Waste Object Detection with Subtype-Aware Fusion and Efficient Re-Ranking (SAFE-R) for Hazardous Waste Identification and Recycling in Waste Streams.

by

Iftekhar Ahmed
21201032
Rashik Bin Haider
21201546
Nafisa Naznin Choudhury
21201072
Mosammat Sumaiya Haque Subah
21201222

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Rashik Bin Haider
(21201546)

Iftekhar Ahmed
(21201032)

Nafisa Naznin Choudhury
(21201072)

Mosammat Sumaiya Haque Subah
(21201222)

Approval

The thesis titled "Waste Object Detection with Subtype-Aware Fusion and Efficient Re-Ranking (SAFE-R) for Hazardous Waste Identification and Recycling in Waste Streams." submitted by

1. Rashik Bin Haider (21201546)
2. Nafisa Naznin Choudhury (21201072)
3. Mosammat Sumaiya Haque Subah (21201222)
4. Iftekhar Ahmed (21201032)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer 2025.

Examining Committee:

Supervisor:
(Member)

Md. Tanzim Reza
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Waste management is essential for a multitude of reasons, including protecting the environment and the people who are employed for this job, especially in developing countries like Bangladesh where primitive practices of waste-collection and landfilling still persist. Improper classification and disposal of waste can lead to many issues like air and water pollution. This study aims to address the challenges of waste management by proposing an automated system that classifies waste into seven different categories like sharpened objects, plastic, metal, glass, etc., using computer vision. To that end, multiple modifications of the popular YOLOv8 model were tested to ascertain the ideal balance between accuracy and latency for real-world implementation. Therefore, after filtering the set of 2,444 images (24,715 annotated objects) and trying various architectural variations, we discovered that the YOLOv8m detector has the balance of being the fastest and most accurate (mAP50 = 0.591, mAP50-95 = 0.332, around 9 ms per image).Furthermore, our study aims to identify hazardous objects like syringes amongst medical waste and nails amongst sharpened objects using query-aware reranking. When identifying dangerous subclasses our pipeline, too, showed significant improvement compared to an open-vocabulary baseline (OWL-ViT). To illustrate, with an IoU of 0.5, syringe detection had an AP50 of 0.619 versus 0.064 with OWL-ViT and nails had 0.538 versus 0.214. These findings indicate that we can successfully detect the presence of dangerous objects such as syringes in real time, which makes waste collection not only quicker and more efficient but also significantly safer to the people working on it. Implementation of proper waste classification using our proposed pipeline can give the waste-collectors safety and promote practices like composting and recycling. In this thesis, we propose to introduce an automated system that will help classify and separate garbage into superclasses and further distinguish subclasses by the combination of YOLOv8m and CLIP algorithms which can aid garbage collectors in their task and foster better recycling and landfill practices.

Keywords: Object Detection, YOLO, CLIP, Query Based Detection, Hazardous Identification, Recycling, Waste management, Contrastive learning, Waste Dataset, OWL-ViT, CNN, Deep Learning.

Acknowledgement

The gratitude due for the completion of this thesis is manifold, starting with the Almighty for gracing us with resilience, patience and knowledge to successfully finish our research. We would also like to extend our heartfelt appreciation to our supervisor Md Tanzim Reza and our co-supervisor Dr. Md. Golam Rabiul Alam for their steadfast support and feedback during this endeavor. Next, we would like to sincerely thank our families without whose love and support we wouldn't have been able to accomplish what we have. Finally, we want to thank all the people who helped us, by giving us company on this journey that we all gave our utmost to.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The adoption of innovative technologies has revolutionized a wide range of industries and sectors, bringing about progress in leaps and bounds. However, in the field of waste management, primitive practices of waste-collection and landfilling still persist, especially in developing countries like Bangladesh. Concerningly, Bangladesh's rapid population growth, urbanization and industrialization all contribute to the severe increase in the country's annual waste generation; the waste generation rate increased from 1.1 million tonnes in 1970 to 5.2 million tonnes in 2015 [5]. More particularly, in 2005, it was recorded that approximately 7700 tons of municipal solid waste was generated daily in the six major cities, with the capital city Dhaka generating the highest [1]. The average urban waste composition recorded in 2014 Bangladesh Waste Database 2014 [3] documents 77.70% organic waste (vegetable and food waste), 7.35% plastic waste, 4.84% paper waste, 2.72% wood waste, 2.56% fabrics waste, 0.64% e-waste (discarded electronic components), 0.44% metals, and 3.74% other wastes.

The staggering amount of solid waste generated daily is dealt with by an inadequate municipal waste management system where solid waste is dumped indiscriminately into water bodies and landfills, or else burnt [11]. Increasing population and the scarcity of land were identified as major issues in waste management by Shams et al.. Regimens that are currently in place harm the environment and surrounding populace in a multitude of ways, as evidenced by Parvin and Tareq's systemic review [18] of the public health risks posed by toxic byproduct of landfill, leachate, which can pollute water bodies; additionally, heavy metals may also be present in leachate which can affect edible plants near landfill sites that cause cancer upon consumption. Other health risks include low birth weight, congenital anomalies, and respiratory diseases.

Besides the health risks faced by the population residing near landfills, extant practices also pose many risks for waste-collectors and landfill-workers. These include risk of injury caused by sharp objects. The health and safety hazards faced by these employees is dire due to the lack of proper equipment as well as the indiscriminate and unclassified nature of the waste they deal with at every step of the system, from collection to disposal.

To tackle these pressing issues perpetuated by the conventional management of municipal waste, Urme et al. (2021) [19] conclude that it's imperative to improve primary waste disposal techniques to reduce solid waste accumulation in landfills and suggest incorporating the creation of refuse-derived fuel (RDF) to effectively handle municipal waste and promote a green environment. To produce RDF from solid waste, necessary components must be separated from the rest, namely combustible materials such as plastics, paper materials and wood; promoting the use of RDF would reduce the reliance on fossil fuels for energy as well as decrease the emission of greenhouse gasses in the environment [28].

However, to establish an effective and proactive waste management system, waste segregation must be performed at the source, ie. at the household level as well as at community level [2] (Matter et al., 2013). Waste collection from households is generally carried out by poor, unskilled, uneducated waste pickers employed by private contractors who go door-to-door to collect waste from residential buildings, shops, restaurants and other establishments. As such, the waste collected is not separated into categories to make segregation easier, as is common in many Asian cities [21].

To summarise, enormous amounts of solid waste is generated in Bangladesh but it is not managed effectively, resulting in adverse health effects for waste-collectors, landfill-workers and the general population as well as causing extensive environmental harm. Hence, resolving how to manage the municipal waste is a pressing concern in order to ensure safeguard public health and reduce environmental degradation.

1.2 Rational of the Study or Motivation

Solid waste management poses such great challenges globally, motivating research into this field in an effort to mitigate the harmful effects to the environment and public health, yielding many solutions in the way of policy changes [5], improving landfill infrastructure and engineering, implementing holistic waste management models [12], and much more. In particular, Akther et al. (2024) [30] emphasise the importance of source separation in reducing waste volume and enforcing sustainability by promoting the 3R strategy (reduce, reuse, recycle). Alongside traditional approaches, it is prudent to leverage technology amidst existing systems to find effective solutions to these challenges.

The primary contributor to the dismal state of the waste management system is the lack of separation or segregation of the solid waste collected. Cutting-edge technology can be utilised to bridge this gap and enforce sorting and categorisation of waste at its source. Machine learning and deep learning models can process vast amounts of data, especially images, to facilitate the automation of the classification system to accurately sort solid wastes into suitable categories, ie. recyclable and non-recyclable, in order to reduce waste in landfills. Further classification can be done to separate waste based on material (eg. paper, plastic, glass, etc) as well as identifying medical waste or hazardous sharp edges to prevent harm to waste-collectors and landfill-workers.

1.3 Problem Statement

Management of municipal solid waste is a significant challenge that can be tackled by employing Computer Vision, a branch of Artificial Intelligence which allows machines to process visual data from the real world. The intersection of computer vision and deep learning models can be leveraged for object recognition, classification, and segmentation, thus facilitating easier and automatic classification of waste at the point of collection. Solid waste should be sorted into the following categories– organic waste (food and other organic matter), paper, cardboard, glass, plastic, metal, sharp/broken waste, medical waste, and other. As such, each category of waste can be disposed of appropriately: organic waste can be used as compost; recyclables can be processed to reuse; sharp/broken waste can be segregated to protect waste-collectors and handlers; and the remaining can be sent for incineration to convert to energy. Therefore, object detection and classification must be carried out with emphasis on accuracy as well as on efficiency to process images in real-time for a deployable solution to this problem. Furthermore, distinct detection of hazardous objects amongst these classes is also required to keep waste-collectors safe, for instance, identifying syringes within medical waste, nails among metal waste, and so on. Thus, it is imperative that a suitable model is employed for the automated classification system for waste classification and segregation of hazardous objects, one that is trained on a robust dataset for greater reliability.

1.4 Objective

The aim of this study is to optimize the municipal waste management system by developing a robust and effective deep learning model or pipeline by employing computer vision to automate the process of solid waste classification into suitable categories as well as implementing a query-based system for identification of hazardous objects within the superclasses. This includes training the model on a vast and varied dataset to improve accuracy, efficiency and prevent overfitting. While numerous datasets are already available, previous research has relied heavily on these; for a deployable system in Bangladesh, we need a robust dataset that is representative of the solid waste produced in the country, with a balanced exemplification of relevant classes. To that end, we aim to create a repository of annotated images of solid waste that can be utilised here.
Below are the principal objectives of our thesis:

1. To develop a robust and comprehensive waste dataset for research into waste detection and classification, with relevant categories for recyclable and hazardous waste.

2. To evaluate the performance of modified YOLOv8 models to determine the ideal configuration for real-life applications.

3. To design a query-based object detection pipeline for distinguishing specific hazardous objects within a superclass.

4. To address the lack of municipal waste management in Bangladesh by introducing an automated backbone for a technology-driven solution.

1.5 Methodology in Brief

To achieve the goal of automated waste classification, the deep learning model must be trained on a robust dataset. The overall workflow of this study includes creation of the dataset and training of the model. Furthermore, we introduced a garbage-detection pipeline, SAFE-R, that operates in the cluttered, real-world environment and still executes in real time by combining YOLO detector with CLIP, an OpenAI neural network trained on a variety of image, text pairs. This pipeline leverages superclass filtering and CLIP's semantics for accurate and fast identification of hazardous and harmless waste objects.

1.5.1 Creating the Dataset

Our initial step was to assemble our dataset, containing approximately 2.4k pictures of seven types of waste (glass, medical, metal, organic, paper, plastic, sharp-object). We filtered the data, retained only helpful images and divided them into 70/15/15 train/val/test. We letterboxed all the images to 640×640, and we did gentle, object-friendly augmentations (mosaic, mixup, mild geometry and colour changes) so that thin objects such as syringes and nails would not be distorted.

1.5.2 Training the Model

In the case of the model, we selected YOLOv8m (pretrained on COCO) as it provides a good balance of accuracy and time ratio. We held the training recipe constant (50 epochs, batch 8, SGD, warmup, AMP) and then experimented with specific adjustments to understand what, in fact, works: we added a P2 head for tiny objects, replaced the neck with a BiFPN with lightweight ECA attention, or added Residual-CBAM to focus more attention on significant features. To make comparisons fair, we tested all of them with identical settings of inferences. Subclass identification (e.g., distinguishing between a syringe and a test tube) is difficult to solve using a detector only, and hence we included a second step. Our detector suggests boxes and we cut them by superclass and then re-score the cropped regions with text prompts by CLIP. We combine the score of CLIP with the confidence of the detector (soft shape prior in the form of long, thin tools) and retain only boxes that match the query. In a nutshell: YOLOv8m is fast and reliable at localization; the CLIP reranker provides the fine-grained labels. Collectively they still satisfy our requirements: good enough to be useful, fast enough to get implemented.

1.5.3 Evaluation and Comparison

Performance of the YOLO variants were carried out using the metrics precision, recall, mAP50, mAP50-95. It was observed that the baseline YOLOv8m outperformed its variants, although there were some class-to-class differences in mAP and recall, indicating class-sensitivity and background-object confusion that couldn't be bridged by architectural modifications.
For our SAFE-R pipeline, we observed performance metrics that outstripped our set baseline of OWL-ViT (an open-vocabulary object detection network similarly trained on a variety of image, text pairs) across a set of eight subclasses. To quantitatively assess the significance of the improvements in AP@0.5 of our model, we used

the Wilcoxon Signed Rank test which revealed statistically significant improvement of SAFE-R over OWL-ViT.

1.6 Scopes and Challenges

The development of an automated waste classification pipeline involved multiple challenges. Most importantly, it is integral to ensure the new dataset is diverse and comprehensive in order to uphold data quality and availability for the model to perform accurately. Moreover, despite the size of the dataset, there is much variability in the types of waste with distinct size, shape, texture, colour, condition, etc which poses a significant challenge in classifying them accurately. It is also crucial to take into consideration the real-world conditions that the system must operate in where environmental factors like lighting, visibility, etc. can hamper the reliability of computer vision. We went over the entire dataset with a fine-toothed comb to correctly classify the waste with bounding boxes that met quality standards. Next, it was imperative to ensure that the evaluation of our architectural variants were carried out without bias so we could ascertain the best performing detector reliably. To that end, we opted for standardised evaluation metrics of mAP50, mAP50-95, Precision, Recall. Despite our hopes for the architectural modifications to yield more promising results, the evidence showed that YOLOv8m was the ideal detector for a pipeline to be suitable for deployment in conditions requiring a good balance between accuracy and speed. Lastly, our query-aware pipeline for hazardous object identification had to be evaluated with a baseline to accurately assess its performance. Hence, we chose to compare it with OWL-ViT, an open-vocabulary detection architecture under identical conditions to draw conclusions as to the pipeline's performance [37].

Chapter 2

Literature Review

2.1 Preliminaries

In the rapid advancement of society through both production and consumption, the management of the enormous amounts of waste generated is an oft-neglected issue that must be addressed in order to embrace the progress of our community in actuality. In fact, mismanagement and inefficient disposal of solid waste engenders several challenges like environmental pollution, resource depletion, adverse health effects, and much more. This thesis is an effort to resolve this pressing issue. In this section, key concepts and theories used to construct our proposal of an automated waste classification system are explored.

