to be sent to local authorities, thereby facilitating timely intervention and effective waste management. These works could contribute to more efficient waste management in smart cities.
The Faster R-CNN framework, enhanced with ResNet for automatic garbage detection in urban environments, was proposed in [19]. This model was trained on a dataset comprising urban images containing garbage. The study introduced a data fusion and augmentation strategy to achieve high detection accuracy, which involved combining images of urban

scenes with and without garbage. The results showed that this method possesses strong generalization and detection capabilities, making it suitable for intelligent urban management applications.

The study [20] demonstrated an enhanced YOLOv4 model for real-time trash segregation in urban environments. The model incorporates a refined network structure with Spatial Pyramid Pooling (SPP) and a Path Aggregation Network (PANet) to improve feature extraction and object detection. The results show that this improved model outperforms the baseline and can generalize to different types of waste. This system, similar to the approach proposed in [21], where a robot detects trash with high accuracy, highlights the potential of AI-powered solutions to enhance waste management efficiency.

2.3.2 Water Waste Detection

There are systems specifically designed to address the challenges of waste detection in underwater environments. These systems are equipped to identify and classify various types of trash submerged in aquatic settings, ranging from plastics and metals to organic waste.

In [22], the use of deep learning to develop a model for real-time underwater trash detection is explored to assist Autonomous Underwater Vehicles (AUVs) in this task. This research employs four state-of-the-art object detection algorithms: YOLOv2, Tiny-YOLO, Faster R-CNN, and SSD, which are trained on the JAMSTEC E-Library of Deep-sea Images (J-EDI) dataset [23]. The study demonstrates that deep learning-based detection models can accurately identify marine litter in controlled environments. Detection accuracy varies depending on the type of litter and environmental conditions, with plastics being the material most successfully identified. This paper highlights the potential of combining robotics with deep learning to address the growing issue of marine litter.

The study [24] focuses on detecting visible trash on urban water surfaces, addressing environmental issues such as contamination and blockages in water channels. The authors propose an attention-based neural network, incorporating a novel attention layer to enhance the detection of small objects commonly found in such settings. A dataset with object-level annotations of trash in water channels is introduced, one of the first datasets tailored for this application. The proposed model outperforms state-of-the-art object detectors, particularly in identifying smaller objects.

In [25], the authors address the pressing issue of water pollution by developing a deep learning-based model, AquaVision, to detect and classify pollutants in oceans and seashores. They introduce a dataset named AquaTrash, derived from the Trash Annotations in Context (TACO) [26] dataset, containing images of various waste items in aquatic environments. AquaVision employs a RetinaNet architecture with a ResNet-50 backbone and FPN to achieve a mAP of 0.8148 in identifying waste objects. This approach facilitates the localization of waste, aiding in the cleanup of water bodies and contributing to the maintenance of aquatic ecosystems.

2.3.3 Domestic Waste Detection

Based on YOLOv3, [27] presents a real-time model for detecting domestic waste using a region-specific dataset called the Taiwan Recycled Waste Dataset (TRWD). Unlike the commonly used TrashNet[28] dataset, which contains single waste objects, TRWD includes images with multiple objects and is tailored to Taiwan's unique waste characteristics. The study highlights the importance of region-specific datasets for optimizing object detection models in waste management.

An enhanced version of YOLOv3, integrating dense convolutional blocks and skip connections for superior domestic garbage detection in diverse and complex environments, named Skip-YOLO, is proposed in [29]. The model improves feature extraction by combining shallow and deep semantic information, addressing challenges such as recognizing objects with similar characteristics and detecting multiple waste types in cluttered scenes. This model achieves better performance than the baseline YOLOv3.

To address the challenges of kitchen waste detection in industrial environments, [30] proposes AL-DETR. This novel object detection model combines active learning with transformer-based detection (DETR). The model selectively annotates the most informative samples from unlabeled data, reducing labeling costs while maintaining high accuracy. Trained on a custom dataset, AL-DETR outperformed state-of-the-art models, particularly in detecting large objects. A robotic sorting system was also implemented, achieving successful results in real-world tests. This approach demonstrates strong potential for scalable and cost-effective waste management automation.

2.3.4 Waste Detection

The study [31] addresses the lack of large datasets by augmenting a custom dataset with TrashNet [28] images and generating synthetic waste piles for training. The model used, Faster R-CNN, automates waste categorization into three categories: landfill, recycling, and paper. The model achieved a mAP of 0.683, demonstrating its effectiveness in detecting and classifying waste.

In research [32], an enhanced Single Shot Multibox Detector (SSD) called L-SSD has been proposed to improve garbage detection efficiency. Key innovations include a lightweight feature fusion module that combines features from different layers into a more refined feature map, improving the detection of various sizes of objects. This approach was evaluated on a custom dataset designed for garbage detection and provides an effective solution for intelligent waste classification in real-world applications, reducing manual sorting burdens and improving waste management systems.

A novel deep learning model designed for real-time trash detection and classification is proposed in [33]. Built on YOLOv4, YOLO-Green introduces architectural optimizations, including a fire module and upsampling/downsampling techniques, to enhance accuracy and efficiency while reducing model complexity. This model was trained on the TrashX and TrashNet datasets with seven waste categories, outperforming YOLOv3, YOLOv4, and other popular models.

In [34], the authors propose a hierarchical deep-learning approach for waste detection and classification in food trays. The model combines Faster R-CNN[1] for bounding box generation with a second-stage CNN to classify waste into material-based and shape-based categories. This study uses a custom dataset, Labeled Waste in the Wild (LWW), which contains 1,000 images of waste in food trays under real-world conditions.

More recently, in 2024, two significant improvements to YOLO for waste detection have been proposed. The first, DSYOLO-Trash [35], is built upon YOLOv5 and incorporates dual attention mechanisms— Convolutional Block Attention Module (CBAM) and Contextual Transformer Networks (CotNet) —to enhance the extraction of channel and spatial attention features. These mechanisms improve the model's ability to detect small, partially obscured, and overlapping waste objects in complex environments, effectively addressing key challenges in waste detection. The paper also introduces a custom dataset comprising 2,332 multi-label images of mixed waste; however, this dataset is unavailable.

The second, MRS-YOLO [36], utilizes YOLOv8 to detect small and complex waste objects, similar to before. The model introduces several innovative techniques, including the Slide-Loss_IOU to prioritize small object detection, a channel pooling with dynamic convolution module for improved feature extraction, and a RepViT Transformer mechanism for more efficient feature utilization. Both studies represent important advancements in state-of-the-art automated waste detection and classification.

The study [37] focuses on identifying contamination in real-world waste collection environments with cluttered scenes, such as inside garbage trucks. A custom dataset was developed using video footage collected across commercial and residential areas in the US and EU. Several state-of-the-art models were trained, with YOLOv8-x utilizing transfer learning, which achieved the best performance. The system demonstrates strong potential for automating contamination detection in complex settings, helping to improve recycling rates and operational efficiency.

These studies contribute to increasing recycling efficiency, reducing manual labor and supporting global sustainability objectives and demonstrate that waste detection is a topical area of research.

2.4 Datasets for Waste Detection

Detecting waste objects has become a relevant research topic, especially in response to rising pollution levels and growing global concern. This topic interests researchers who are looking for technological solutions to mitigate the impacts of waste on the environment. In this context, datasets have been developed to support the development of systems capable of identifying waste based on images.

These datasets usually contain images representing different types of waste, allowing deep learning models to be trained for waste detection tasks. However, there are significant challenges associated with using these datasets. One of them is the influence of the background of the images: unstable backgrounds, which include complex scenes with other objects, make it challenging to identify waste accurately. On the other hand, datasets that use uniform

backgrounds, such as a white background, do not reflect real-world conditions, limiting the models' performance when applied to real situations.
Therefore, to detect residues effectively in varied environments, the quality and representativeness of the datasets used must be considered. Below are some of the main datasets available.

2.4.1 Trash Datasets

TrashNet [28] is a dataset widely used in waste detection. It consists of 2,400 images divided into six categories: metal, paper, plastic, glass, organic waste, and other waste. The images were captured against a white background in different lighting conditions. Despite offering some standardization for the models, this dataset lacks diversity in terms of scenarios since all the photographs were taken in controlled environments.
Garbage in Photos (GINI) [38] is a dataset that contains images of a single class, representing garbage in general. It was created using a Bing image search API and totals around 1,400 images. WADE-AI[39], similar to GINI, includes images collected from Google Street View, offering a diverse set but limited to a single generic waste classification.
Another relevant dataset is TACO [26], designed specifically for waste detection and segmentation tasks. It comprises 1,500 images captured by mobile devices and distributed across 28 categories. TACO presents a variety of contexts and scenarios, such as streets, beaches, parks, and other outdoor environments.
Waste Pictures [40] is a dataset built through Google searches. It contains around 24,000 images and is organized into 34 classes, representing a variety of waste types.
OpenLitterMap [41] is one of the largest datasets available, with over 100k images captured by mobile device cameras. Users from all over the world submitted the images, providing a huge diversity of scenery and photography styles.
There are also specific datasets. The WaDaBa Dataset [42] is a specialized dataset that focuses exclusively on plastic waste collected in domestic environments. It contains photographs of 100 different plastic objects, each photographed 40 times in different conditions and types of lighting. This dataset is particularly useful for research exploring the identification of plastic waste in controlled environments.
Figure 2.7 provides examples of images from several of the datasets discussed.