2.1.1 Waste Management and Classification

The municipal waste management system encompasses many steps starting from waste collection, segregation, recycling, transportation and disposal (Syeda et al., 2017) [6]. This cycle must be carried out in an environmentally friendly and sustainable manner. Focusing on the classification of waste into reusable and harmful categories is crucial for effective recycling and reducing landfill impact and automating the process would improve efficiency and accuracy. Solid waste would be separated into the following categories: organic waste (food and other organic matter), paper, cardboard, glass, plastic, metal, e-waste, sharp/broken waste, medical waste, and other waste.

2.1.2 Role of Deep Learning and Computer Vision in Waste Detection

Deep learning has prompted tremendous progress in several areas of computer vision, for instance, object detection, motion tracking, action recognition, human pose estimation, and semantic segmentation (Voulodimos et al., 2018) [8]. Employing the techniques of object recognition and classification in particular, deep learning can be integrated into waste management systems to identify waste types and categorise them accordingly for an improved management system. In fact, many popular deep learning models, such as CNN, YOLO, Inception, ResNet, etc, have been utilized to significantly refine the performance of computer vision systems in waste classification tasks.

2.1.3 Common Datasets

Over the years, multiple datasets have been created for the purpose of aiding waste classification. Some popular datasets include the following:

- **TrashNet**: Introduced by Yang and Thung (2016) [4], the dataset is composed of solid waste classified into six categories (cardboard, glass, metal, plastic, paper, and trash). Since it was published, this dataset has become the foundation for many papers exploring the categorization of waste using computer vision.

- **TACO (Trash Annotations in Context)**: Created by Proença and Simões (2020) [14], it is an open image dataset for litter detection and segmentation that is growing through crowdsourcing. The images are annotated with bounding boxes and segmentation masks.

- **Detect-Waste**: A merged collection of Extended TACO and other publicly available datasets (Wade-AI, UAVVaste, TrashCan, Drinking-Waste, etc) with a single label 'litter' to overcome the limitations of uneven distribution of classes, lack of background diversity, small number of annotations, etc. (Majchrowska et al., 2022) [23].

- **Classify-Waste**: A merged collection of publicly available datasets proposed as a benchmark by Majchrowska et al. [23] (2022) with eight classification labels (bio, glass, metals and plastics, non-recyclable, paper, unknown waste, other, and an extra class background for images without any litter).

2.2 Review of Existing Research

2.2.1 Object Detection Approaches

To address waste management, Kumar [33] presented an insightful garbage detection method based on the use of YOLOv8. Paper, plastic, cardboard, glass, metal, and rubbish were the six categories into which the 2,519 annotated photos from the TrashNet repository were categorized and used by the system. With mean average precision (mAP) values of 98.1% for cardboard and over 95% for other waste categories, the system showed significant classification accuracy. The integration of the system with advanced technologies like cameras and drones for continual surveillance was highlighted in the study. Future developments will focus on increasing the system's scalability in a variety of circumstances and enhancing the dataset by including more waste categories.

Aishwarya et al. [15] presented a waste management method that identified and separated non-biodegradable garbage using the YOLO algorithm. For metal, the model's accuracy was 75%; for plastic, it was 65%; and for glass, it was 60%. Regardless these encouraging findings, the study admits that improved accuracy can be attained by the use of advanced models or optimal approaches. The work highlights the potential of machine learning in environmental sustainability and is a significant step towards automating waste classification process, even if the authors did not name a specific dataset.

A deep learning-based system was presented by Islam and Alam [29] to enable real-time identification and classification of waste into five categories: paper, plastic, metal, glass, and kitchen waste. Over 135,000 annotated objects taken from the Garbage Annotated in Context (GACO) dataset were utilized to train the system. A YOLOv3-based structure with an SPP-PAN neck and a CSPDarknet53 backbone was used by the authors. At an Intersection over Union (IoU) threshold of 0.35, the model's mean average precision (mAP) was 66.08%; but the system had trouble with accuracy in the glass and paper categories. To make the detection model more broadly applicable in waste management settings, the authors propose expansion of the dataset and improving it.

A deep learning-based garbage detection system designed for both urban and natural environments was proposed by Majchrowska et al. [23]. The study made use of a combined dataset known as Detect-Waste, which merged many publicly accessible datasets, including TrashNet and TACO. Using a single waste class, they assessed three object detection models: EfficientDet, DETR, and Mask R-CNN. They discovered that EfficientDet-D2 had the best detection accuracy, at 65.5%. EfficientNet-B2 performed better than other models for classification. The system's overall accuracy was 73%, and on the training set, it was higher at 86.7%. However, the scientists point out that sorting industrial garbage has additional difficulties that require further research.

Using YOLOv8, Kumar [33] display good performance with a high classification accuracy across an array of waste categories, reaching up to 98.1% accuracy for particular materials such as metal and cardboard. Yet, the model's scalability and generalization to a variety of real-world garbage are limited by its dependence on a very small dataset of 2,500 photos. Similar to YOLO, but with a more focused approach, Aishwarya et al. [15] concentrate on non-biodegradable garbage, such as metals and plastics. Although the model's simplicity enables quicker training, its performance is not well—it can only classify metals with 75% accuracy—which highlights the need for improved model optimization and a wider variety of datasets. Islam and Alam [29], on the other hand, use a more sophisticated architecture with YOLOv3 and a larger dataset (GACO) to accomplish a mean average precision of 66.08%. Although their approach is more resilient and shows dependability in practical settings, it has trouble classifying glass and paper, indicating that multi-class classification could use more work. Majchrowska et al. [23] are notable for their thorough analysis of several detection models, utilizing EfficientDet-D2 to achieve a 73% classification accuracy. Even so, the model's usefulness for waste management activities involving a greater range of waste types is limited by their detection's concentration on a single class (litter). Although all four studies provide insightful information, they all have similar drawbacks, including reliance on short datasets, less accuracy for specific waste types, and difficulties with practical implementation.

2.2.2 Classification Approaches

Deep Neural Networks for Trash Classification (DNN-TC), a distinct deep learning framework, was proposed by Vo et al. [10] (2023) to improve waste classification accuracy using transfer learning. The framework has been verified on two datasets: VN-Trash and TrashNet. By utilizing transfer learning, the model could adjust to different waste classifications and ultimately achieve 94% and 98% accuracy on the datasets, respectively which showed that it excelled in the earlier techniques. Additionally, the author proposed that adding segmentation techniques for input preprocessing would help improve the extracted feature representations by reducing noise in the datasets. And this way, DNN-TC's potential as a reliable automated waste classification system is demonstrated.

Mao et al. [17] 2021 focused on improving the efficiency and accuracy of DenseNet121 and with that view, presented a new Convolution Neural Network architecture for the classification of waste. Furthermore, the research augmented the dataset, so that it is large and varied. The authors also used modified DenseNet21 where Genetic Algorithm was used to optimize its classification layer, by adjusting the hyperparameters to improve efficiency and accuracy. Using Keras, and Python, some tests were done and these tests were profitable as the time required dropped when compared to the unoptimized algorithms. In their tests, the accuracy of the model increased to 99.40% .The authors state that future research might find the reasons to why misclassifications happen in the case of classification, look into multi-object detection in the garbage photos, and improve the algorithms and architecture using the TrashNet dataset.

Combining deep learning and the Internet of Things (IoT), Wang et al. [20] in a different study, proposed a smart municipal waste management system to improve waste classification accuracy. In this study, a modified version of the TrashNet dataset was used which included nine categories like plastic, glass, metal, and hazardous garbage. Here, the classification algorithm MobileNetV3 was used to allow the facilitating waste classification using smartphones and sorting it before its disposal. The model attained a high classification accuracy of 94.24%. because of its minimal complexity (49.5 MB) and quick processing speed (261.7 ms per image). The authors suggested investigating instance segmentation in future research to detect various trash categories in a single image.

In order to efficiently categorize recyclable items, Ahmed et al. [26] proposed an automated waste classification system which utilizes CNN and pre-trained models like DenseNet169, MobileNetV2, and ResNet50V2. A Middle Eastern dataset of 5,000 garbage photos (obtained via Kaggle) was preprocessed using their methods by first scaling the photographs to 224 $\times$ 224 pixels, then normalizing them, and then dividing the data into 70% training and 30% testing sets. To optimize the models, they used a Randomized Search CV for hyperparameter adjustment. While their customary CNN obtained an accuracy of 88.5%, the ResNet50V2 model exceeded the others, with the best accuracy (98.95%), precision (98.35%), recall (98.38%), and F1-score (98.38%). The study did point out that when it comes to identifying unknown or irrelevant trash photos, performance may suffer. They proposed using augmenting data, transfer learning, and fusion learning methodologies to make future advancements.

Malik et al. [24] used a CNN model called EfficientNet-B0 that was pre-trained on ImageNet to offer a trash classification framework for sustainable development. Based on 34 distinct classes, it divides solid waste into 10 higher-order groups.By assigning these groups to different kinds of recycling and composting plants, the framework guarantees effective classification for waste management. In order to maintain pre-trained knowledge, the study used transfer learning, freezing the internal layers of EfficientNet-B0 and fine-tuning additional layers using a secondary dataset specific to regional waste compositions. Some examples are plastic bottles in metropolitan areas or coconut shells in coastal locations. The 11,823 photos in the collection, which came from Kaggle, represented 34 different types of solid waste. With far lower computing requirements, the model's classification accuracy of 81.2% was on par with larger models like EfficientNet-B3. Compared to EfficientNet-B3, which uses 1.8 billion floating-point operations per second, EfficientNet-B0 uses just 0.39 billion. This makes it more economical. Notwithstanding these successes, the study found that managing subclasses within categories had limitations. To further improve classification refinement and accuracy, future research may incorporate subclass identification and segmentation approaches.

A lightweight 5-layer CNN model was designed by Nnamoko et al. [25] in order to maximize computational efficiency and reproducibility when classifying solid waste into two categories: recyclable and organic. Using Sekar's Waste Classification Dataset from Kaggle, which included 25,077 photos (made 345,870 by flipping, rotating, and adjusting brightness), it was used. High resolution (225×264 pixels) and low resolution (80×45 pixels) were the two resolutions used to assess the CNN architecture. A 60% training, 15% validation, and 25% testing data split was used to train the model, which was implemented using Keras with cross-entropy loss and Adadelta optimization. With a smaller size (1.08 MB) and a greater accuracy of 80.88%, the low-resolution model only needed 6.4 hours of training. Alternatively, the high-resolution model achieved 76.19% accuracy, but it was larger (2.35 MB) and took 65.46 hours of training. At 50.05 percent, both models performed significantly better than a baseline random guess classifier. The study showed variations in accuracy during training, validation and testing due to the irregularity of the data set and misclassifications. Here, in order to improve accuracy, the authors suggested to enhance the dropout rates, fine-tuning dataset labels, and investigating alternative resolutions.

Hossen et al. [32] utilizing the TrashNet dataset to classify recyclable waste presented a novel deep learning network named RWC-Net (Recyclable garbage Classification Network). Here the wastes are classified into six categories naming glass, metal, cardboard, plastic, and litter. The study's objective was to develop a solid framework that increases the effectiveness of waste sorting and recycling processes. With an impressive 95.01% overall accuracy and high F1-scores for specific categories, like 97.24% for cardboard and 96.18% for glass, RWC-Net outperformed several advanced models, demonstrating its reliability as well as accuracy.Where as models like AlexNet and optimized VGG16 previously obtained 91% and 93% accuracy respectively. Using Score-CAM saliency maps offering insights into categorization performance, the model's emphasis across various waste types was illustrated.The model used LogSoftMax activation and Cross-Entropy loss functions to provide strong optimization. The study further highlighted the growing urgency for effective waste management by citing statistics indicating that one-third of the world's garbage is not handled properly and that the global production of municipal solid waste could rise from 2.01 billion tons in 2016 to 2.59 billion tons by 2030. Future research might expand the model's potential use for a greater variety of garbage classification tasks through the addition of new datasets, like TACO or VN-trash.

In another instance of optimizing existing classification architectures, Toğaçar et al. [9] put forth a new framework for AlexNet, GoogLeNet and ResNet-50 architectures with AutoEncoder integration in order to reconstruct dataset images to reduce noise and Ridge Regression (RR) application to choose optimal features. The results were promising with increased accuracy in each case, in particular the highest classification accuracy of 99.95% achieved. This showcases the significant potential of achieving better results when utilizing feature selection and refining dataset quality, especially in the case of waste classification.

Additionally, in what can now be considered a foundational project in the undertaking of waste classification, Yang and Thung [4] developed the first open source dataset of images in their attempt at applying Support Vector Machines (SVM) classifier with scale-invariant feature transform (SIFT) features. The dataset contained approximately 2,400 manually-collected images classified into six categories (paper, metal, plastic, glass, cardboard and trash). Their proposed model achieved 63% accuracy with the highest accuracy for paper and metal. To compare the results, they used the Torch7 framework for Lua to construct an 11-layer CNN similar to AlexNet; however, due to hyperparameter tuning issues and disproportionate dataset, the network failed to learn properly, achieving a dismal 22% accuracy. The authors posited that increasing the dataset size and better tuned hyperparameters could yield better results for the CNN, even predicting that it could potentially perform better than the SVM.

Another innovation in classifying waste into recyclable and non-recyclable was proposed by Che et al. [7] where Multilayer Hybrid System (MHS) was combined with AlexNet and Multilayer Perceptron (MLP). Trained on a custom dataset of 5000 images, the MHS model significantly outperformed the CNN-only model, achieving greater rates of accuracy, precision and recall. This was attributed to the integration of sensor-based material detection (namely bridge-sensor and inductor) alongside image-based classification in MHS, especially as it aided in the classification of materials with weak image features, for instance, transparent plastics and small metallic items. It would be worthwhile to further investigate this consolidation of information from both image and sensory channels, in particular by expanding the dataset, enhancing multi-object detection and refining sensor-based classification techniques, in order to improve efficiency and effectiveness or waste categorisation.

2.2.3 Semantic Segmentation Approaches

For classifying the construction waste (CW) into different types, naming Rock/Stone, Bricks/Concrete, and Cardboard/Wood, Lu et al. [22]created a semantic segmentation model with ResNet and Xception as backbone models. For doing this DeepLav3+ is used as the primary architecture. Techniques like MS (multi-scale) input. LR (left-right flip) is used along with optimized hyperparameters such as output stride and resolution. A dataset of 5,366 images of CW is used and annotated for higher accuracy. The demonstration shows a respectable segmentation accuracy where F1 scores of 0.696 and 0.688, performance evaluation on AWS infrastructure, including NVIDIA K80 and Tesla V100 GPUs, revealed mIoU scores of 0.562 and

0.555 on validation and test sets, respectively.

Though positives is indicated by its success but no exception that it has its own demerits. The model had trouble while it was illustrating the edges and also boundaries of the material as certain the materials were very difficult to distinguish due to its complex composition.For that the author suggested SegFix and superpixel segmentation for increasing the segmentation accuracy and automatic edge extraction respectively.Futhermore, it is noticed that the model's efficiency declined while handling up to 30 fps, and segmentation performancec differences within categories is noticed. So, for improving the model and make it more competitive for real-world applications, the focus will be on refining edge detection and preprocessing to increase segmentation accuracy in the future.

Panwar et al. [13] presented AquaVision, in order to assist waste management and the preservation of aquatic life. It is basically a system that detect waste in aquatic environments using deep transfer learning.The study used a database which is an extension of TACO containing curated rubbish photos exclusive to water bodies named AquaTrash alongside TACO.AquaVision used a AquaVision used a transfer learning strategy, utilizing pre-trained models on these dataset to tackle the difficulties of identifying waste in aquatic situations. The contextually aware waste detection capabilities of the model address an urgent requirement for automated solutions in sustainability-focused applications.

The study highlights the importance of accurate waste detection in aquatic environments, where improper waste management puts marine life in danger.Although the AquaTrash and TACO datasets offer a variety of training environments, their reliance on carefully selected sources can restrict their ability to adapt to real-world situations with greater complexity waste compositions. In the future, the dataset might be expanded to include a greater range of aquatic circumstances, and instance segmentation might be used to increase the granularity of detection. AquaVision shows a potential step toward intelligent waste detection systems catered to environmental sustainability concerns by fusing state-of-the-art transfer learning techniques with specific datasets.

2.3 Summary of Key Findings

2.3.1 Key Results and Trends

Striving for more accurate and efficient waste classification, existing models have been optimized and numerous deep learning architectures have been developed. Among these models, RWC-Net performed outstandingly with an accuracy of 95.01% on the TrashNet dataset, surpassing other models like VGG16 and AlexNet [32]. Besides accuracy, there was also a lot of focus on reducing the computation time and cost for smoother implementation in real-world scenarios, namely MobileNetV3 used by Wang et al. [20] that achieved 94.24% accuracy while maintaining low complexity and retaining high inference speed allowing its use in mobile smartphones. Similarly, the lighter architecture of a CNN model used in Nnamoko el al. had low computational cost yet demonstrated 80.88% accuracy [25]. The key was to trade high-resolution images for low-resolution ones to cut down on training time significantly. Such models are more suited for adoption in municipal waste management systems where resources may have constraints.

Several datasets are publicly available and these are commonly used in research, for instance, TrashNet, TACO, and Sekar's Waste Classification dataset are popular choices that were widely used. To accommodate more specific purposes like classification of garbage in water bodies to help prevent the degradation of marine life, Panwar et al. [13] introduced AquaVision. Each dataset contains a different number of classes and varying categories. There are many limitations that should be resolved, starting from limited images, inconsistencies, lack of reliable annotations, and more. In many cases, data augmentation and transfer learning were used to help model generalisation.

Research into semantic segmentation shifted waste classification to more detailed waste analysis allowing for sorting of mixed waste, specifically the use of DeepLabv3+ by Lu et al.. This could further assist the integration into real-world systems by allowing categorisation in a cluttered environment with complex backgrounds.

2.3.2 Implications

Exploration of existing literature reveals the importance of developing light-weight models that can yield accurate results efficiently through low computational costs that are most suitable for real-world implementation. Furthermore, it's critical to develop and utilize a high-quality dataset with sufficient and relevant classes that is robust and diverse. For applications in specific regions, for instance in developing countries like Bangladesh, utilising datasets specific to the region can improve real-world performance. Future research is required to find the best method to integrate this new technology into practice, including exploring the intersection of robotics and waste collection, as well as delving into revising extant policies of waste management. Overall, promising research suggests that the challenges faced in municipal waste management systems can be overcome by employing these innovative technologies.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

Upon completion of literature review, we narrowed down the specifications and requirements based on the existing technology and the circumstances in which our project is to be utilised. Such requirements represent the constraints of deployment (real-time work with cluttered environments) as well as between versions (similar training and inference conditions).First, we require a deep learning based object detection model, using computer vision techniques for real-time classification of images into seven categories: organic, plastic, glass, paper, metal, medical, and sharp objects. The model aims for high precision and recall, as well as more than 90% mAP. The dataset that will be used to train the model must be balanced, accurately labelled and representative of Bangladesh's municipal waste.

The dataset used to work with consists of 2,444 images having 24,715 YOLO-formatted object annotations across seven classes, divided into 70/15/15 (train/val/test). Policy - labels are strictly annotated (tight box, overlap-dealing), and images are screened based on licensing/ethics which are web sourced. During training and inference, data will be saved in the convention of Ultralytics folder (images/, labels/) and re-scaled to 640x640 by letterboxing.

The system detects seven waste superclasses (*glass, medical, metal, organic, paper, plastic, sharp-object*) and supports query-based subclass localization (e.g., *knife, nail, syringe, test tube*). The detector is **YOLOv8m** (COCO-initialized) trained for 50 epochs at 640×640, with training-time augmentation only; post-processing is fixed at `conf=0.25` and `NMS IoU=0.45`. Target performance focuses on maximizing $\text{mAP}_{50/95}$ while maintaining a latency of $\leq$ 10–12 ms/image on a single GPU, along with class-wise AP, precision/recall, PR curves, and confusion matrices. A single evaluation approach is adopted for all models, ensuring identical data splits, post-processing settings, and training schedules. To maintain reproducibility, all artifacts—including confusion matrices, per-class AP, and PR curves—are preserved.

The **SAFE-R stage** reranks YOLO proposals using **CLIP** (positive/negative prompts, light shape prior) to improve hazardous subclass precision, with **OWL-ViT** used as a comparative baseline. It employs curated positive and negative prompts along with a light aspect-ratio prior to execute CLIP (`ViT-B/32`) for subclass retrieval. Only YOLO proposals are utilized to preserve the detector's runtime and recall.

Software Requirements: Python ≥ 3.10, PyTorch ≥ 2.1 with CUDA 11.8/cuDNN 8, and Ubuntu 22.04 LTS or Windows 11 comprise our software stack. For analysis, we make use of `NumPy`, `Matplotlib`, `CLIP/transformers`, and `Ultralytics YOLOv8` (8.x).

Hardware Requirements: One CUDA-capable GPU (tested on NVIDIA GeForce RTX 3070 Ti, 8 GB VRAM) with 16 GB RAM and at least 50 GB of free SSD storage. We utilize a deterministic dataloader, fix random seeds, and version all configurations (weights, thresholds, and other run settings) to ensure reproducibility. At 640×640 resolution, inference latency remains below 10–12 ms per image on the above GPU.

The dataset follows a 70/15/15 train/validation/test split with YOLO-format annotations.

3.2 Societal Impact

Waste management is chaotic and unsafe in Bangladesh, particularly in cities such as Dhaka and Chattagram. Home, hospital, market, and industrial waste are generally not sorted out beforehand, so collection and disposal of these types of waste is inefficient and dangerous. The most vulnerable people in the process are informal waste-pickers and landfill workers, most of whom do not have the right protective gear. They are exposed to infected objects like the used masks, gloves, syringes, and broken glasses on a daily basis, which can result in injuries and the spread of life-threatening illnesses like Hepatitis or HIV.

The solution to this societal challenge that our proposed system proposes is to provide automated identification of waste into seven categories (plastic, glass, metal, paper, sharpened objects, medical and organic). In addition to classification, the pipeline is able to identify harmful subgroups, for instance, from the medical waste it can identify syringes. This helps to protect frontline waste workers by avoiding direct contact with underlying sharp and infectious materials.

On a larger scale, effective segregation helps to support recycling sectors, enhance dignity and safety of waste-pickers and create a channel through which waste management in Bangladesh can be formalized. Eventually, this helps make the urban environment cleaner, livelihoods are safer and public health is more secure.

3.3 Environmental Impact

The effect of environmental implications of this technology is equally significant. In Bangladesh, unsegregated dumping now contaminates the soil and groundwater by leachate and causes air pollution by burning waste in the open air. Ecosystems are degraded by medical and plastic wastes, especially, which contain toxic chemicals and microplastics.

Our pipeline will minimize the amount of hazardous waste that makes it to the landfills and water bodies since waste is categorized and isolated at an early stage by the pipeline. Plastic and glass can be directed to be recycled, organic waste can be composted and the dangerous sharps or medical waste can be burned under controlled conditions. This reduces unwanted burning and reduces the mixture of greenhouse gasses.

The bigger picture is how the system relates to the United Nations Sustainable Development Goals (SDG 11: Sustainable Cities and Communities and SDG 13: Climate Action) as the system leads to cleaner cities, safer waste generation, and climate-sensitive disposal practices. Not only is it a technical solution but a direction to sustainable urban development.

3.4 Ethical Issues

The ethical considerations on the implementation of automated waste classification systems should be strong. To start with, any images or information used in training and testing should not violate privacy and appropriate anonymization and consent should be used should they be captured in sensitive or semi-private places. This is to make sure that the system does not inadvertently violate the rights of people.

Second, although automation can enhance safety and efficiency, it should not eliminate the livelihoods of informal waste workers who are already a major component of the waste system in Bangladesh. Implementing ethics should hence include reskilling programs where workers can be incorporated into safer areas like system operation, monitoring and maintenance rather than being an outsider to the process.

Lastly, there should be fairness and reliability of the model. The data collection must be non discriminatory in that it includes different types of waste, environmental factors, and rural and urban situations to avoid any bias. Otherwise, the system might not be able to detect certain objects or will not work in various situations. Ethical care will make the model resistant, socially just, and contextually relevant to the waste management issues in Bangladesh.

3.5 Project Management Plan

To keep the progress under control and easy to understand, we divided the work into five milestones: (P1) the curation and labeling of the data (policy, split, QA), (P2) the training and validation on the YOLOv8m (baseline), (P3) the architectural ablations (P2, BiFPN, CBAM), (P4) the writing of the SAFE-R pipeline by using YOLO+CLIP with prompts and score fusion, and (P5) the ultimate evaluation and thesis packaging (figures, tables, error analysis). The tasks were monitored in a Kanban board; at the conclusion of each stage, an artifact review was completed (metrics, confusion matrices, qualitative montage). The budget had been assumed for one CUDA-enabled workstation (8–12 GB VRAM), open-source framework (Ultralytics/PyTorch/CLIP), and no commercial datasets. Our resources were allocated based on GPU time, with bulky training being done overnight, and light analysis being done during the day; all experiments (configs, seeds, thresholds) were versioned in order to make findings reproducible and easy to compare between variants.

3.6 Risk Management

We were aware of a number of technical threats at the beginning: strong class imbalance (medical/sharp under-represented), false positives in the background in cluttered images, overfitting with the larger variants, compute and data/licensing limitations. Our way of dealing with the imbalance was a hard policy in terms of annotation, and application of moderate augmentation. We used a conservative operating point, that is, `confidence=0.25` and `NMS IoU=0.45`, and overlaid a SAFE-R reranker on this to eradicate spurious detections. To control our overfitting, we reformatted the training protocol and slated negative results instead of performing tuning on a model-by-model basis. In order to keep our compute/time constraints down, we executed small, well-scoped experiments and used the checkpoints of all models that were pretrained on COCO. On the information part, we filtered out only the allowed sources and deleted sensitive photos to comply with the ethical standards. We still have certain risks, like the transparent materials (glass, in particular) and the domain shift, however, we capture them and offer feasible solutions such as the increased data collection, calibration of scores, and class-conscious NMS.

3.7 Economic Analysis

The biggest costs are the GPU time (electricity and amortized hardware) and student labor used to annotate/experiments; the software is open-source. The system is able to reduce the risk of injuries (sharp/medical exposure) compared to sorting alone, accelerate the sorting throughput, and enhance the yield of recycling (cleaner streams), leading to downstream savings on the part of municipalities and privately recycled materials. Depending on the operating point selected (around 10–12 ms per image), real-time operation is possible on either a single workstation or a small edge GPU to maintain the hardware cost at low levels. The SAFE-R stage enhances the SAFE-R re-training without retraining the hazardous subclasses resulting in low maintenance expenses. And at last, scale is what will make the payoff: a small reduction in hazardous or mis-sorted material reduced number of incidents, reduced sick leave, clean streams, etc. can pay back the computation costs in safer operations and more valuable recyclables.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

Our methodology was guided by two key objectives: (i) reliably detecting seven waste superclasses in cluttered images, and (ii) retaining a real-time throughput for deployment in real-world scenarios. The first step included compiling our multi-source dataset with strict annotation rules for minimum label mismatch and background confusion. The images were resized to 640x640, while small-object fidelity was maintained through controlled augmentations (mosaic 0.5, mixup 0.1, mild geometric and HSV jitter). This dataset was the basis of our hypothesis-driven modelling cycle where our focal objectives guided the ablation of model architecture.

To begin with, we chose YOLOv8m (pretrained on COCO) because of its promising balance of both accuracy and latency. The training parameters were kept constant across all variants to reliably isolate the effects of architectural modification, specifically at 50 epochs, batch 8, SGD with lr0=0.01 and weight decay 0.0005, 3-epoch warmup, AMP enabled. Initial assessment of the validation set revealed high confusion between object and background as well as many misses on small, elongated hazardous objects (like knives, needles, syringes). Thus, we focused our architecture modifications on overcoming this limitation rather than completely swapping models.