2.4.1.1 ZeroWaste Dataset

Previously available datasets were limited and did not accurately reflect reality, which made it challenging to develop effective solutions for industrial environments.
ZeroWaste [7] is a dataset created to address the lack of quality datasets in real environments, aiming to train and evaluate classification and segmentation algorithms to classify industrial waste.
Therefore, this dataset is developed to facilitate the study of waste detection and segmentation in industry, including cluttered scenarios and deformable and translucent objects, as il-

Figure 2.7: Examples of dataset images, grouped by columns. Each column contains three images from the same dataset, illustrating variations within each category.

lustrated in Figure 2.8. This comprehensive dataset can be used for various types of learning. It is divided into three distinct parts: ZeroWaste-f for fully supervised detection, ZeroWaste-s for semi-supervised learning, and ZeroWaste-w, which includes images before and after object removal for weakly supervised learning.
Detecting deformable, translucent, and disordered objects in this dataset represents a unique challenge in computer vision, highlighting the relevance of ZeroWaste [7] to advances in this area.

Figure 2.8: Examples of images from ZeroWaste Dataset [7].

2.5 Summary

This chapter began with a comprehensive analysis of the methods used for object detection, which can be divided into three main categories: two-stage, one-stage, and transformer-based models. Next, a detailed review of the state-of-the-art in waste detection was presented, with the approach organized into four groups: systems developed for detecting waste in urban areas, systems for detecting waste in aquatic environments, systems for detecting

Dataset	Images	Classes	Task	Method	Application
TACO [26]	1500	28	Segmentation	Mask R-CNN	Litter detection
TrashNet [28]	2400	6	Classification	SVM	Recycling waste
WaDaBa [42]	4000	6	Classification	-	Recycling waste
WADE - AI [39]	1396	1	Detection + Segmentation	Mask R-CNN	Litter detection
LWW [34]	1002	19	Detection	Faster R-CNN	Recycling waste
OpenLitterMap [41]	100.000+	100+	Classification	-	Litter detection
J-EDI [23]	5720	3	Detection	YOLOv2, Tiny YOLO, Faster R-CNN, SDD	Marine debris
AquaTrash [25]	369	4	Detection	Faster R-CNN, RetinaNet	Marine debris
TRWD [27]	6233	6	Detection	YOLOv3	Waste detection
Waste Pictures [40]	23633	34	Classification	-	Trash objects
GINI [38]	1400	1	Classification	-	Litter detection
ZeroWaste - f [7]	4503	4	Detection + Segmentation	RetinaNet, Mask R-CNN, TridentNet	Industrial recycling waste
Proposed Dataset	5261	8	Detection	Faster R-CNN, RetinaNet, TridentNet, YOLOv11	Industrial recycling waste

Table 2.1: Comparison of public waste datasets.

domestic waste, and, finally, systems that do not fit into any of the previous groups. Subsequently, some existing datasets on this subject were discussed.

Although several interesting waste detection approaches exist, none focus specifically on waste detection in MBT environment. This project, therefore, aims to be part of a new group of waste detection systems aimed at the industrial environment. The industry presents specific challenges, such as overlapping, damaged, or dented objects, which make the detection task more complex. In addition, except for ZeroWaste [7], existing datasets for waste detection are not developed for industry, as shown in Table 2.1. They are mainly composed of images obtained from Google or captured by smartphones and do not accurately reflect the real-world conditions in the industry. Many of these datasets contain images with only one object on a white background, which limits the applicability of these models in real-world scenarios.

Therefore, creating more representative datasets that capture the complexity and diversity of conditions found in industrial environments is crucial for developing waste detection models suitable for this context. This dissertation aims to fill this gap and contribute to the evolution of waste detection systems adapted to the specificities of the industrial environment.

Chapter 3

SortWaste Dataset

3.1 Introduction

This chapter introduces SortWaste, the dataset developed and utilized in this dissertation. It describes the data collection process, the annotation methodology, and the preprocessing steps applied to prepare the dataset for training and evaluation. It also provides statistical insights into the dataset and includes a comparative analysis with the state-of-the-art ZeroWaste-f [7] dataset.

The creation of SortWaste was crucial to achieving the goals of this work, as no publicly available datasets adequately reflected the specific requirements and real-world conditions relevant to the Portuguese context. So, SortWaste contributes a valuable resource to the research community in the field of waste detection.

3.2 Data Collection

The data used in constructing this dataset were collected at an MBT facility in Portugal. This facility receives MSW, the contents of common containers, typically green or grey. Although it is intended for unsorted waste, recyclable materials are often present, as not everyone separates their waste properly at source.

The manual sorting on this line aims to identify and remove items that can still be recovered, such as different types of plastics, before the remaining waste is sent to landfills. A nonintrusive approach was used to collect the data, ensuring that the normal operation of the sorting line and the workers' activities were not disrupted.

To collect the videos, a tripod with a smartphone (iPhone 14) was mounted on the side of the sorting line to capture the conveyor belt from a top-down perspective, as illustrated in Figure 3.1. The smartphone was positioned approximately 100 cm above the conveyor belt, ensuring the framing of the area of interest. The existing lighting conditions in the facility were stable and adequate, making the use of additional light sources unnecessary.

A total of 120 minutes of video footage was recorded, with a resolution of 1920×1080 pixels at 60 frames per second. The recording took place at the beginning of the sorting line, before any direct human intervention with the waste. Nevertheless, it is essential to note that various mechanical and automated processes occur between the arrival of the waste and the point of manual sorting.

This methodology enabled the collection of data representative of the real waste sorting environment while ensuring the integrity of operations and compliance with safety and privacy regulations on-site.

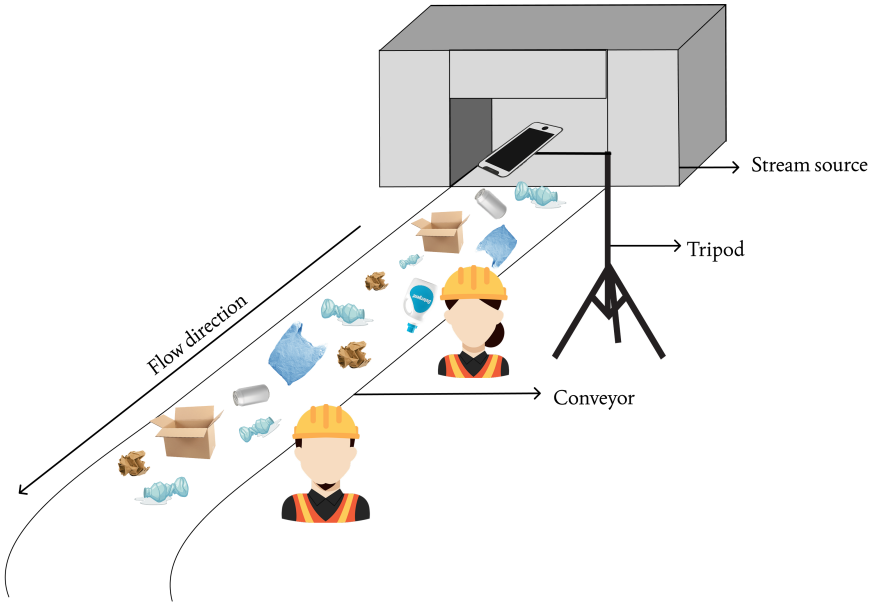


Figure 3.1: Illustration of the data collection setup, showing the arrangement of waste on the conveyor belt, the flow direction, the position of the workers, and the image capture system using a smartphone on a tripod.

3.3 Annotation and Preprocessing

Although a total of 120 minutes of video was collected, as previously mentioned, only 18 minutes were annotated for this project due to the time-consuming nature of the manual annotation process. To reduce redundancy and streamline this process, the original frame rate of 60 frames per second was downsampled to 5 frames per second, significantly decreasing the number of frames requiring annotation while preserving the variability and representativeness of the data.

The annotation of the selected frames was performed manually using the open-source tool Computer Vision Annotation Tool (CVAT) [43], a widely used framework for labeling data in computer vision tasks. All annotations were carried out solely by the author of this work over approximately two months. On average, each frame required approximately six minutes to annotate, depending on its complexity and the number of objects present. This careful annotation process was crucial in ensuring the accuracy and quality of the labeled dataset, which serves as a fundamental component for training and evaluating state-of-the-art models in the context of plastic detection in MSW.

Eight types of materials were annotated, based on the categories of packaging waste defined in Despacho No. 15370/2008 [44]:

- **Polyethylene Terephthalate (PET):** Rigid, transparent or green objects, usually bottles, jars, and other containers previously used to package water, soft drinks, or other beverages.
- **High-Density Polyethylene (HDPE):** Opaque, colored objects that are less flexible

and denser, such as yogurt cups, bottles, and jars used for food products, hygiene items, detergents, fabric softeners, or alcohol.

- **Liquid Food Carton Packaging (ECAL):** Multilayer packaging composed of at least 75% cardboard, intended for containing liquid foods (e.g., milk, juice).
- **PET Oil:** PET containers specifically used for packaging edible oils. Although this subcategory is not defined as an independent category in the Despacho, it can be considered a subdivision of PET due to its typical contamination.
- **Mixed Soft Plastic:** Flexible and compressible plastics, such as cookie wrappers, potato chip bags, and plastic bags.
- **Mixed Rigid Plastic:** Rigid plastics that do not fall under the HDPE category, often transparent, such as molded packaging, boxes, and other hard containers.
- **Cardboard:** Corrugated or flat cardboard packaging used for storing, transporting, and distributing products.
- **Metal:** Metallic packaging made of steel or aluminum, such as beverage cans or food tins.