Mainly, we tried three modifications. First, a P2 head was integrated into YOLOv8 for smaller spatial scales, to test whether more low-level resolution would succeed in detecting tiny metal or sharp objects (e.g. nails, syringes, etc). Secondly, the default PAN neck was replaced with a BiFPN layer augmented by ECA attention to investigate whether more salient features could be identified by this fusion despite the cluttered images. Lastly, we incorporated Residual Convolutional Block Attention Module (ResCBAM) blocks in the neck to further make it more "context-aware" through channel attention and spatial attention helping identify and localise more important features. For comparison, we also evaluated the performance of YOLOv12m to gauge the tradeoff between accuracy and latency. All adjustments were made on the COCO-pretrained YOLOv8m model.

During post-processing, we set the confidence threshold to 0.25 and opted for NMS IoU = 0.45 in order to decrease the number of almost identical bounding boxes in densely cluttered scenes without compromising recall. These settings were kept con-

stant during comparative evaluation to make sure that changes in mAP, precision, recall, and per-class AP were reliably a result of model modification.
In order to go beyond superclass detection, we designed a query-aware reranking pipeline (SAFE-R) for subclass discernment for hazardous objects (for instance, distinguishing a syringe from a test tube within the medical-waste class). The YOLOv8m detector would propose high-recall bounding boxes at low confidence which would be class-gated by the relevant superclass; only candidate cropped regions would then be scored by CLIP (Contrastive Language-Image Pretraining) against positive and negative text prompts. Upon fusing the scores with the detector's confidence, only query-consistent boxes would be retained. This pipeline was expected to leverage the strong localisation ability of YOLOv8m while introducing semantic flexibility through CLIP, achieving more fine-grained labels without retraining.

The above approaches outline our incremental, evidence-driven process in testing targeted hypotheses guided by the objectives to retain the ideal balance between accuracy and speed, as well as introducing a more robust query-aware semantics to add measurable value to our waste-classification pipeline.

4.2 Model Specification

During the design phase, we considered several object detection architectures including YOLO, EfficientDet, Faster R-CNN, SSD and Detectron2. YOLOv8 and YOLOv11 were considered particularly suitable for the proposed task as they are optimized for real-time detection while maintaining a strong balance between speed and accuracy, making them effective for identifying many small and overlapping waste objects. EfficientDet, with its compound scaling and BiFPN structure, is highly efficient in resource-constrained environments and can identify small objects with relatively low computational cost. Faster R-CNN provides strong detection accuracy, especially for small objects, but its slower inference speed limits its applicability in real-time scenarios. SSD offers fast detection with acceptable accuracy, making it useful for real-time tasks, although it tends to underperform on small or compact waste objects compared to other models. Detectron2 is a flexible framework that supports both detection and segmentation, but its complexity and higher computational requirements reduce its suitability for lightweight, real-time applications. After analyzing these alternatives, the model ultimately adopted in this research is YOLOv8m, selected for its proven balance of speed, accuracy, and practicality for waste detection.

4.2.1 YOLOv8m

The starting point was an anchor-free YOLOv8m model consisting of 3 main components: a backbone, neck and decoupled detection head.

- YOLOv8 comprises a sophisticated **C2f backbone** with an **SPPF block** at the end for extracting multi-scale features from input images, designed to minimize computational overhead while retaining representational power. All the C2F blocks are built from Convolution, BatchNorm and SiLU(CBS) blocks. At the end of the backbone, Spatial Pyramid Pooling - Fast (SPPF) block aggrigates multi-scale context feature.

- The **neck module** is an optimized version of the Path Aggregation Network (PANet) structure that combines top-down and bottom-up feature fusion (similar to FPN + PAN). It receives multi-scale feature maps from the backbone (P3, P4, P5) and refines them for small, medium, and large-object detection, while optimizing memory and computational efficiency.
- The **decoupled-head module** uses three detection heads (P3–P5) to generate the final predictions, including bounding box coordinates. Unlike YOLOv5, in YOLOb8 decoupled-head module, classification and box regression are separated branches. Aditionally, the head predicts center points and offsets instead of anchors, using BCE + CIoU + DFL.

This configuration provided the ideal balance between accuracy and latency based on our dataset, with robust localisation in cluttered scenarios, competitive mAP, and sub-10-12ms/image inference on our GPU. The baseline was the default deployment model as well as the proposal generator for the SAFE-R pipeline.

4.2.2 YOLOv8m + P2 head

To make the model more receptive to objects with thin edges (like, nails, syringes, etc), we extended the baseline with a P2 detection head where the default YOLOv8 architecture only predicts at three pyramid levels: P3/P4/P5. Adding the fourth head at P2 level allows a higher-resolution feature extraction for the very small objects prevalent in our dataset. We expected greater recall on small objects with this enhancement. The rest of the model and training schedule were left unchanged.

4.2.3 YOLOv8m + P2 head + BiFPN

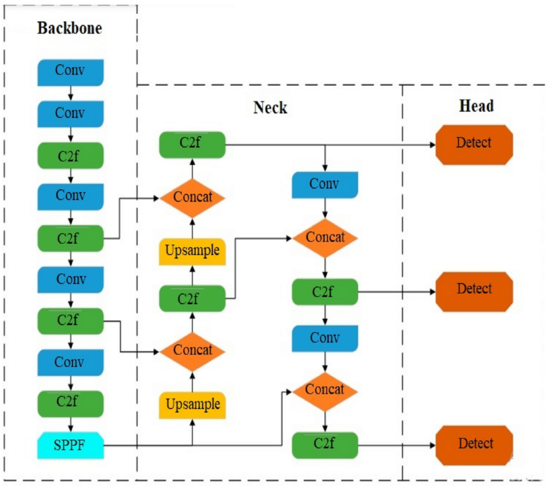
In order to compensate for the drop in accuracy, next, we tried replacing the standard PAN neck of YOLOv8 with a **Bi-directional Feature Pyramid Network (BiFPN)**. The BiFPN, augmented with ECA attention at fusion nodes, introduces weighted feature fusion, allowing the network to adaptively learn which feature scales contribute most to final predictions. This is particularly useful in scenes with dense, overlapping objects where different scales may contain complementary information. The hypothesis was that content-adaptive routing would propagate the correct scale(s) for elongated objects and decrease over-reliance on any one pyramid level, but at the risk of amplifying false positives in heavily cluttered situations due to capacity-data mismatch.

4.2.4 YOLOv8m + Residual CBAM

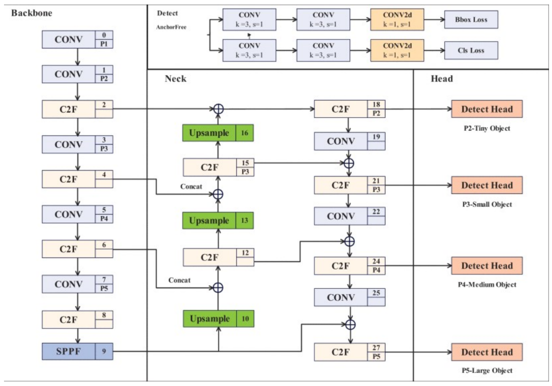
We also experimented with **Residual Convolutional Block Attention Module (ResCBAM)**, integrated in place of some convolutional blocks in the head pathway. ResCBAM combines channel and spatial attention, encouraging the model to highlight both important feature channels and spatial locations with minimal overhead. The hypothesis was improved class-dependent gains on hazardous regions, especially in cases where background clutter is high.

4.2.5 YOLOv12m

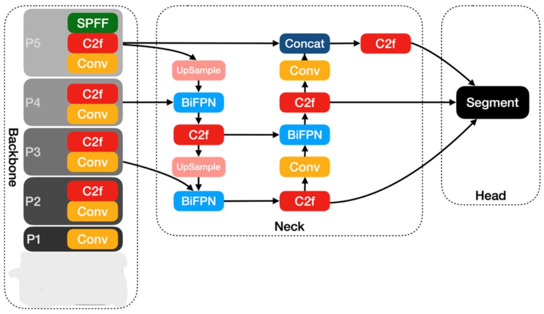
For in-context comparison with the baseline, the newer, larger configuration of YOLOv12m was included to establish an upper-bound reference for accuracy that could be achieved on our dataset. The aim was to quantify the tradeoff between accuracy and latency under the same conditions of preprocessing, augmentations, and post-processing.



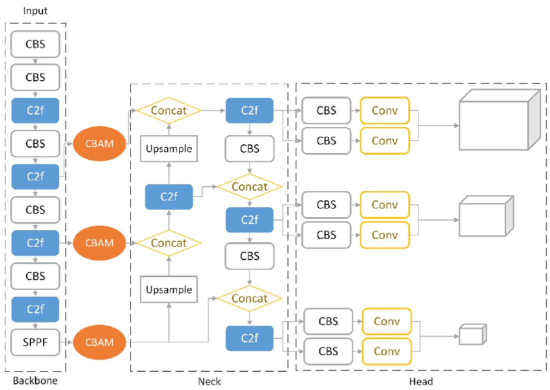
(a) YOLOv8 from Elhanashi et al., 2024 [31]



(b) YOLOv8+P2 from Fang et al., 2025 [36]



(c) YOLOv8+BiFPN from Wang et al., 2024 [35]



(d) YOLOv8+CBAM from Liu et al., 2024 [34]

Figure 4.1: YOLOv8 architectures.

4.2.6 Post-processing and Evaluation

For all models, we used the same inference settings for reliable comparison: confidence = 0.25, NMS IoU=0.45, max_det = 300. For performance evaluation, metrics included mAP50/95, precision, recall, per-class AP, latency, confusion matrices, and qualitative audits; the SAFE-R pipeline was analysed separately to determine sub-class behaviour.

4.3 Data Collection

Our custom dataset comprises **2,444 images** with **24,715 bounding-box** annotations across **seven classes**: glass, medical, metal, organic, paper, plastic, and sharp-object. Images were collected from three sources:

- Smartphone self-taken pictures at local waste sites
- Python scripts to web-scrape Google images
- The dataset that was annotated before in YOLO format on Kaggle (five classes) [38]

We tried to ensure the data reflected geographically-relevant, real-world waste-sorting conditions (indoor/outdoor, cluttered backgrounds, varied lighting, etc). They were organized under the standard Ultralytics directory convention (train/valid/test with YOLO TXT labels). Class distribution prior to any augmentation is imbalanced (e.g., organic 5,587; plastic 5,053; paper 4,131; metal 3,998; glass 2,670; medical 1,862; sharp-object 1,414), which motivated future decisions on augmentation approaches and evaluation emphasis for the hazardous categories, namely medical, sharp-object and glass.

4.3.1 Data Cleaning and Annotation

A thorough cleaning procedure was undertaken to guarantee quality data. Blurred, low-quality and redundant photos were deleted. Poorly lit images or images that had blurred object boundaries were also eliminated.
To ensure proper labelling and bounding box quality assurance, we followed the steps proposed by Pičuljan and Car (2023) [27] wherein first, it was determined that the images collected complied with all regulations, with the data being representative of the real-world scenario without irrelevant data clogging up the set. Next, the structural attributes like image size, quality, etc, were visually inspected for data quality assurance, images that did not meet the standards were removed. Following that, we started labelling the images (and relabelling existing dataset images) according to the seven classes previously mentioned. Extra care was taken for annotation, especially for images that included many overlapping objects, ensuring that the bounding box properly enclosed the object without leaving too much space around it.

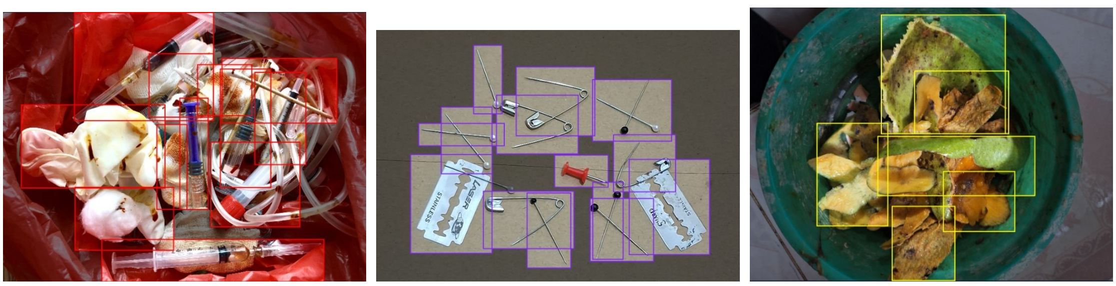


Figure 4.2: Some annotated images from our custom dataset.

4.3.2 Data Transformation

During preprocessing, we opted for letterbox resizing for all images in the dataset to the size 640x640 to retain the origin aspect-ratios and spatial context. Moreover, we used the normalisation pipeline native to Ultralytics/YOLO during training. For data augmentation, we attempted various light-weight image processing approaches like grayscale → CLAHE, RGB → Laplacian but these did not produce consistent gains and ultimately, weren't used in the primary training set. The decision was made against using any offline data augmentations. On the contrary, we adopted training-time augmentations like mixup, mosaic, mild geometric and HSV jitter for incorporating diversity without compromising the integrity of small objects.

4.3.3 Data Integration

All labels and images were incorporated under a single taxonomy in data.yml (nc: 7, names: 'glass','medical','metal','organic','paper','plastic','sharp-object'). During compiling the images for multiple sources, we synchronized the label conventions, resolving issues with synonymous classes to ensure that any external annotation aligned with our superclasses as well as the YOLO TXT format (class id + normalised box). The unified structure projected the train, val, test sets to their respective images/ folders along with the labels/ folders with mirrored filenames, enabling easy reproducibility across models.

4.3.4 Data Reduction

To prevent overfitting and reduce redundancy, we deleted near identical images and prioritised those with hazard-rich content for validation/test sets, and retained only one representative from many similar image clusters. We decided against any form of aggressive dimensionality reduction or offline cropping to ensure global context that would aid in detection amongst clutter. We relied on augmentations and class-aware sampling in an effort to undercut the imbalance in class distribution.