Figure 3.2: Examples of images from the SortWaste dataset.

After the annotation process, a total of 5396 frames were labeled, containing dirty, deformed, broken, and overlapping objects, as illustrated in Figure 3.2.
To create the training, validation, and test subsets while minimizing the risk of similar data appearing across different partitions, the dataset was first grouped into scenes of 200 consecutive frames. For each scene, the first five frames were removed to reduce temporal redundancy and ensure a more precise separation between adjacent scenes.
Following this preprocessing step, the scenes were grouped to approximate a 70%/15%/15% split of the dataset into training, validation, and test sets, respectively. This distribution was performed to maintain a similar class distribution across subsets. However, since the number of objects per frame varies and not all classes appear uniformly, achieving exact proportions per class was not feasible. Even so, the final splits closely approximate the target percentages, ensuring similar proportions of the classes in the different sets.

Split	Images	HDPE	ECAL	PET	Mixed Soft Plastic	Mixed Rigid Plastic	Cardboard	Metal	PET Oil	Objects
Train	3705	16803	13649	11976	9077	7066	1524	945	802	61842
Validation	780	4972	2552	2108	1443	1120	425	277	168	13065
Test	776	3269	3026	2722	1817	1230	207	215	132	12618
Total	5261	25044	19227	16806	12337	9416	2156	1437	1102	87252

Table 3.1: Summary of object counts within each dataset split: training, validation, and test.

Table 3.1 presents the number of images in the dataset after its division into training, validation, and test subsets. It also details the number of annotated objects per class within each subset.

3.4 Statistics

This section presents statistical insights into the SortWaste dataset and provides a brief comparison with the only existing dataset for this purpose in the state-of-the-art.

3.4.1 SortWaste Dataset

3.4.1.1 All Classes

In Figure 3.3, the bar chart shows the total number of annotated bounding boxes for each class present in the dataset. It can be observed that the HDPE class has approximately 25000 annotations, while the ECAL and PET classes each have over 15000 boxes. On the other hand, the Cardboard, Metal, and PET Oil classes are less frequent, each with fewer than 5000 annotations.

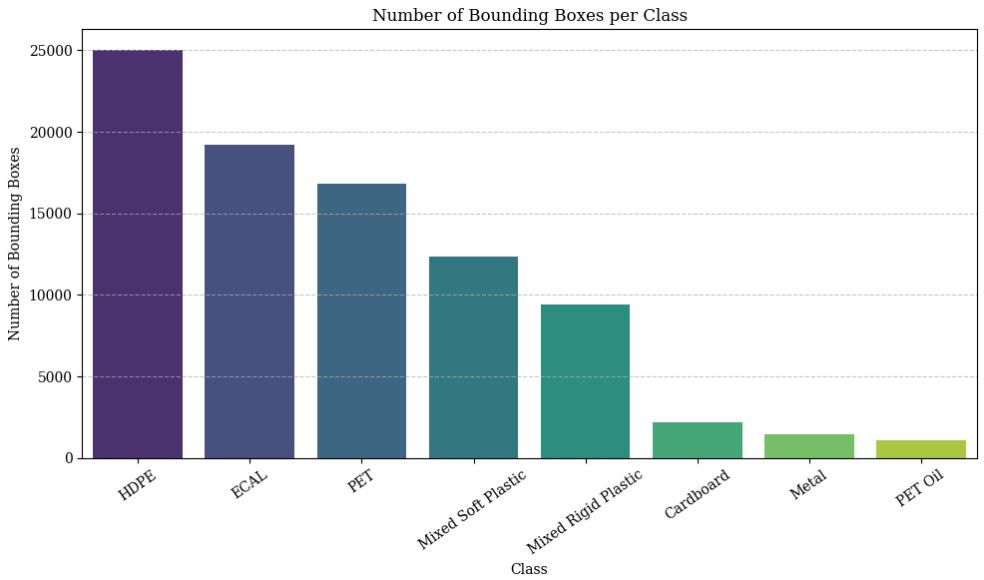


Figure 3.3: Statistics of the number of instances per class in the SortWaste dataset.

This disparity can be attributed to the operational context of the unsorted waste sorting line where the data were collected. Metallic objects are removed mainly by a magnet during the

process, reducing their occurrence in the videos. Similarly, Cardboard tends to disintegrate due to the existing moisture. As for the PET Oil class, it is a subcategory of the PET class and, therefore, naturally appears less frequently. This distribution suggests an imbalance among the classes, reflecting the actual frequency of the objects in the collected data.

3.4.1.2 Plastics Classes

As previously mentioned, the SortWaste dataset comprises eight object classes. However, to better align with the primary objective—detecting plastics in MSW sorting lines—the original classes were regrouped into four plastic categories.

The HDPE and ECAL classes were maintained without modifications. The PET class was expanded to include instances originally labeled as PET Oil, reflecting their similar material properties. A new category, Mixed Plastic, was created by merging the Soft Plastic and Rigid Plastic classes. On the other hand, the Metal and Cardboard classes were excluded from the analysis, as they are not plastic.

Figure 3.4 shows the distribution of bounding boxes per category after this regrouping. The HDPE class has the highest number of annotations, followed by Mixed Plastic, ECAL, and PET, each exceeding 18000 annotated instances. This relatively balanced distribution contributes to a representative training process for object detection models.

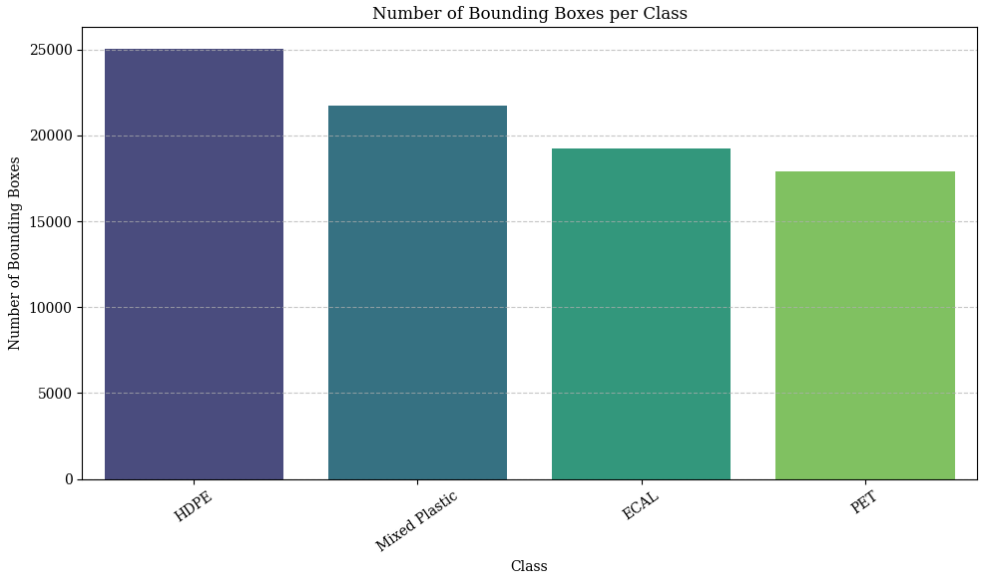


Figure 3.4: Statistics on the number of instances per plastic class in the SortWaste dataset.

3.4.1.3 ZeroWaste Classes-Matching

To enable comparative experiments with the state-of-the-art reference dataset, it was necessary to group the eight original SortWaste classes, allowing for a direct comparison between the two datasets. In this mapping, the Metal class was retained, as it is present in

both datasets. The Cardboard class was merged with ECAL, as the packaging in the latter is primarily composed of cardboard. The Mixed Soft Plastic class was renamed to Soft Plastic, retaining the same elements. The HDPE, PET, PET Oil, and Mixed Rigid Plastic classes were grouped under the new label Rigid Plastic. As shown in Figure 3.5, there is a predominance of the Rigid Plastic class, which is expected, since it was previously identified that the HDPE class (now included in this grouping) is one of the most frequent in the dataset.

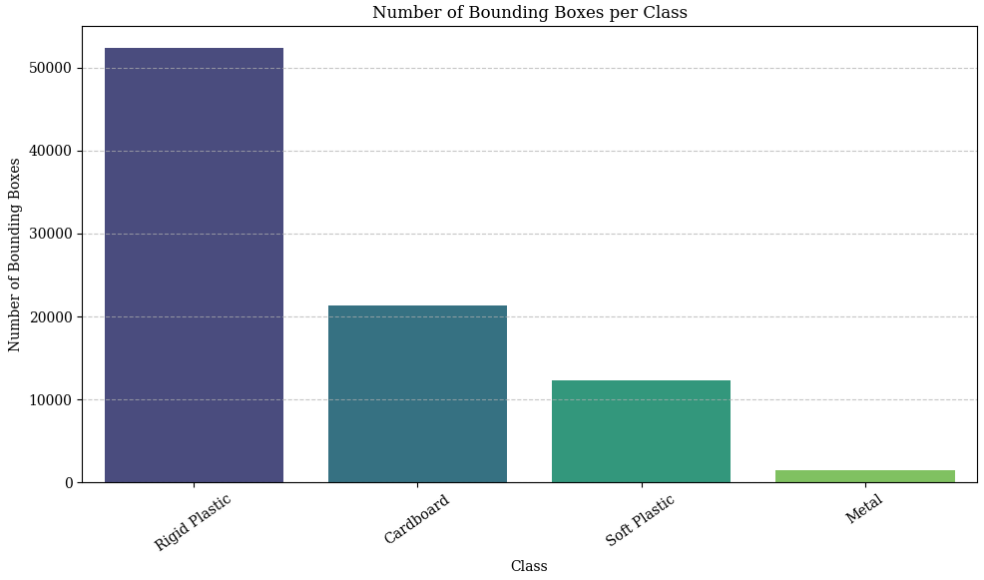


Figure 3.5: Statistics on the number of instances per class after class remapping, as in ZeroWaste.