4.3.5 Summary of Preprocessed Data

The final dataset used was subjected to 70/15/15 split for train/validation/test with augmentation applied only during training. Before augmentation, the number of annotations per class stood at the following: 2,670 glass; 1,862 medical; 3,998 metal; 5,587 organic; 4,131 paper; 5,053 plastic; and 1,414 sharp-objects. This brought the total count to 24,715 annotations in 2,444 images. The labels were in YOLO-format TXT files paired with each image. The goal of the chosen configuration along with the letterbox resizing and augmentation strategy was to provide a stable and reproducible base for any architectural experimentation as well as for being the foundation of any waste-related computer vision initiative.

4.4 Implementation of Selected Design

4.4.1 Initialization and Training Schedule

The YOLOv8m detector was not trained from scratch, rather warm-started from COCO pretrained weights to provide a robust foundation for general object detection, which would stabilise early training, especially in terms of minority classes like medical and sharp-object, and reduce data hunger. To ensure reliability in comparison of results across ablations, the schedule and parameters were kept constant at the values enumerated below.

Parameters	**Values**
Weights	yolov8m.pt and variants
Epoch, Batch, Image Size	Epoch = 50, Batch = 8, Image Size (*imgsz*) = 640
Warmup	3 epochs
Optimizer	SGD
	Momentum = 0.90, Weight Decay = 0.0005
Learning Rate	Initial $lr_0 = 0.01$ with linear decay schedule
Loss	Box = 7.5, Class = 0.5, DFL = 1.5
GPU	NVIDIA GeForce RTX 3070 Ti

Table 4.1: Training Configuration and Parameters

The configuration was chosen for the optimal balance between convergence stability, minority-class recall, and throughput on our hardware.

4.4.2 Data Augmentation (Training-time)

To promote diversity without degrading small, thin objects (for instance, needles, syringes, nails), we chose the following augmentations:

- Geometric: mosaic 0.5, mixup 0.1, degrees 10°, translate 0.10, scale 0.50, shear 2.0, perspective 0.0005, horizontal flip 0.5.
- Color: HSV jitter (h=0.015, s0.6–0.7, v0.4).
- Resizing: letterbox to 640×640 (no aspect-ratio stretch)

The selected augmentation strategy retains the integrity of small, elongated objects by avoiding heavy warping or cropping. This was applied consistently across all iterations for reliable comparison of model performance.

4.4.3 Architectural Variants

The baseline configuration is a standard YOLOv8m detector with a C2f backbone that ends with an SPPF block, a regular PAN/FPN neck, and three separate detection heads working at P3, P4, and P5. In our YAML, the head uses features from layers [15, 18, 21] and sends them to Detect. We selected this model as the anchor because it consistently balanced accuracy and latency on our dataset while remaining stable under moderate augmentation. It also acted as the proposal generator for the subclass pipeline in later experiments, providing a consistent basis for analysis. The training and validation loss curves for baseline YOLOv8m model are included below (Fig. 4.3) (Fig. 4.4).

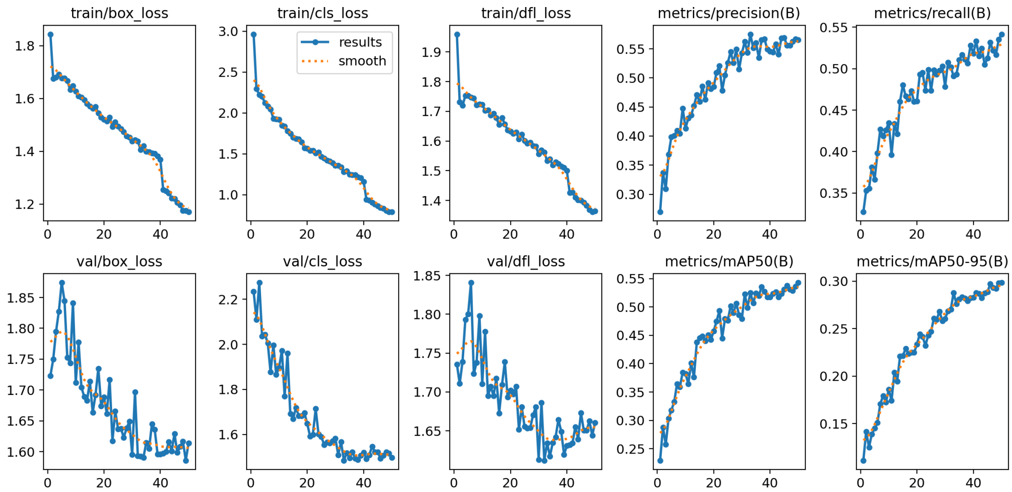


Figure 4.3: The training and validation loss curves for **baseline YOLOv8m**.

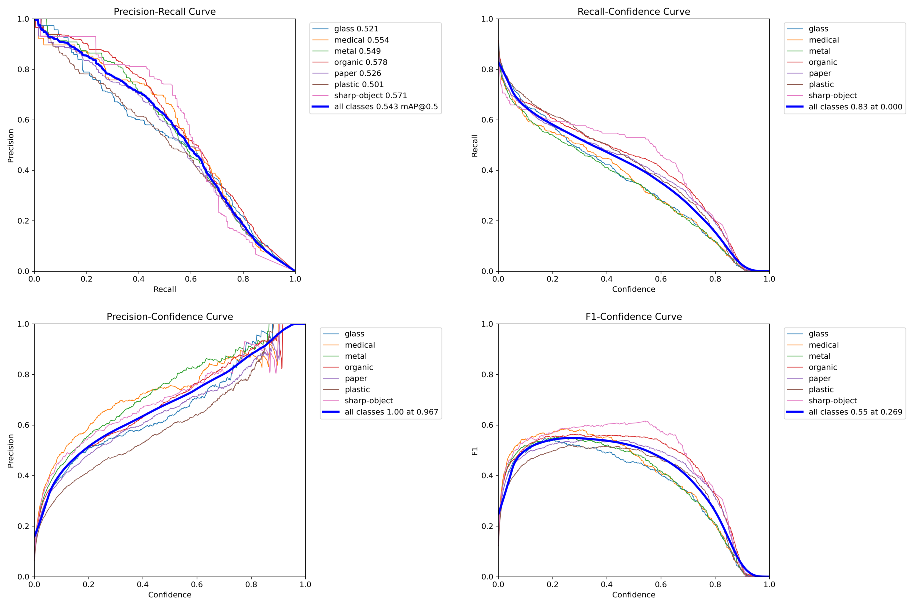


Figure 4.4: Confidence–performance Curves for **baseline YOLOv8m**.

To address small-object misses, especially for narrow, elongated hazards like nails and syringe shafts, we added a P2 head on top of the baseline model. Specifically, we extended the neck by adding an upsample-and-merge path to expose higher-resolution features, resulting in a four-head detector at P2, P3, P4, and P5. In the corresponding YAML, Detect takes features from [18, 21, 24, 27]. We believed that earlier pyramid features would recover small structures poorly captured at P3, which would improve recall for ”needle-like” objects. The known risk was that without enough clean and plentiful small-object supervision, the extra head could introduce noisy activations and lower overall mAP.

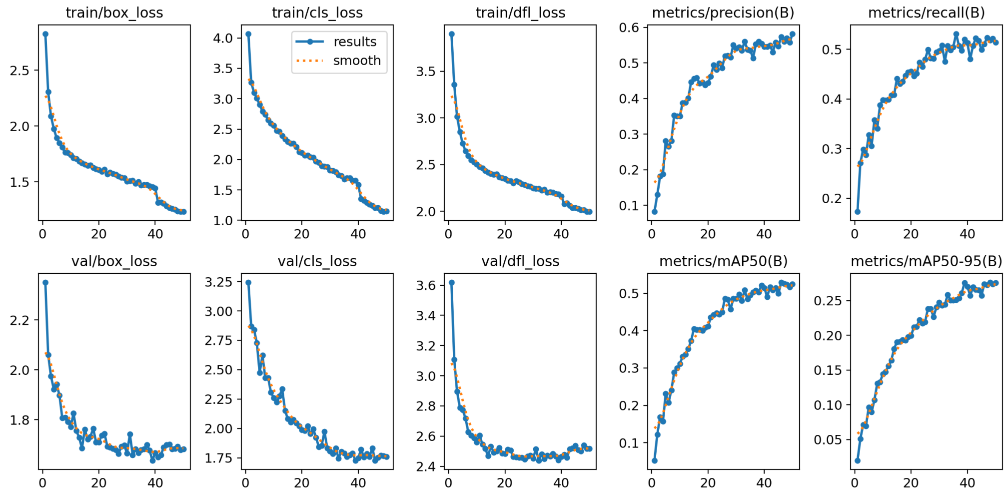


Figure 4.5: The training and validation loss curves for **YOLOv8m with P2 extension**.

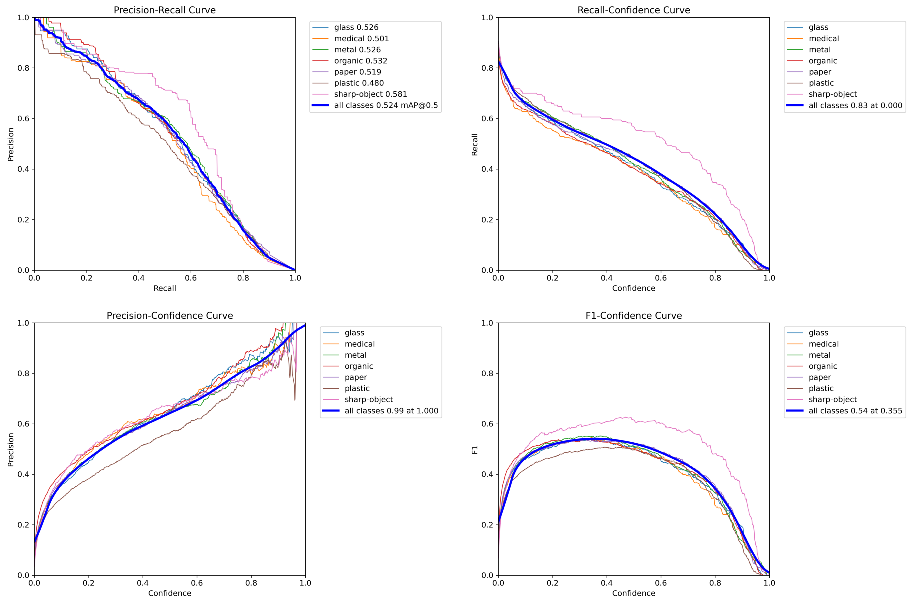


Figure 4.6: Confidence-performance Curve for **YOLOv8m with P2 extension**.

For the second variant, we replaced the PAN neck with a BiFPN module that performs normalized, learnable weighted fusion across scales, reinforced with ECA attention at the fusion nodes. The inputs to the BiFPN were the P2 to P5 features provided by the neck, and the fused outputs were split back into four streams for the detection heads. In our YAML, Detect uses the post-fusion tensors labeled as [29, 30, 31, 32]. We hypothesized that content-adaptive routing would highlight the right scale(s) for elongated tools and lessen the reliance on any single pyramid level in cluttered scenes. The risk was a mismatch between capacity and data, where the added fusion flexibility could overfit or amplify background responses when hazardous samples are limited.

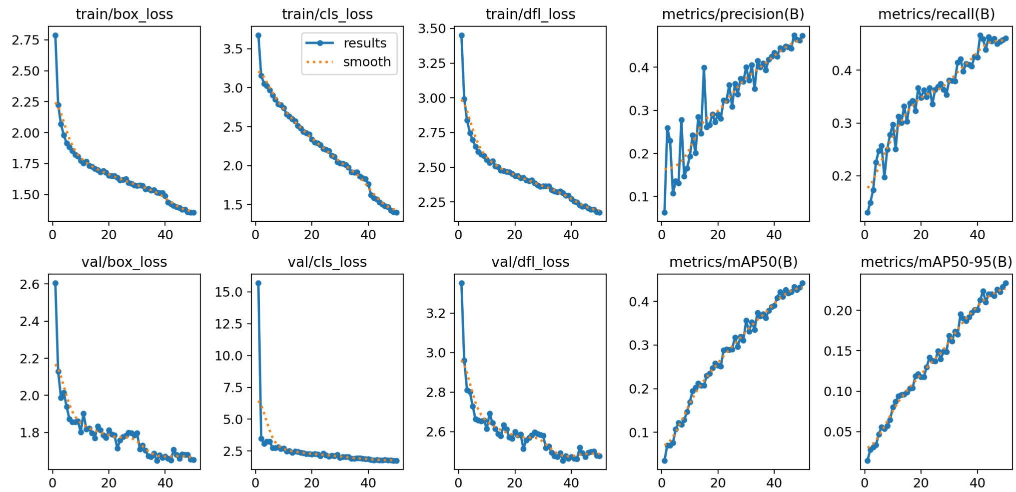


Figure 4.7: The training and validation loss curves for **YOLOv8m with P2 extension and BiFPN (+ECA) neck**.

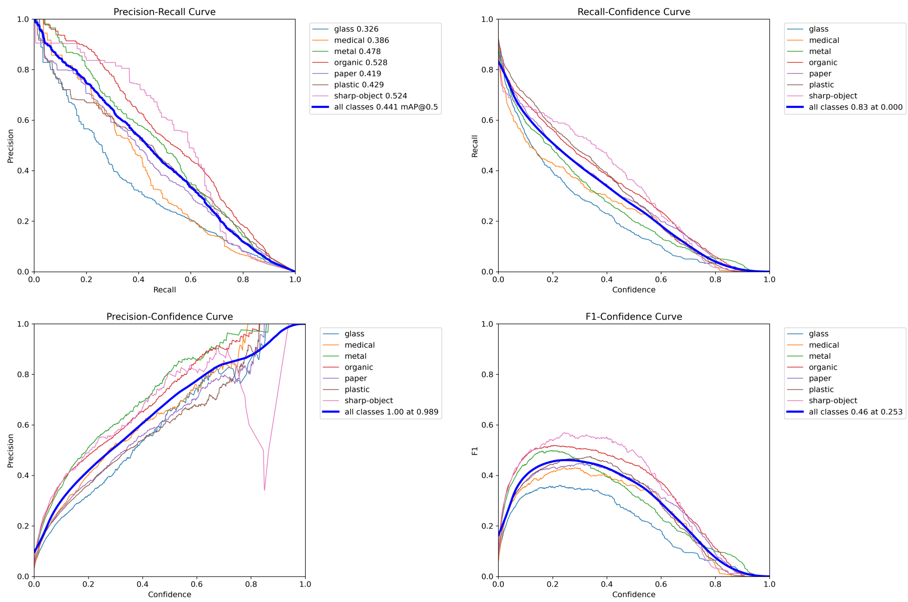


Figure 4.8: Confidence-performance Curve for **YOLOv8m with P2 extension and BiFPN (+ECA) neck**.