3.4.2 ZeroWaste-f Dataset

This section presents statistics from the state-of-the-art reference dataset to establish a comparison with the dataset developed in this dissertation. It is observed, in Figure 3.6, that the predominant class in this dataset is Cardboard, which can be explained by the fact that it was constructed from a sorting line where that material was dominant. Conversely, the Metal class is the least represented — a pattern also observed in the SortWaste dataset.

Table 3.2 presents the number of images in the dataset after it has been divided into training, validation, and test subsets. It also details the number of annotated objects per class within each subset.

Split	Images	Cardboard	Soft Plastic	Rigid Plastic	Metal	Objects
Train	3002	12940	4862	1160	263	19225
Validation	572	2167	855	305	60	3387
Test	929	3428	1236	315	63	5042
Total	4503	18535	6953	1780	386	27744

Table 3.2: Summary of object counts within each dataset partition: training, validation, and test in ZeroWaste-f [7].

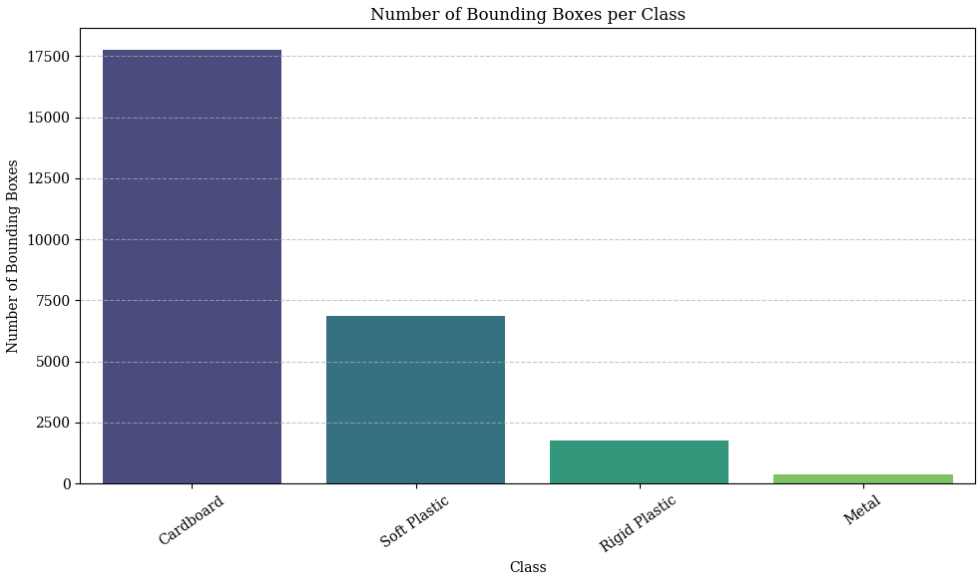


Figure 3.6: Statistics on the number of instances per class in ZeroWaste-f [7].

3.4.3 Comparison between SortWaste and ZeroWaste-f

To enable a comparison between the two datasets, Figure 3.7 and Table 3.3 are presented.

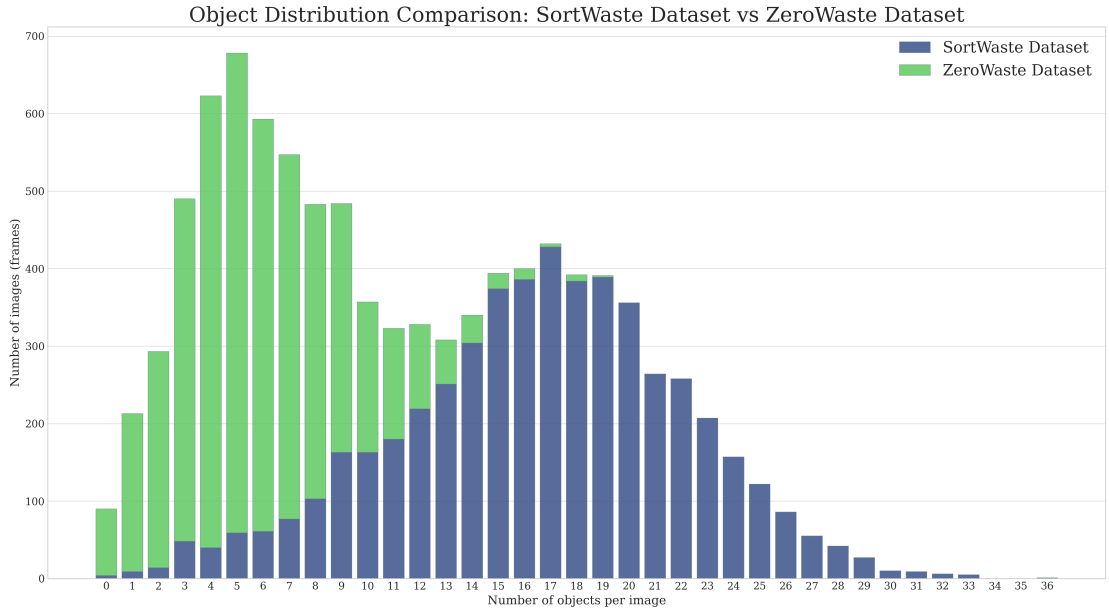


Figure 3.7: Statistics of the total number of annotated objects per frame in the SortWaste dataset and the ZeroWaste-f dataset [7].

Figure 3.7 illustrates the distribution of the number of objects per image in both the SortWaste dataset developed in this work and the publicly available ZeroWaste dataset. A clear difference in the complexity profile of the two datasets can be observed.

The ZeroWaste-f [7] set shows a distribution mainly concentrated between 2 and 10 objects

per image, with a peak in the range of 4 to 6 objects. At first glance, this distribution suggests that most images contain a small number of objects, indicating relatively simple scenes with lower visual density. However, it is important to note that, despite containing fewer annotated objects, the ZeroWaste-f [7] set exhibits visual complexity, as many images feature paper in the background, with only contamination being annotated - i.e., elements that do not correspond to high-quality paper.

In contrast, the SortWaste set exhibits a skewed distribution towards higher object counts, with most images containing between 12 and 22 objects and a peak at around 17. This pattern reflects denser scenes that more accurately represent real-world MSW environments. There is also a partial overlap between the two sets in the range of 10 to 14 objects.

These differences in object distribution and annotation policy are relevant because they can directly influence the performance and generalization capacity of machine learning models, especially in contexts with variable visual density.

Characteristic	SortWaste	ZeroWaste-f
Collection Location	Portugal	Massachusetts, USA
Sorting Line Type	Municipal solid waste	High-quality paper stream
Stream	Single	Single
Number of Classes	8	4
Number of Images	5261	4503
Total Bounding Boxes	87252	27744
Material Diversity	High – various types of plastics and waste	Low – predominantly paper

Table 3.3: Comparison between the SortWaste dataset and the ZeroWaste-f [7] dataset.

Table 3.3 presents a comparative overview between the SortWaste dataset, developed in this work, and the ZeroWaste-f [7] dataset, a commonly referenced benchmark in the waste detection domain.

The ZeroWaste-f [7] dataset was collected from a high-quality paper sorting line in Massachusetts, USA. As a result, its contents consist primarily of paper materials. A predominance of cardboard and limited material diversity characterizes the dataset. It contains 27744 annotated objects distributed across 4503 images, categorized into four main classes: Cardboard, Soft Plastic, Rigid Plastic, and Metal.

On the other hand, the SortWaste dataset was collected under real-world conditions in Portugal at the beginning of a MSW sorting line at MBT facility. This setting captures the inherent complexity and variability of real waste streams. SortWaste comprises 5261 images and a significantly higher number of annotations, 87252, spread across eight more detailed classes, including HDPE, ECAL, PET, and PET Oil, among others. This increased material diversity makes SortWaste a representative and challenging dataset for developing and evaluating object detection models in realistic, heterogeneous environments.

3.5 Conclusion

The development of the SortWaste dataset represents a key contribution to this work, addressing the lack of annotated waste datasets adapted to industrial conditions. By capturing real-world scenarios at MBT facility, SortWaste provides a realistic benchmark for waste classification systems in MSW.

Its specificity, including diverse contamination levels and complex object occlusions, offers challenges that better reflect operational environments. Compared to existing datasets such as ZeroWaste [7], SortWaste enhances diversity and contextual relevance, particularly for European waste management systems.

This dataset not only enabled the training and evaluation of models within this project but also serves as a resource for the broader research community, promoting advancements in sustainable waste processing through machine learning.

Chapter 4

Experiments and Results

4.1 Introduction

In this chapter, we present the benchmark results of experimental evaluation. Four state-of-the-art models were tested on two datasets: SortWaste, a dataset developed as part of this dissertation, and ZeroWaste-f [7], the only publicly available state-of-the-art dataset designed for industrial waste classification. We conducted four experiments to assess the performance of these models on the SortWaste dataset. All experiments were evaluated using the metrics detailed in Section 4.3.