A third variant added lightweight saliency through Residual CBAM blocks along the head pathway while keeping the baseline three-head structure intact. Essentially, we inserted three RES-CBAM modules before Detect, and the detector used features from [17, 21, 24]. This setup aimed to reduce recurrent distractions—such as wires, rods, and glossy textures—by improving channel and spatial selectivity without changing the scale hierarchy. We expected the results to vary by class: if background confusion was mostly representative, CBAM could enhance contours and provide minor gains; if confusion was driven by annotations or data issues, results would likely revert to baseline performance.

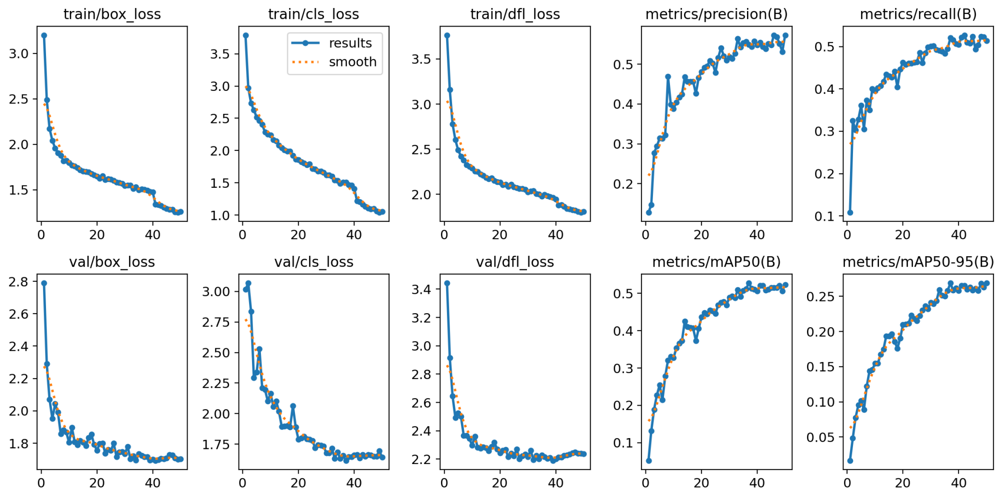


Figure 4.9: The training and validation loss curves for **YOLOv8m with ResCBAM module**.

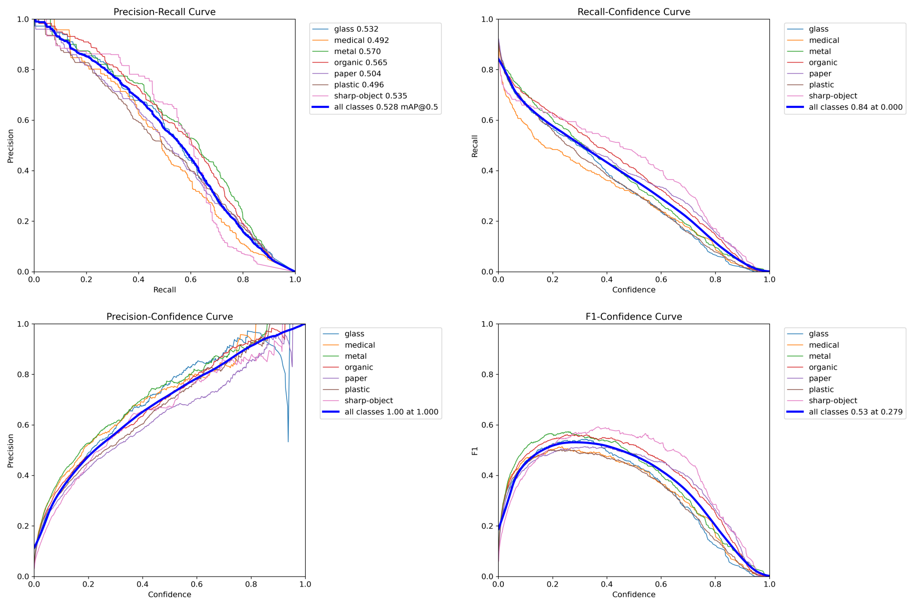


Figure 4.10: Confidence-performance Curves for **YOLOv8m with ResCBAM module**.

Finally, YOLOv12m was included as a reference for capacity to understand gains due to model size rather than changes in architecture. All preprocessing, augmentation, and post-processing settings remained unchanged so that any differences would mainly reflect capacity and inductive bias. Across all variants, warm starts from COCO within the same family were used to retain the advantages of pretraining and maintain uniform initialization effects.

4.4.4 Post-processing / Inference Settings

We used the baseline model to tune the post-processing settings then kept it constant for all other modified iterations to avoid bias. The confidence threshold was set to

0.25 and non-maximum suppression used an IoU of 0.45– both to retain recall on small, narrow objects as well as to avoid near-duplicate boxes on densely cluttered scenes. Maximum detections per image were limited to 300 to avoid crowding. To keep latency representative of deployment, we disabled test-time augmentation. Inputs were also letterbox resized to 640x640 to mirror the training images and eliminate the risk of aspect-ratio disruptions to narrow, elongated objects.
The aforementioned configuration was kept constant for YOLOv8m, the P2 enhancement, the BiFPN with ECA fusion model, the ResCBAM alteration, as well as the YOLOv12m reference. As such, it was assured that the performance evaluation and metrics reflect changes in architecture rather than mismatch in settings.

4.4.5 Query-Aware Reranking (SAFE-R: YOLO + CLIP)

The deeper distinction of subclass object identification in cluttered waste scenes, for instance, distinguishing a syringe from a test tube amongst the medical waste superclass– was challenging to achieve on purely detector modification or enhancement. Hence, we opted for a two-stage pipeline that leverages a proposal → rerank design in which the YOLOv8m detector first generates high-recall proposals, allowing CLIP to use text-conditioned semantics to produce query-consistent boxes.
CLIP was developed with the intention of integrating object detection and image captioning to extract information about object categorisation and location. With two distince encoders sharing the same embedding space, one for images and one for text, CLIP is capable of zero-shot classification of images by looking at the similarities between texts and images [16]. We leveraged this for our proposed pipeline intended to further distinguish hazardous and harmless objects amongst superclasses.
Firstly, YOLO runs to maximise candidate recall and generates a set of bounding boxes. Next, the superclass filter ensures only those from the relevant category are propagated, thereby omitting the obvious out-of-scope candidates, i.e. “knife” query limits the scope to the “sharp-object class” while “syringe” limits it to the “medical” class. Crops from the remaining candidates are encoded using CLIP’s image encoder, while the given query is projected onto a positive prompt set consisting of synonyms and phrasings, and a negative prompt set consisting of visually confusing non-targets. For each candidate, the text-image similarity is scored using the following equation:

$$s_i^{\text{clip}} \;=\; \max_{t \in \mathcal{P}} \,\cos\big(f_{\text{img}}(b_i),\, f_{\text{text}}(t)\big) \;-\; \beta \,\max_{t^- \in \mathcal{N}} \,\cos\big(f_{\text{img}}(b_i),\, f_{\text{text}}(t^-)\big).$$

The scores must be normalised that the detector and CLIP scores operate on differing scales. The equation to apply normalisation to each channel and compute a balanced fusion:

$$\hat{s}_i \;=\; \lambda\, \widetilde{s}_i^{\text{clip}} \;+\; (1-\lambda)\, \widetilde{s}_i^{\text{yolo}}.$$

Lastly, the final outcome simply thresholds the fused score and outputs the resulting boxes in descending order of fused score.

$$\text{keep } b_i \;\text{ if }\; \hat{s}_i \geq \tau$$

(we use tau 0.50) This reranking stage is meant to constrain the localisation of the detector rather than replace it. Multiple considerations prove the combination to be precision-stable, namely the detector's anchor of geometry and recall being supplemented by CLIP's text-image matching based on query-specific distinction allowing to circumvent previously prevalent confusions (e.g. shiny rods mislabelled as knives, or narrow glassware misclassified as test tubes).

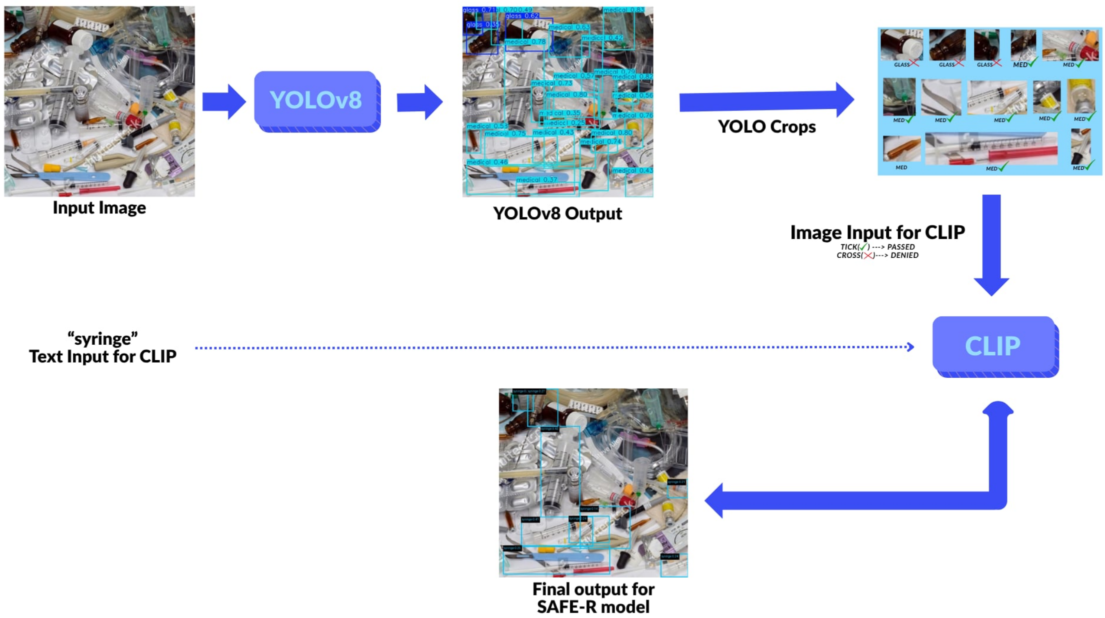


Figure 4.11: SAFE-R pipeline flowchart.

4.4.6 Design Rationale

Performance evaluation revealed baseline YOLOv8m yielded the most reliable balance on our dataset following careful data cleaning, augmentation and post-processing. To summarise, in spite of the limited boost in small-object identification with the P2 head extension, it compromised over mAP. Replacing the PAN with BiFPM increased degrees of freedom without consistent gains under clutter. Similarly, despite improvement in attention to salient contours with ResCBAM implementation, there was no significant reduction in object as background confusion, maybe as a result of representational scarcity. Therefore, for the query-aware pipeline, we opted for the YOLOv8m detector integrated with CLIP for subclass distinction. By utilizing a constant training schedule and post-processing steps across architectural variants, we avoided bias in evaluating the performance metrics. The resulting pipeline is easy to explain, cheap to implement, and effective where it matters– making it especially suitable for deployment in real-world situations, specifically in our country.

Chapter 5

Result Analysis

5.1 Performance Evaluation

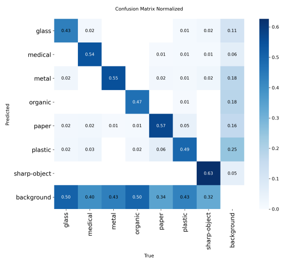
According to the configuration described in section 4.4, we evaluated the performance of the models using the following metrics: mAP50/90, precision/recall, and class-wise AP for accuracy; per-image latency for efficiency, and confusion matrices and qualitative audits for reliability.
On the test split, the baseline YOLOv8m achieved mAP50 $\approx$ 0.591, mAP50-95 $\approx$ 0.332, Precision $\approx$ 0.642, Recall $\approx$ 0.510, with inter-class differences reflecting stronger performance on sharp-object and paper compared to plastic and metal, likely due to intra-class variation in data.

Class	**Images**	**Instances**	**Precision (B)**	**Recall (B)**	**mAP@0.5**	**mAP@0.5-0.95**
All	377	4546	0.642	0.510	0.591	0.332
Glass	35	654	0.700	0.393	0.563	0.308
Medical	59	331	0.693	0.526	0.623	0.356
Metal	65	885	0.624	0.446	0.552	0.319
Organic	90	1024	0.640	0.464	0.572	0.327
Paper	130	632	0.584	0.578	0.608	0.358
Plastic	122	795	0.541	0.507	0.518	0.293
Sharp-Object	40	225	0.712	0.658	0.702	0.362

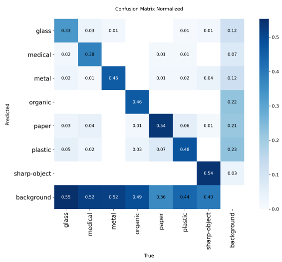
Table 5.1: Baseline YOLOv8m Test Results on the Waste Dataset.

Qualitative analysis revealed two main areas of failure: (i) high confusion between object and background in cluttered scenes and (ii) recall sensitivity for small, narrow hazardous items (like syringe, nail).

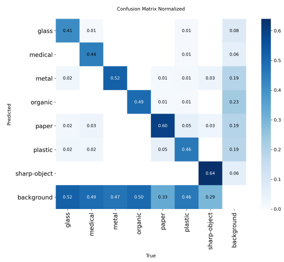
Figure 5.1: Normalized confusion matrix for baseline YOLOv8m.



(a) YOLOv8m + P2



(b) YOLOv8m + P2 + BiFPN + ECA



(c) YOLOv8m + ResCBAM

Figure 5.2: Comparison of normalized confusion matrices for YOLOv8m architectural variants. Each matrix shows per-class precision–recall distribution normalized across all test instances.

Figure 5.3: YOLOv8m predictions.

5.2 Analysis of Design Solutions

With all the architectural variants undergoing training with the same configuration and inference settings, the results are compared below.

YOLOv8m+P2: The P2 enhancement for a fourth head modestly improved AP and recall for "metal" class but decreased overall scores with a latency increase. The metrics stood at $mAP_{50} \approx 0.569$, $mAP_{50–95} \approx 0.308$. Although gains were seen on texture-heavy small objects which aligned with our hypothesis, the same could not be said for other hazard classes, for instance, AP for sharp-object class decreased. The additional low-level scale introduced noisy activations where tiny, clean annotations were limited and the resulting trade-off was not promising for deployment.

YOLOv8m + P2 + BiFPN: Switching the PAN neck with BiFPN plus ECA attention yielded results inferior to the baseline while increasing latency: $mAP_{50} \approx 0.517$, $mAP_{50–95} \approx 0.283$. Although a small increase on metal AP was observed, it was offset by regressions on metal and plastic, likely due to the capacity–data mismatch in clutter as predicted in Section 4.2.3.

YOLOv8m + ResCBAM: This ablation produced results similar to baseline at comparable speeds: $mAP_{50} \approx 0.571$, $mAP_{50–95} \approx 0.300$. Unlike our expectations, it failed to reduce background confusion and did not yield significant gains for hazardous classes.

Table 5.2: Comparison of YOLO Model Variants on the Waste Dataset

Model	Precision	Recall	mAP@0.5	mAP@0.5–0.95
YOLOv8m	0.642	0.510	0.591	0.332
YOLOv8m + P2	0.597	0.520	0.569	0.308
YOLOv8m + P2 + BiFPN	0.581	0.452	0.517	0.283
YOLOv8m + CBAM	0.626	0.496	0.571	0.300
YOLOv12m	0.654	0.529	0.610	0.362

To summarise, under standardised conditions, YOLOv8m offers the ideal accuracy-speed balance. P2 modification helps specific small-object cases but lowers over mAP; BiFPN(+ECA) and CBAM do not yield net advantages on the dataset.
YOLOv12m improves accuracy due to its increased complexity, deeper backbone and by compromising latency. Its inclusion is purely comparative, to provide an upper bound for demonstrating the effect of network capacity on model performance. The heavy-weight model is unsuitable for real-time waste-separation task and hence, not part of our proposed design.

5.3 Final Design Adjustments

The configuration for YOLOv8m detector was ultimately set at `conf=0.25` and `NMS IoU=0.45` for three key reasons: (i) a threshold sweep revealed this as the Pareto point (mAP_{50} vs. FP load); (ii) it ensured the least number of near-identical boxes in cluttered scenes without compromising small, narrow objects; (iii) it maintains $\leq$ 10–12 ms/image throughput in deployment-like runs. We also conformed the artifact generation (confusion matrices, PR curves, qualitative montages) under this operating point to allow reliable comparison across models for bias-free analysis.
Furthermore, the query-aware reranking pipeline's detector was also anchored at the same baseline, so any improvements observed in subclass retrieval could be attributed to semantics and gathering rather than any changes in detector configuration.

5.4 Statistical Analysis

To quell the possibility of stochasticity influencing our architectural conclusions, we retained a fixed seed, a deterministic dataloader, and identical schedules across training of all ablations. Moreover, hyperparameter values were kept constant when all models were trained. To quantitatively assess the observed gains of our architectural variants, we chose to statistically compare them to the run-to-run mAP50 variability– only values obtained beyond the ±0.01–0.02 range were considered statistically significant.

YOLOv8m vs Architectural Variants

The aggregated results for statistical comparison are summarised in the table below. The baseline YOLOv8m served as the reference model, with mAP50 = 0.591 and mAP50-95 = 0.332.
The P2 extension marginally improved recall (+0.01) but decreased precision and overall mAP, indicating a recall-precision tradeoff for fine-scale feature additions. BiFPN resulted in the largest diminishing of mAP at -12.5%, while CBAM showed minor gain in precision but no qualitative improvement in mAP50 when compared to the baseline.

Model	Precision	Recall	mAP@0.5	mAP@0.5–0.95	ΔmAP@0.5 vs Baseline	Δ%
YOLOv8m	0.642	0.510	0.591	0.332	—	—
+P2	0.597	0.520	0.569	0.308	−0.022	−3.7%
+P2+BiFPN	0.581	0.452	0.517	0.283	−0.074	−12.5%
+CBAM	0.626	0.496	0.571	0.300	−0.020	−3.4%

Table 5.3: Performance comparison of YOLOv8 variants under identical training configuration.

YOLO+CLIP (SAFE-R) vs. OWL-ViT

Evaluation of subclass disambiguation was carried out using a set of eight hazardous and harmless objects more prevalent in our dataset (knife, nail, syringe, test tube, etc). The SAFE-R pipeline yielded substantially greater scores than the open-vocabulary OWL-ViT model across all metrics, as seen in the table below.

Query	**SAFE-R**		**OWL-ViT**	
	AP@0.5	**best_F1**	**AP@0.5**	**best_F1**
knife	0.5536	0.7353	0.4113	0.5146
mask	0.4182	0.5823	0.0000	0.0000
nail	0.4015	0.5688	0.2997	0.4527
plastic bottle	0.6414	0.7316	0.0887	0.1658
scissors	0.5934	0.6942	0.4692	0.5310
screw	0.3994	0.5502	0.0006	0.0141
syringe	0.2572	0.4891	0.0615	0.1667
test tube	0.2161	0.4118	0.0000	0.0000
Macro-Avg	**0.4351**	**0.5954**	**0.1664**	**0.2306**

Table 5.4: Comparison of AP@0.5 and best F1 scores for SAFE-R vs. OWL-ViT across all eight subclass queries.

With our sample size of eight subclasses, all paired differences favoured SAFE-R, resulting in a z-value of -2.52 and corresponding p-value 0.0059. This demonstrates clearly that there is a statistically significant difference between SAFE-R and OWL-ViT's performance. More particularly, that SAFE-R performs statistically better at identifying hazardous subclasses under uniform testing conditions.

Query	**SAFE-R AP@0.5**	**OWL-ViT AP@0.5**	$\mathbf{d = x - y}$	**rank**
knife	0.5536	0.4113	0.1423	3
mask	0.4182	0.0000	0.4182	7
nail	0.4015	0.2997	0.1018	1
plastic bottle	0.6414	0.0887	0.5527	8
scissors	0.5934	0.4692	0.1242	2
screw	0.3994	0.0006	0.3988	6
syringe	0.2572	0.0615	0.1957	4
test tube	0.2161	0.0000	0.2161	5
Macro-Avg	**0.4351**	**0.166375**		

Table 5.5: Wilcoxon signed-rank inputs: per-query AP@0.5, differences d, and ranks.

From the signed ranks:

$$S^+ = 36, \qquad S^- = 0, \qquad W = \min(S^+, S^-) = 0.$$

For $n = 8$ pairs,

$$\mu_W = \frac{8(8+1)}{4} = 18, \qquad \sigma_W = \sqrt{\frac{8(8+1)(2 \cdot 8+1)}{24}} = 7.14.$$

The test statistic and p-value are:

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{0-18}{7.14} = -2.52, \qquad p = 0.0059.$$

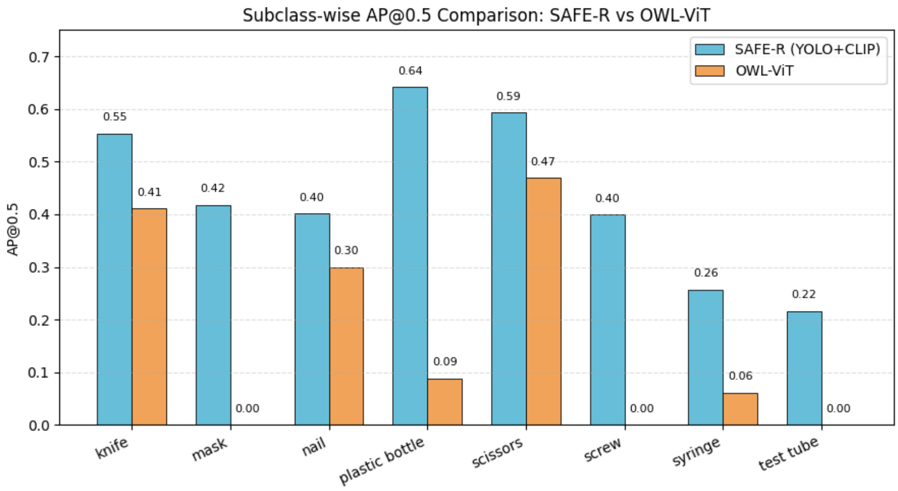


Figure 5.4: AP@0.5 comparison for SAFE-R vs. OWL-ViT.

5.5 Comparisons and Relationships

The following comparisons can be made of our ablations:

1. **YOLOv8 architectural relationships:** Increasing fusion complexity or adding head-level attention did not result in increased mAP in our cluttered dataset; in fact, in certain cases, it amplified false positives. In particular, the P2 head brought improvement to a subset of small object class (namely metal) but diminished sharp-object, clearly portraying the tendency of architectural gains being class-dependent and data-sensitive.

2. **Query-aware Reranking vs. Open-Vocabulary Baseline:** Using the pipeline wherein YOLOv8m detector yields proposals at low conf, undergoes superclass gating followed by CLIP scoring with positive and negative scores, resultant fusion shows substantially better performance in comparison to OWL-ViT on distinguishing subclasses. At IoU 0.5, YOLO+CLIP reached a macro-average of 0.4351 AP@0.5 in contrast to OWL-ViT's 0.166375 AP@0.5. Our hypothesis is validated by the Wilcoxon Signed Rank test enumerated

above, clearly demonstrating the pipeline's superior ability to precisely detect subclass objects without compromising recall.

Moreover, OWL-ViT is much more resource heavy in comparison to SAFE-R which uses 33% lower GPU utilization (SAFE-R GPU memory usage stood at 0.8 GB in contrast with OWL-ViT's 1.2 GB), a significant reduction that makes it more ideal for wide-scale real-life implementation. Besides this, computational advantages of SAFE-R still outweigh OWL-ViT at the constant input size 640x640: the average latency achieved by SAFE-R is 189 ms per image (mean FPS = 5.3) compared to OWL-ViT's 355 ms per image (FPS = 2.8). Due to the lightweight nature of both YOLO detector and CLIP reranking, our pipeline remains substantially more efficient, with approximately 26% of the total latency attributed to YOLO and 65% to CLIP.

Furthermore, the class-gating and CLIP's semantic advantage enabled SAFE-R to produce an average of 1.4 final detections per image from YOLO's initial 11.2 proposals, effectively filtering background clutter while preserving relevant hypotheses. In contrast, OWL-ViT yielded only 0.4 detections per image under identical conditions, demonstrating weaker localisation consistency as well as lower recall. Overall, we can conclude that SAFE-R allows a markedly better trade-off between accuracy, throughput, latency as well as memory efficiency while yielding precise subclass disambiguation in real-time detection scenarios.

(a) Knife (b) Mask (c) Syringe (d) Screw (e) Plastic Bottle

(f) OWL-ViT predictions

(a) Knife (b) Mask (c) Syringe (d) Screw (e) Plastic Bottle

(f) SAFE-R predictions

Figure 5.5: SAFE-R vs OWL-ViT predictions.

Table 5.6: Efficiency benchmark (@imgsz=640, query = syringe)

Model	Mean (ms)	p50 (ms)	p95 (ms)	FPS	Peak GB	Preds/img
SAFE-R (YOLO+CLIP)	189.4	52.3	602.5	5.3	0.80	1.4
OWL-ViT	355.8	352.7	388.4	2.8	1.20	0.4

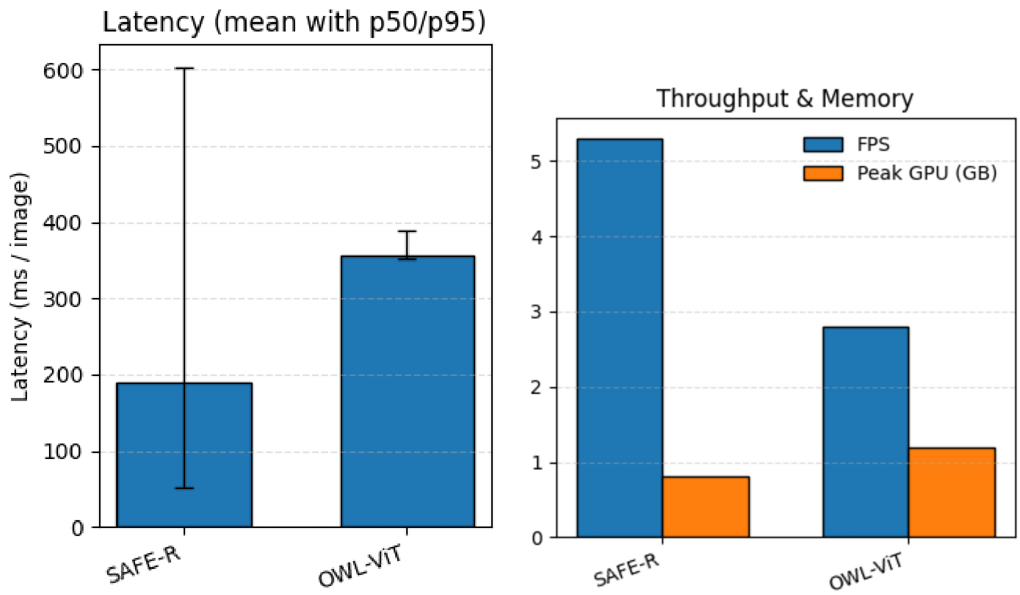


Figure 5.6: Computational Comparison of SAFE-R vs. OWL-ViT

Interpretation

The magnitude and consistency of the proposed SAFE-R pipeline across IoU thresholds and subclasses makes it the most positive result of the study relative to the established open-vocabulary baseline. The targeted objective for classification of hazardous objects was satisfied by the ideal division of labour in this pipeline, with the YOLOv8m detector used for geometry and the CLIP model used for semantics resulting in greater precision profile especially in cluttered scenes, as per the study's intention.