4.2 Implementation Details

The Faster R-CNN [1], TridentNet [3], and RetinaNet [4] models were implemented using the default configurations provided by the Detectron2 framework [45], with model weights initialized from pre-training on the COCO dataset. For each model, only two hyperparameters—the learning rate and optimizer—were varied to identify the optimal configuration. The batch size was fixed at eight across all experiments. Early stopping was applied based on two criteria: a patience of 15 epochs or a maximum of 80000 iterations. In contrast, the YOLOv11 [5] model was implemented using the Ultralytics framework [46], with similar tuning limited to the learning rate and optimizer. Early stopping for YOLOv11[5] was configured with patience of 15 epochs and a maximum of 300 training epochs. The objective of these experiments was to determine the best-performing configuration for each model. The results presented correspond to the highest performance achieved by each model on the test set.

4.3 Evaluation Metrics

Relevant metrics were chosen to evaluate the models' performance in object detection tasks. The main metrics used are the Average Precision (AP), the mAP, and the Precision-Recall (PR) curve, which offer a comprehensive insight into the model's accuracy and reliability.

4.3.1 Precision-Recall Curve

The PR curve is a commonly used metric for assessing the performance of models. In the context of object detection, it offers a comprehensive visualization of a model's performance by analyzing the balance between precision and recall across various confidence thresholds.

Precision measures the ratio of true positive detections to all predicted positive detections. It is expressed as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}.$$

In waste detection, high precision means that the model almost always correctly identifies an object as waste. In other words, the model rarely misclassifies a non-waste item as waste, resulting in few false positives.

Recall indicates the ratio of true positive detections to the total number of actual positive cases. It is given by:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}.$$

High recall means the model correctly identifies waste items, minimizing the number of missed detections, false negatives. Alternatively, if an object is indeed waste, the model is highly likely to detect it as such.

The PR curve is created by changing the confidence threshold, which is required to consider whether the prediction is a positive detection. The model's precision and recall are calculated and plotted for each confidence level on a curve that reflects the trade-off between these two metrics.

The closer the curve is to the top right corner, the better the model's performance, indicating an optimal balance between precision and recall. The threshold can be adjusted to achieve the desired balance by evaluating the trade-off between precision and recall while considering the application's context.

4.3.2 Average Precision and Mean Average Precision

The mAP is another metric used to evaluate object detection models. The AP is based on three key metrics: precision, recall, and Intersection-over-Union (IoU), with IoU being important for determining detection accuracy.

The IoU quantifies the overlap between the predicted and ground truth regions by calculating the ratio of their intersection to their union. It is commonly used to determine whether a predicted bounding box should be classified as a true positive or a false positive. Formally, IoU is defined as:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|},$$

where A represents the predicted region, B represents the ground-truth region, $|A \cap B|$ is the area of intersection between the predicted and ground-truth regions, and $|A \cup B|$ is the area of the union of the predicted and ground-truth regions.

The AP is computed for each class by evaluating the PR curve at discrete intervals, using various IoU thresholds. The AP reflects how well a model balances precision and recall for a given class, considering different IoU thresholds. Mathematically, it is given by:

$$AP = \int_0^1 P(r)\, dr,$$

where $P(r)$ is the precision at the recall r. This integral can be approximated by discretizing the recall values and averaging the precision at those points.
The mAP is then obtained by averaging the APs of all the classes involved in the model. Formally, the mAP can be expressed as:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i,$$

where N is the number of classes, and AP_i is the AP for class i. The mAP is a useful metric because it summarizes the model's performance across all classes, accounting for various confidence thresholds and the overlap of predicted bounding boxes with ground truth.

4.4 Experiments

The experiments were structured into four distinct phases:

- **Full-Class Evaluation:** State-of-the-art models were trained and evaluated on the SortWaste dataset using all annotated classes.
- **Plastic-Only Evaluation:** The models were trained and evaluated on the SortWaste dataset, considering only plastic-related classes.
- **Class-Matching Evaluation:** The models were trained and evaluated on the SortWaste dataset using only the classes that are also present in the ZeroWaste-f [7] dataset, ensuring a fair comparison.
- **Cross-Dataset Evaluation:** The models were trained on one dataset (SortWaste or ZeroWaste-f [7]) and tested on the other, and vice versa, to assess generalization and transferability between datasets.

4.4.1 Full-Class Evaluation

In this first experiment, state-of-the-art object detection models were trained on the SortWaste dataset using all available classes.

	PET	HDPE	Mixed Soft Plastic	ECAL	Metal	Cardboard	Mixed Rigid Plastic	PET Oil	AP	AP50	APs	APm	APl
Faster R-CNN [1]	0.870	0.700	0.460	0.795	0.470	0.093	0.541	0.652	0.415	0.573	nan	0.136	0.431
TridentNet[3]	0.854	0.702	0.444	0.778	0.419	0.123	0.547	0.802	0.407	0.584	nan	0.177	0.419
RetinaNet[4]	0.844	0.723	0.455	0.785	0.517	0.108	0.562	0.755	0.435	0.594	nan	0.178	0.443
YOLOv11 [5]	**0.880**	**0.712**	**0.470**	**0.808**	**0.330**	**0.044**	**0.568**	**0.725**	**0.451**	**0.567**	-	-	-

Table 4.1: mAP results on the SortWaste test set for COCO-pretrained state-of-the-art models fine-tuned on SortWaste. The results are reported per class and overall, using standard COCO evaluation metrics.

In Table 4.1, RetinaNet [4] achieves the highest AP50 score of 0.594, indicating superior performance under an IoU threshold of 50%. However, YOLOv11[5] demonstrates the highest overall AP of 0.451, which reflects performance averaged across multiple IoU thresholds

ranging from 0.5 to 0.95. This suggests that YOLOv11[5] maintains more consistent localization accuracy across a broader range of IoU thresholds, making it more robust for precise object detection in complex, real-world scenarios.
It is also observed that the PET class is the easiest to detect, followed by the ECAL class. In contrast, the Cardboard class proves to be the most challenging, likely due to its lower representation in the dataset. The strong performance in detecting PET and ECAL can be attributed to their distinctive and easily recognizable visual features.
Overall, all evaluated state-of-the-art models achieve an AP50 greater than 55%, indicating good performance. This outcome suggests that, despite challenges such as class imbalance, the models are capable of delivering reliable object detection results within the experimental context.

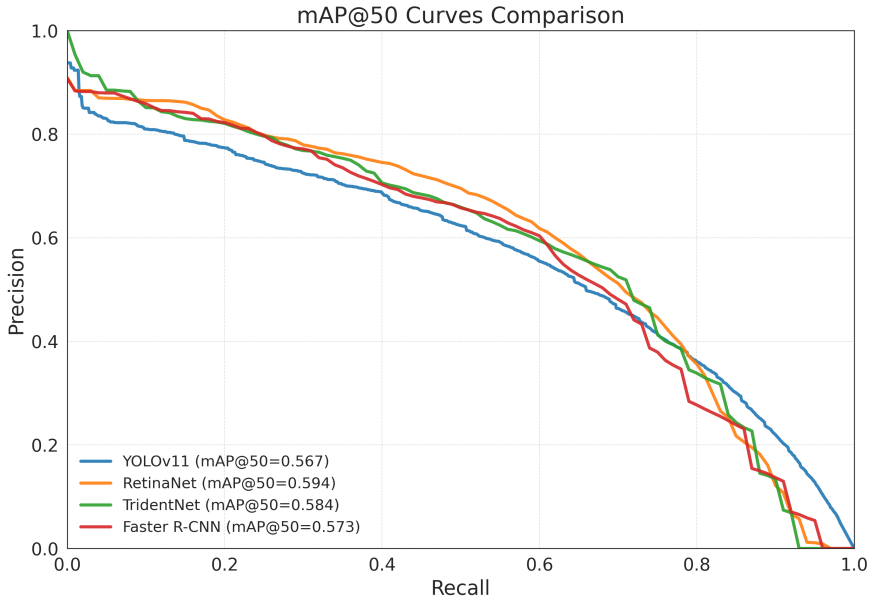


Figure 4.1: Precision-Recall curves for mAP@50 of state-of-the-art object detection models. Each curve represents the performance of a different model.

Figure 4.1 shows the PR curves for mAP@50 across all trained object detection models. RetinaNet [4] achieves the highest mAP@50 of 0.594, maintaining strong precision across a wide range of recalls, which indicates effective localization at an IoU threshold of 0.5.
YOLOv11 [5], while slightly lower at mAP@50, achieves the highest overall AP. Its smoother PR curve reflects balanced performance across varying recall levels, consistent with its strength across stricter IoU thresholds (0.5–0.95).
The PR curves confirm that RetinaNet [4] excels under lenient overlap conditions, while YOLOv11 [5] offers more consistent performance, making it better suited to practical applications.

4.4.2 Plastic-Only Evaluation

The second experimental study represents the main focus of this dissertation. This study was conducted to achieve the primary objective. As previously mentioned, some classes were

merged, and others were excluded because they did not represent plastic materials. As a result, four plastic-related classes were used in this experiment. Table 4.2 presents the results for the best performance achieved by each model.

	PET	HDPE	Mixed Plastic	ECAL	AP	AP50	APs	APm	APl
Faster R-CNN [1]	0.862	0.706	0.623	0.789	0.545	0.745	nan	0.230	55.264
TridentNet [3]	0.836	0.714	0.643	0.774	0.519	0.742	nan	0.189	0.527
RetinaNet [4]	0.858	0.692	0.632	0.775	0.550	0.739	nan	0.217	0.558
YOLOv11 [5]	**0.872**	**0.729**	**0.624**	**0.786**	**0.597**	**0.753**	-	-	-

Table 4.2: mAP results on the SortWaste test set for COCO-pretrained state-of-the-art models fine-tuned on SortWaste. The results are reported per plastic class and overall, using standard COCO evaluation metrics.

As shown in Table 4.2, the YOLOv11 [5] model achieved the best performance among the evaluated models. YOLOv11 [5] is a one-stage object detector known for its speed and efficiency in real-time applications. In this set of experiments, the classes were relatively balanced, as illustrated in Figure 3.4, which helped to reduce bias during training and evaluation.
The AP50 scores across all models were quite similar, indicating consistent performance in detecting plastics. Among the four classes, the PET class was the easiest to detect, as was the case in the previous experiment, likely due to its easily recognizable appearance and consistent visual features. In contrast, the Mixed Plastic class proved to be the most challenging, possibly due to its heterogeneous appearance and less consistent features. Nevertheless, the performance for all classes remained high, demonstrating the models' effectiveness in detecting different types of plastics in MSW.

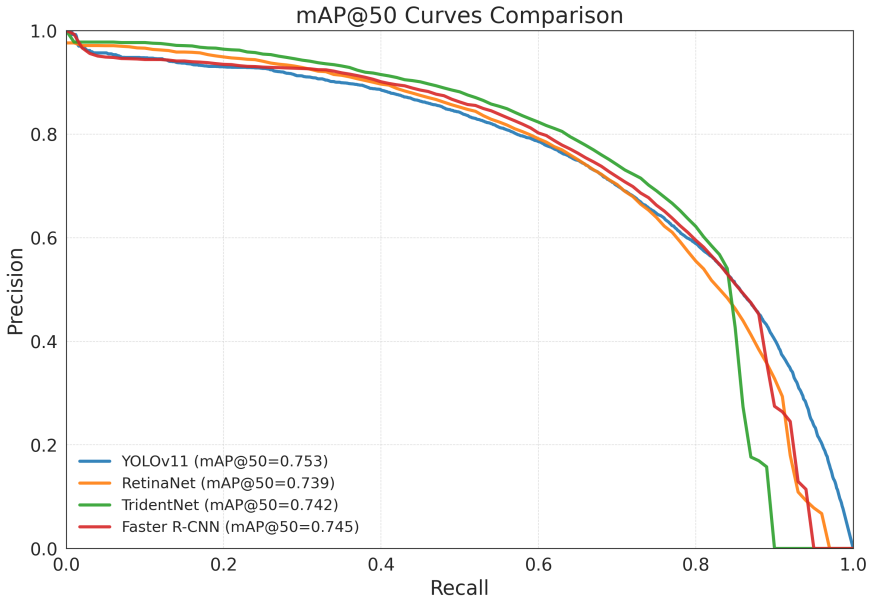


Figure 4.2: Precision-Recall curves for mAP@50 of state-of-the-art object detection models. Each curve represents the performance of a different model trained with plastic classes.

Figure 4.2 presents the PR curves for mAP@50 for the four state-of-the-art object detection models evaluated in this study. Among them, YOLOv11 [5] achieved the highest mAP@50 score of 0.753, as shown in Table 4.2, followed closely by Faster R-CNN [1], TridentNet [3], and RetinaNet [4]. These results demonstrate a relatively narrow performance gap among

the models, indicating that all architectures are capable of effectively detecting plastic waste in MSW under the given experimental conditions.
The PR curves are generally smooth and maintain high precision values across a broad range of recall values, which suggests that the models are both accurate and consistent in their predictions. YOLOv11 [5], in particular, shows a slightly better balance between precision and recall, maintaining higher precision at higher recall levels compared to the other models. Overall, the shape and position of the curves highlight strong detection capabilities across all models, with YOLOv11 [5] standing out as the most robust, especially in scenarios where both high precision and recall are required.

4.4.3 Class-Matching Evaluation

The third experimental study consists of a comparative analysis between models trained on the SortWaste dataset and those trained on the ZeroWaste-f [7] dataset. To enable a meaningful comparison, a class-matching procedure was performed, aligning the class labels from SortWaste with the corresponding classes in ZeroWaste-f [7]. As a result, the classes in the SortWaste dataset were regrouped to match the class structure of the ZeroWaste-f [7] dataset. The distribution of instances per class for each dataset is illustrated in Figures 3.5 and 3.6. Table 4.3 presents the results obtained using the SortWaste-trained models, while Table 4.4 shows the results for the models trained on ZeroWaste-f [7].

	Rigid Plastic	Soft Plastic	Cardboard	Metal	AP	AP50	APs	APm	APl
Faster R-CNN [1]	0.831	0.457	0.752	0.488	0.457	0.632	nan	0.181	0.466
TridentNet [3]	0.801	0.494	0.748	0.433	0.422	0.619	nan	0.128	0.437
RetinaNet [4]	**0.824**	**0.497**	**0.754**	**0.469**	**0.466**	**0.635**	**nan**	**0.144**	**0.484**
YOLOv11 [5]	0.844	0.455	0.748	0.394	0.463	0.610	-	-	-

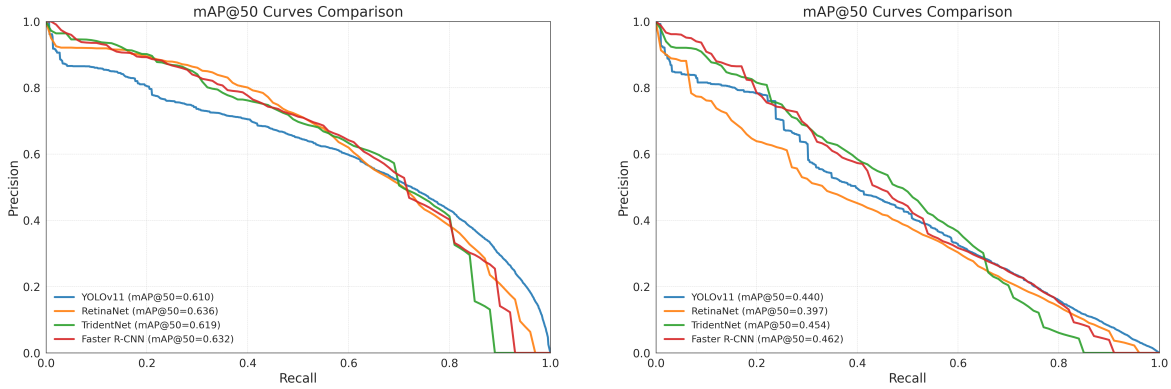
Table 4.3: mAP results on the SortWaste test set for COCO-pretrained state-of-the-art models fine-tuned on SortWaste. The results are reported per class, as ZeroWaste, and overall, using standard COCO evaluation metrics.

	Rigid Plastic	Cardboard	Metal	Soft Plastic	AP	AP50	APs	APm	APl
Faster R-CNN [1]	0.420	0.574	0.361	0.492	0.307	0.462	0.092	0.184	0.332
TridentNet [3]	0.333	0.550	0.428	0.506	0.252	0.454	0.498	0.113	0.277
RetinaNet [4]	0.304	0.558	0.211	0.513	0.265	0.397	0.111	0.176	0.278
YOLOv11 [5]	**0.297**	**0.580**	**0.351**	**0.529**	**0.339**	**0.439**	**-**	**-**	**-**

Table 4.4: mAP results on the ZeroWaste-f [7] test set for COCO-pretrained state-of-the-art models fine-tuned on ZeroWaste-f. The results are reported per class and overall, using standard COCO evaluation metrics.

An analysis of Table 4.3 reveals that among the models trained on the SortWaste dataset, RetinaNet [4] achieved the best performance. As previously discussed, RetinaNet [4] is particularly effective in handling imbalanced datasets, a characteristic present in this experiment. Among the classes, Rigid Plastic and Cardboard were the easiest to detect, likely due to their higher number of instances. In contrast, Soft Plastic and Metal, which had fewer training examples, were more difficult for the models to detect accurately.
Regarding Table 4.4, which presents results for the ZeroWaste-trained models, YOLOv11 [5] demonstrated the highest overall performance. Similar to the SortWaste experiment,

classes with more instances, such as Cardboard and Soft Plastic, achieved higher average precision scores, while classes with fewer examples were more challenging to detect. This trend was consistent across both datasets: classes with greater representation in the training data generally resulted in higher detection accuracy.
Analyzing the two tables, the Cardboard class had approximately 20000 instances in SortWaste and around 17500 in ZeroWaste-f [7], comparable quantities. The corresponding AP50 scores differed significantly: 0.754 for SortWaste and 0.492 for ZeroWaste-f [7]. This suggests that, in addition to class frequency, other dataset-specific factors, such as image quality, annotation consistency, or intra-class variability, influence model performance.



(a) Precision-Recall curves for models trained on the SortWaste dataset.

(b) Precision-Recall curves for models trained on the ZeroWaste-f dataset.

Figure 4.3: Comparison of mAP@50 Precision-Recall curves between models trained on SortWaste and ZeroWaste-f datasets, respectively.

Figure 4.3 compares the mAP@50 PR curves for models trained on the SortWaste and ZeroWaste datasets. In the SortWaste dataset (a), RetinaNet [4] achieved the highest performance, followed closely by Faster R-CNN [1], TridentNet [3], and YOLOv11[5].
In contrast, on the ZeroWaste-f dataset (b), all models showed reduced performance, with Faster R-CNN [1] achieving the highest mAP@50. The lower precision across recall levels highlights the increased difficulty of this dataset, possibly due to its smaller number of instances and greater variability.
In general, the models trained on the SortWaste dataset performed better, confirming that both dataset quality and the number of examples have a significant impact on detection accuracy.

4.4.4 Cross-Dataset Evaluation

To assess the generalization capability of the models across distinct real-world waste datasets, a cross-dataset evaluation was conducted. In this setting, the models were trained on one dataset and evaluated on another, and vice versa. The results of the models trained on the SortWaste dataset and evaluated on the ZeroWaste-f [7] dataset are presented in Table 4.5, while the results of the models trained on ZeroWaste-f [7] and tested on SortWaste are shown in Table 4.6.

	Rigid Plastic	Soft Plastic	Cardboard	Metal	AP	AP50
Faster R-CNN [1]	0.209	0.047	0.099	0.048	0.0659	0.101
TridentNet [3]	0.128	0.031	0.089	0.064	0.044	0.078
RetinaNet [4]	**0.241**	**0.063**	**0.087**	**0.024**	**0.069**	**0.104**
YOLOv11 [5]	0.0806	0.050	0.186	0.0431	0.059	0.090

Table 4.5: mAP results for cross-dataset evaluation. The models were trained on the SortWaste dataset and tested on the ZeroWaste-f [7] dataset. Results are reported per class and overall.

	Rigid Plastic	Cardboard	Metal	Soft Plastic	AP	AP50
Faster R-CNN [1]	0.378	0.220	0.017	0.108	0.109	0.181
TridentNet [3]	0.238	0.160	0.044	0.084	0.066	0.132
RetinaNet [4]	0.401	0.210	0.019	0.119	0.118	0.187
YOLOv11 [5]	**0.355**	**0.303**	**0.050**	**0.110**	**0.14**	**0.205**

Table 4.6: mAP results for cross-dataset evaluation. The models were trained on the ZeroWaste-f [7] dataset and tested on the SortWaste dataset. Results are reported per class and overall.

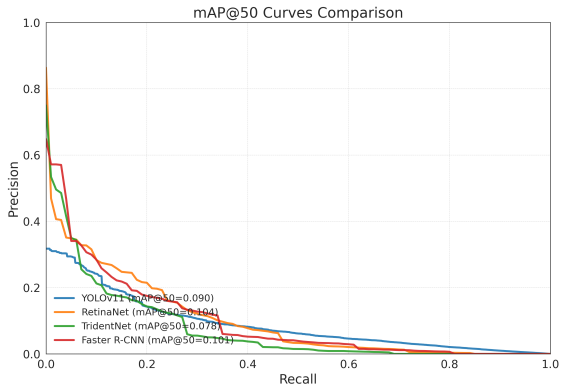
The results highlight a significant drop in performance when the models are applied across domains without any adaptation. For models trained on the SortWaste dataset and tested on ZeroWaste-f [7], RetinaNet [4] achieved the best performance, consistent with its performance when evaluated within the SortWaste dataset. In contrast, when trained on ZeroWaste-f [7] and tested on SortWaste, YOLOv11 [5] demonstrated superior performance.

The highest AP50 in Table 4.5 is 0.104, whereas in Table 4.6 it is 0.204. These values are substantially lower than those obtained in within-dataset evaluations, underscoring a pronounced degradation in performance. This decline highlights the models' limited ability to generalize across domains, indicating a significant shift in the domain between the visual characteristics, object representations, and contextual elements present in the ZeroWaste [7] and SortWaste datasets.

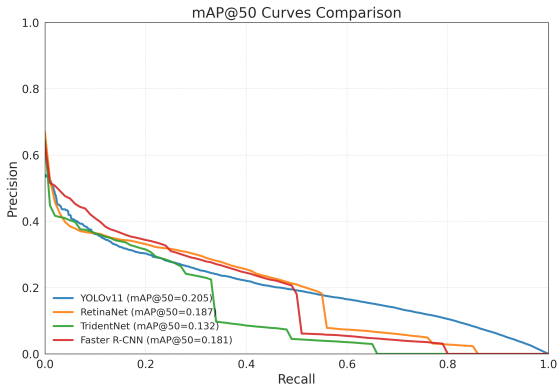
Across all cross-dataset experiments, the AP50 values for generalization from ZeroWaste-f [7] are higher than those from SortWaste. This asymmetry can be attributed to the visual and contextual differences between the datasets. As illustrated in Figure 2.8, images from the ZeroWaste-f [7] dataset often feature complex backgrounds, typically paper surfaces, with only the contaminant objects annotated. Although these images may contain fewer annotated objects, the diverse and noisy background introduces additional variability that the model learns to handle during training.

In contrast, the SortWaste dataset generally includes a higher number of annotated objects per image but presents simpler and more uniform visual scenes. As a result, models trained on SortWaste may struggle to adapt to the more cluttered and visually diverse environments found in ZeroWaste-f [7]. Conversely, models trained on ZeroWaste-f [7] benefit from a broader exposure to background variability, which may enhance their ability to generalize to the simpler visual context of SortWaste.

Figure 4.4 graphically confirms the trends observed in the numerical evaluation: models trained on ZeroWaste-f [7] outperform their counterparts trained on SortWaste when applied across domains. The PR curves exhibit a notable decline in both precision and recall



(a) Precision-Recall curves for models trained on the SortWaste dataset and tested in ZeroWaste-f.



(b) Precision-Recall curves for models trained on the ZeroWaste-f dataset and tested in SortWaste.

Figure 4.4: Comparison of mAP@50 Precision-Recall curves between models trained on SortWaste and ZeroWaste-f datasets, respectively.

under domain shift. These results highlight the need for domain adaptation or dataset augmentation techniques to enhance the robustness and transferability of practical waste classification systems.

4.5 Conclusion

In conclusion of this chapter, a summary of the main insights from each experiment is presented. In the first experiment, it was observed that classes with a higher number of examples were generally easier to detect, or specific classes exhibited distinctive visual characteristics, such as PET Oil, which aligns with expectations. Despite variability across classes, all models achieved a mAP@50 greater than 55%, indicating consistent overall performance.

In the second experiment, the datasets presented a more balanced distribution of class instances. However, performance differences still emerged among classes. For example, the PET class was easier to detect, likely due to its shiny surface and its frequent association with similarly shaped objects, such as bottles. Similarly, the ECAL class demonstrated strong performance due to its distinct visual features, much like the HDPE class, which is often associated with colored plastics. In contrast, the Mixed Plastics class exhibited inconsistent visual characteristics, which may explain its lower detection scores.

When comparing models trained and tested on two different datasets, it became apparent that classes with more training instances generally achieved higher mAP@50 scores. However, the most frequent classes differed between the datasets. Cardboard was the only class with a relatively similar instance count in both, yet its detection performance varied significantly—better in SortWaste than in ZeroWaste. This discrepancy may be attributed to the merging of the ECAL class into Cardboard in the SortWaste dataset, which could enhance the model's ability to learn from distinctive ECAL features.

Models trained on ZeroWaste-f [7] demonstrated better generalization capabilities when applied to the other dataset. This may be attributed to the more complex and varied back-

grounds in ZeroWaste-f [7] images, which could encourage the learning of more robust and transferable features. On the other hand, the greater number of examples in SortWaste likely allowed models trained on it to learn more specific characteristics of each class. However, this may have limited their ability to generalize to datasets not focused on MSW.
In general, all models, regardless of their architecture, achieved similar levels of performance, as evidenced by the PR curves.

Chapter 5

Conclusion and Future Work

5.1 Impacts and Limitations

This work demonstrates the potential of object detection models to enhance waste sorting processes, contributing to both environmental sustainability and operational efficiency. By automatically identifying different types of plastics in MSW, such systems can improve the quality of recyclable materials, reduce worker exposure to hazardous waste, and improve productivity in Material Recovery Facilities (MRFs).
A key contribution is the creation of the SortWaste dataset, which was explicitly developed for this research due to the lack of suitable public datasets in real industrial settings. Although limited in the number of images, the dataset includes a high density of annotated objects and will be made publicly available to support further research in intelligent waste management.
One notable limitation is the class imbalance in the dataset. This is difficult to avoid, as the presence of materials in MSW is unpredictable. Operational factors, such as cardboard degradation due to moisture and the removal of metals by magnets, contribute to this imbalance. While challenging to control, the results still show promising model performance under realistic conditions.

5.2 Future Work

This dissertation establishes a foundation for intelligent waste detection in MBT facilities; however, several avenues for future research remain.
Expanding the dataset to include more samples and material classes could enhance model generalization. Future work may also explore the use of advanced architectures, such as transformer-based models, which have shown promising results in object detection tasks.
Additionally, improving the annotation process through semi-automated techniques could reduce manual effort and enable faster expansion of the dataset. Incorporating video data to leverage temporal information may also contribute to increased detection accuracy, particularly in dynamic conveyor belt environments where occlusions and motion are frequent.

5.3 Conclusions

This dissertation aimed to develop a system capable of detecting valuable materials, specifically plastics, which are among the most polluting components—in MSW streams at MBT facilities. The motivation behind this research derives from the urgent need to enhance waste sorting processes, in real-world industrial settings.

To achieve this objective, a novel dataset was created, addressing a significant gap in the current state-of-the-art: the lack of publicly available datasets collected under realistic industrial conditions, as shown in Table 2.1. The dataset was acquired at an MBT facility in Portugal, where MSW is sorted. It was manually annotated using CVAT [43] and consists of eight waste material classes, with a focus on various types of plastic.

Following the development of the dataset, several state-of-the-art object detection models were trained and evaluated to establish performance benchmarks. Four experiments were conducted: a full-class evaluation, a plastics-only evaluation, a class-matching evaluation, and a cross-dataset evaluation to assess model generalization. The results of the plastics-only evaluation demonstrated that the proposed system is capable of achieving a mAP of up to 59.7%, which is promising and exceeds initial expectations.

This research represents a significant innovation in the field of waste management and automated sorting. Based on the information found, no previous system has been developed under these specific industrial conditions, and no similar dataset existed before this work. To support further scientific advancement, the dataset will be made publicly available, providing resources to the research community focused on intelligent waste sorting systems.

The strong performance of the proposed system highlights the feasibility of implementing intelligent technologies in MBT facilities. Beyond improving sorting efficiency and material recovery rates, such systems can also enhance workers' safety by reducing human exposure to hazardous materials and working conditions.

In summary, this dissertation presents a novel dataset, demonstrates the practical application of deep learning for plastics detection in MSW, and opens up new directions for intelligent automation in waste treatment infrastructures.

Bibliography

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://arxiv.org/abs/1506.01497 xvii, 6, 7, 8, 9, 12, 15, 29, 31, 33, 34, 35, 36

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870 xvii, 7, 8, 9

[3] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," 2019. [Online]. Available: https://arxiv.org/abs/1901.01892 xvii, 8, 9, 29, 31, 33, 34, 35, 36

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018. [Online]. Available: https://arxiv.org/abs/1708.02002 xvii, 9, 10, 29, 31, 32, 33, 34, 35, 36

[5] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," 2024. [Online]. Available: https://arxiv.org/abs/2410.17725 xvii, 10, 11, 29, 31, 32, 33, 34, 35, 36

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020. [Online]. Available: https://arxiv.org/abs/2005.12872 xvii, 11, 12

[7] D. Bashkirova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko, "Zerowaste dataset: Towards deformable object segmentation in cluttered scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 115–21 125. xvii, xix, 16, 17, 18, 19, 24, 25, 26, 27, 29, 31, 34, 35, 36, 37, 38

[8] M. Triassi, R. Alfano, M. Illario, A. Nardone, O. Caporale, and P. Montuori, "Environmental pollution from illegal waste disposal and health effects: A review on the "triangle of death"," *International Journal of Environmental Research and Public Health*, vol. 12, no. 2, pp. 1216–1236, 2015. [Online]. Available: https://www.mdpi.com/1660-4601/12/2/1216 1

[9] S. Kaza, L. C. Yao, P. Bhada-Tata, and F. Van Woerden, *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050.* Washington, DC: World Bank, 2018, urban Development. [Online]. Available: http://hdl.handle.net/10986/30317 1

[10] A. Feil, T. Pretz, M. Jansen, and E. U. T. van Velzen, "Separate collection of plastic waste, better than technical sorting from municipal solid waste?" *Waste Management & Research*, vol. 35, no. 2, pp. 172–180, 2017. 1

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. [Online]. Available: https://arxiv.org/abs/1311.2524 5

[12] R. Girshick, "Fast r-cnn," 2015. [Online]. Available: https://arxiv.org/abs/1504.08083 5, 6

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385 7, 10

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. [Online]. Available: https://arxiv.org/abs/1506.02640 10

[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767 10

[16] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024. [Online]. Available: https://arxiv.org/abs/2405.14458 10

[17] B. D. Carolis, F. Ladogana, and N. Macchiarulo, "Yolo trashnet: Garbage detection in video streams," *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 1–7, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220071608 12

[18] Rania and D. Kumar, "Smart garbage detection system for sustainable waste management using deep learning techniques," in *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*. IEEE, 2024, pp. 253–257. [Online]. Available: https://doi.org/10.1109/ICWITE59797.2024.10503301 12

[19] Y. Wang and X. Zhang, "Autonomous garbage detection for intelligent urban management," in *MATEC Web of Conferences*, vol. 232. EDP Sciences, 2018, p. 01056. 12

[20] P. Zhou, Z. Zhu, X. Xu, X. Liu, B. He, and J. Zhang, "Towards the urban future: A novel trash segregation algorithm based on improved yolov4," *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1526–1531, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:247797188 13

[21] S. Hossain, B. Debnath, A. Anika, M. Junaed-Al-Hossain, S. Biswas, and C. Shahnaz, "Autonomous trash collector based on object detection using deep neural network," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 1406–1410. 13

[22] M. Fulton, J. Hong, M. J. Islam, and J. Sattar, "Robotic detection of marine litter using deep visual detection models," 2018. [Online]. Available: https://arxiv.org/abs/1804.01079 13

[23] J. A. for Marine-Earth Science and T. (JAMSTEC), "J-edi (japan e-library of deep-sea images)," 2022. [Online]. Available: https://www.godac.jamstec.go.jp/jedi/e/ 13, 18

[24] M. Tharani, A. W. Amin, M. Maaz, and M. Taj, "Attention neural network for trash detection on water channels," 2020. [Online]. Available: https://arxiv.org/abs/2007.04639 13

[25] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, S. Sharma, and I. H. Sarker, "Aquavision: Automating the detection of waste in water bodies using deep transfer learning," *Case Studies in Chemical and Environmental Engineering*, vol. 2, p. 100026, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666016420300244 13, 18

[26] P. F. Proença and P. Simões, "Taco: Trash annotations in context for litter detection," 2020. [Online]. Available: https://arxiv.org/abs/2003.06975 13, 16, 18

[27] W.-L. Mao, W.-C. Chen, H. I. K. Fathurrahman, and Y.-H. Lin, "Deep learning networks for real-time regional domestic waste detection," *Journal of Cleaner Production*, vol. 344, p. 131096, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652622007284 14, 18

[28] G. Thung and M. Yang, "Classification of trash for recyclability status," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:27517432 14, 16, 18

[29] L. Zhao, Y. Pan, S. Wang, L. Zhang, and M. Islam, "Skip-yolo: Domestic garbage detection using deep learning method in complex multi-scenes," 07 2021. 14

[30] H. Qin, L. Shu, L. Zhou, S. Deng, H. Xiao, W. Sun, Q. Liang, D. Zhang, and Y. Wang, "Active Learning-DETR: Cost-Effective Object Detection for Kitchen Waste," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024, article ID: 2509115. 14

[31] O. Awe and R. Mengistu, "Final report : Smart trash net : Waste localization and classification," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:36189518 14

[32] W. Ma, X. Wang, and J. Yu, "A lightweight feature fusion single shot multibox detector for garbage detection," *IEEE Access*, vol. 8, pp. 188 577–188 586, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225079169 14

[33] W. Lin, "Yolo-green: A real-time classification and object detection model optimized for waste management," *2021 IEEE International Conference on Big Data (Big Data)*, pp. 51–57, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245934319 14

[34] J. Sousa, A. Rebelo, and J. S. Cardoso, "Automation of waste sorting with deep learning," in *2019 XV Workshop de Visão Computacional (WVC)*, 2019, pp. 43–48. 15, 18

[35] W. Ma, H. Chen, W. Zhang, H. Huang, J. Wu, X. Peng, and Q. Sun, "Dsyolo-trash: An attention mechanism-integrated and object tracking algorithm for solid waste detection," *Waste Management*, vol. 178, pp. 46–56, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0956053X24000990 15

[36] Y. Ren, Y. Li, and X. Gao, "An mrs-yolo model for high-precision waste detection and classification," *Sensors*, vol. 24, no. 13, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/13/4339 15

[37] D. Mewada, C. Agnew, E. M. Grua, C. Eising, P. Denny, M. Heffernan, K. Tierney, P. van de Ven, and A. Scanlan, "Contamination detection from highly cluttered waste scenes using computer vision," *IEEE Access*, vol. 12, pp. 129 434–129 446, 2024. 15

[38] G. Mittal, K. B. Yagnik, M. Garg, and N. C. Krishnan, "Spotgarbage: Smartphone app to detect garbage using deep learning," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 940–945. [Online]. Available: https://doi.org/10.1145/2971648.2971731 16, 18

[39] L. D. I. World, "Wade ai: An ai algorithm for detecting trash in geolocated images," https://github.com/letsdoitworld/wade-ai, 2019, accessed: 2025-01-17. 16, 18

[40] Z. Wang, "Waste pictures dataset," https://www.kaggle.com/datasets/wangziang/waste-pictures, accessed: 2025-01-17. [Online]. Available: https://www.kaggle.com/datasets/wangziang/waste-pictures 16, 18

[41] OpenLitterMap, "Openlittermap web repository," https://github.com/OpenLitterMap/openlittermap-web, accessed: 2025-01-17. [Online]. Available: https://github.com/OpenLitterMap/openlittermap-web 16, 18

[42] J. Bobulski and J. Piątkowski, "Pet waste classification method and plastic waste database - wadaba," in *International Conference on Image Processing and Communications Challenges*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:25529137 16, 18

[43] OpenCV Team, "CVAT: Computer vision annotation tool," https://github.com/opencv/cvat, 2020, accessed: 2025-05-27. 20, 40

[44] Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional, "Despacho n.º 15370/2008," 2008, diário da República, 2.ª série, n.º 106, de 3 de junho de 2008. [Online]. Available: https://diariodarepublica.pt/dr/detalhe/despacho/15370-2008-2516665 20

[45] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019, accessed: YYYY-MM-DD. 29

[46] G. Jocher, A. Chaurasia, J. Fang, Laughing, A. V, C. Garcia, J. Nadar, Imyhxy, K. Michael, J. Borovec, SkalskiP, and Z. Wang, "ultralytics/ultralytics: v8.0.0 - yolov8," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7705639 29