5.6 Discussions

Upon evaluating the findings, we can draw three key conclusions.
Firstly, with a robust and well-curated dataset, a stable training schedule and moderate augmentation, YOLOv8m detector remains ideal in terms of accuracy to latency balance for cluttered waste collections. Attempts to modify by adding scales (P2) or replacing the neck with attention blocks (BiFPN/CBAM) did not overcome the pressing issue of object-background confusion; rather, they worsened the results with false positives in some cases.
Secondly, the key to targeted subclass disambiguation was separating concerns: utilising detector's robust localisation and leveraging CLIP's contrastive semantics allowed the SAFE-R pipeline to outperform OWL-ViT at practical IoUs.
Thirdly, in the case of architectural modifications, improvements are largely class-dependent: P2 was observed to yield better results for metal. Aligned with our initial objective, combining YOLO's high recall and localization precision with CLIP's query-based semantic reasoning yielded desirable results in the fine-grained detection of hazardous subtypes, like knife, nail, syringe, etc.
However, there are limitations to acknowledge. Despite attempts to build a comprehensive and robust dataset, hazardous subclasses remain under-represented and

transparent materials like glass tend to drive residual FNs/FP spikes. We ascertained that future work should expand upon the existing dataset to increase hazardous annotation.

Chapter 6

Conclusion

6.1 Summary of Findings

In the first stage of our thesis, we tried various ways to adapt the YOLOv8m model for greater performance in waste classification, particularly for the detection of small, narrow objects that might pose a risk to workers in the field. We tried three variants: first, a P2 addition, then BiFPN with ECA augmentation replacement, and lastly, a ResCBAM addition in the neck. We compared the performance of these variants with the original YOLOv8m as baseline. Performance metrics revealed that YOLOv8m could not be outperformed by its variants, leading us to conclude that using the original YOLOv8m detector in our SAFE-R pipeline would yield best results.

Upon evaluating the results of SAFE-R (YOLOv8m + CLIP pipeline) in comparison to that of OWL-ViT on our custom dataset, it was seen that SAFE-R surpassed OWL-ViT at identifying all eight hazardous and non-hazardous waste objects we tested for. Quantitative assessment reiterated the statistical significance of the gains achieved by the SAFE-R pipeline. Not only that, the computation costs are also significantly lower for SAFE-R.

6.2 Contributions to the Field

As stated in our original objectives, we strove to advance the municipal waste management system of Bangladesh to curtail the adverse effects to the environment. The resultant SAFE-R pipeline would make significant strides in attempts to segregate waste at the source, improving waste management, especially in overcrowded cities like Dhaka. Its ideal equilibrium between speed and accuracy would make it suitable for deployment and the lightweight model would allow implementation using edge devices. Use cases include waste-collection from pharmacies, factories, supershops where a dangerous mix of hazardous and harmless unsegregated waste is discarded– SAFE-R can aid in identifying and safely separating hazardous objects from other waste classes, safeguarding waste-collectors from harm. Furthermore, city corporations can use SAFE-R at final waste disposal sites for separating recyclables from non-recyclables– even removal of organic wastes from the dumping grounds can greatly reduce water and soil contamination with the added benefit of allowing the organic waste to be used as compost. NGOs are already working

in Dhaka for improving waste collection and management, this technology can be leveraged to make them more sustainable and effective in their initiatives.
Moreover, our thesis delved into the possible modifications to YOLOv8m allowing for more minute object detection for greater accuracy in real-world conditions consisting of great variability, not to mention the risk of many small hazardous waste objects. However, the attempted variants did not perform as well as the original YOLOv8m model, demonstrating the robustness of the existing model. However, the takeaway from the attempts reveals the challenges faced in balancing the extraction of higher-resolution features and limiting the resultant noisy activations and lower overall mAP. It is possible to overcome this with more clean small-object supervision in a more representative dataset. Most significantly, the impressive gains of the YOLOv8m and CLIP integration opens the doors for wider applications of this vision-based, query-aware object detection pipeline.

6.3 Recommendations for Future Work

Our custom dataset forms a strong basis for a diverse, robust collection of waste data representative of real-world conditions, specifically of the Asian subcontinent, categorised with relevant classes suitable for use in research work. It can be further improved by more contributions, especially for the less represented classes, like sharp-objects with greater focus on small hazardous objects. Alongside this, our pipeline can be further improved by incorporation of semi-supervised labelling to reduce annotation cost. Further experimentation with other vision-language fusion models might be promising in terms of leveraging semantics for more accurate waste classification. Additionally, incorporating context-awareness into the detector model promises a post-hoc avenue for further improvement in accuracy of hazardous waste detection without harming recall.
For implementation, automation integration is the natural next step to translate the decisions of our pipeline to automated sorting of waste objects while maintaining real-time throughput and safety constraints. Thus, further steps include investigating suitable sensors, control strategies, actuators as well as evalutation protocols to really establish a scalable, automation-based foundation for municipal and industrial waste management solution in Bangladesh.

Bibliography

[1] M. Alamgir and A. Ahsan, "Municipal solid waste and recovery potential: Bangladesh perspective," *Iranian Journal of Environmental Health, Science and Engineering (ISSN: p-ISSN: 1735-1979) Vol 4 Num 2*, vol. 4, Jan. 2007.

[2] "Improving the informal recycling sector through segregation of waste in the household – the case of dhaka bangladesh," *Habitat International*, vol. 38, pp. 150–156, 2013, ISSN: 0197-3975. DOI: https://doi.org/10.1016/j.habitatint.2012.06.001 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0197397512000276

[3] *Wasteconcern.org*, https://wasteconcern.org/wp-content/uploads/2024/03/Waste-Data-Base_2014_Draft-Final.pdf, [Accessed 29-01-2025], 2014.

[4] M. Yang and G. Thung, "Classification of trash for recyclability status," Stanford University, Tech. Rep., 2016, Stanford University, CS229 Project Report. [Online]. Available: cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf

[5] S. Shams, J. Sahu, S. S. Rahman, and A. Ahsan, "Sustainable waste management policy in bangladesh for reduction of greenhouse gases," *Sustainable Cities and Society*, vol. 33, pp. 18–26, 2017, ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2017.05.008 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221067071730149X

[6] A. B. Syeda, A. Jadoon, and M. Nawaz Chaudhry, "Life cycle assessment modelling of greenhouse gas emissions from existing and proposed municipal solid waste management system of lahore, pakistan," *Sustainability*, vol. 9, no. 12, 2017, ISSN: 2071-1050. DOI: 10.3390/su9122242 [Online]. Available: https://www.mdpi.com/2071-1050/9/12/2242

[7] Y. Chu, C. Huang, X. Xie, B. Tan, S. Kamal, and X. Xiong, "Multilayer hybrid deep-learning method for waste classification and recycling," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–9, Nov. 2018. DOI: 10.1155/2018/5060857

[8] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, Feb. 2018. DOI: 10.1155/2018/7068349

[9] M. Toğaçar, B. Ergen, and Z. Cömert, "Waste classification using autoencoder network with integrated feature selection method in convolutional neural network models," *Measurement*, vol. 153, p. 107 459, Dec. 2019. DOI: 10.1016/j.measurement.2019.107459

[10] A. Vo, L. Son, M. Vo, and T. Le, "A novel framework for trash classification using deep transfer learning," *IEEE Access*, vol. 7, pp. 178 631–178 639, 2019. DOI: 10.1109/ACCESS.2019.2959033 [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2959033

[11] O. Alam and X. Qiao, "An in-depth review on municipal solid waste management, treatment and disposal in bangladesh," *Sustainable Cities and Society*, vol. 52, p. 101 775, 2020, ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2019.101775 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670719307061

[12] M. Maharjan and S. Lohani, "Municipal solid waste management in nepal: Opportunities and challenges," *Journal of the Institute of Engineering*, vol. 15, pp. 222–226, Oct. 2020. DOI: 10.3126/jie.v15i3.32185

[13] H. Panwar et al., "Aquavision: Automating the detection of waste in water bodies using deep transfer learning," *Case Studies in Chemical and Environmental Engineering*, vol. 2, p. 100 026, 2020. DOI: 10.1016/j.cscee.2020.100026 [Online]. Available: https://doi.org/10.1016/j.cscee.2020.100026

[14] P. F. Proença and P. Simões, *Taco: Trash annotations in context for litter detection*, 2020. arXiv: 2003.06975 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2003.06975

[15] A. Aishwarya, P. Wadhwa, O. Owais, and V. Vashisht, "A waste management technique to detect and separate non-biodegradable waste using machine learning and yolo algorithm," in *Confluence 2021: International Conference on Cloud Computing, Data Science & Engineering*, 2021, pp. 443–447. DOI: 10.1109/Confluence51648.2021.9377163 [Online]. Available: https://doi.org/10.1109/Confluence51648.2021.9377163

[16] F. Bianchi, G. Attanasio, R. Pisoni, S. Terragni, G. Sarti, and S. Lakshmi, *Contrastive language-image pre-training for the italian language*, 2021. arXiv: 2108.08688 `[cs.CL]`. [Online]. Available: https://arxiv.org/abs/2108.08688

[17] W.-L. Mao, W.-C. Chen, C.-T. Wang, and Y.-H. Lin, "Recycling waste classification using optimized convolutional neural network," *Resources, Conservation and Recycling*, vol. 164, p. 105 132, 2021. DOI: 10.1016/j.resconrec.2020.105132 [Online]. Available: https://doi.org/10.1016/j.resconrec.2020.105132

[18] F. Parvin and S. M. Tareq, "Impact of landfill leachate contamination on surface and groundwater of bangladesh: A systematic review and possible public health risks assessment," *Applied Water Science*, vol. 11, p. 100, 2021. DOI: 10.1007/s13201-021-01431-3

[19] S. A. Urme et al., "Dhaka landfill waste practices: Addressing urban pollution and health hazards," *Buildings and Cities*, vol. 2, no. 1, pp. 700–716, 2021. DOI: 10.5334/bc.108

[20] C. Wang, J. Qin, C. Qu, X. Ran, C. Liu, and B. Chen, "A smart municipal waste management system based on deep-learning and internet of things," *Waste Management*, vol. 135, pp. 20–29, 2021. DOI: 10.1016/j.wasman.2021.08.028 [Online]. Available: https://doi.org/10.1016/j.wasman.2021.08.028

[21] P. J. D. Gamaralalage, S. K. Ghosh, and K. Onogawa, "Source separation in municipal solid waste management: Practical means to its success in asian cities," *Waste Management & Research*, vol. 40, no. 3, pp. 360–370, 2022, PMID: 34608835. DOI: 10.1177/0734242X211049606 eprint: https://doi.org/10.1177/0734242X211049606. [Online]. Available: https://doi.org/10.1177/0734242X211049606

[22] W. Lu, J. Chen, and F. Xue, "Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach," *Resources, Conservation & Recycling*, vol. 178, p. 106 022, 2022. DOI: 10.1016/j.resconrec.2021.106022 [Online]. Available: https://doi.org/10.1016/j.resconrec.2021.106022

[23] S. Majchrowska et al., "Deep learning-based waste detection in natural and urban environments," *Waste Management*, vol. 138, pp. 274–284, 2022. DOI: 10.1016/j.wasman.2021.12.001 [Online]. Available: https://doi.org/10.1016/j.wasman.2021.12.001

[24] M. Malik et al., "Waste classification for sustainable development using image recognition with deep learning neural network models," *Sustainability*, vol. 14, no. 12, p. 7222, 2022. DOI: 10.3390/su14127222 [Online]. Available: https://doi.org/10.3390/su14127222

[25] N. Nnamoko, J. Barrowclough, and J. Procter, "Solid waste image classification using deep convolutional neural networks," *Infrastructures*, vol. 7, no. 4, p. 47, 2022. DOI: 10.3390/infrastructures7040047 [Online]. Available: https://doi.org/10.3390/infrastructures7040047

[26] M. I. B. Ahmed et al., "Deep learning approach to recyclable products classification: Towards sustainable waste management," *Sustainability*, vol. 15, no. 14, p. 11 138, 2023. DOI: 10.3390/su151411138 [Online]. Available: https://doi.org/10.3390/su151411138

[27] N. Pičuljan and Ž. Car, "Machine learning-based label quality assurance for object detection projects in requirements engineering," *Applied Sciences*, vol. 13, no. 10, 2023, ISSN: 2076-3417. [Online]. Available: https://www.mdpi.com/2076-3417/13/10/6234

[28] K. Sarquah et al., "Characterization of municipal solid waste and assessment of its potential for refuse-derived fuel (rdf) valorization," *Energies*, vol. 16, no. 1, p. 200, 2023. DOI: 10.3390/en16010200

[29] S. M. Yeaminul Islam and M. G. Rabiul Alam, "Computer vision-based waste detection and classification for garbage management and recycling," in *The Fourth Industrial Revolution and Beyond*, ser. Lecture Notes in Electrical Engineering, M. S. H. et al., Ed., vol. 980, Springer Nature Singapore Pte Ltd, 2023, pp. 389–404. DOI: 10.1007/978-981-19-8032-9_28 [Online]. Available: https://doi.org/10.1007/978-981-19-8032-9_28

[30] S. Akther, J. Evans, and N. Millington, "Why do cities in the global south pursue waste incineration rather than source segregation and recycling? insights from bangladesh," *Frontiers in Sustainability*, vol. 5, 2024, ISSN: 2673-4524. DOI: 10.3389/frsus.2024.1496075 [Online]. Available: https://www.frontiersin.org/journals/sustainability/articles/10.3389/frsus.2024.1496075

[31] A. Elhanashi, P. Dini, S. Saponara, and Z. Qinghe, "Telestroke: Real-time stroke detection with federated learning and yolov8 on edge devices," *Journal of Real-Time Image Processing*, vol. 21, Jun. 2024. DOI: 10.1007/s11554-024-01500-1

[32] M. M. Hossen et al., "A reliable and robust deep learning model for effective recyclable waste classification," *IEEE Access*, vol. 12, p. 3 354 774, 2024. DOI: 10.1109/ACCESS.2024.3354774 [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3354774

[33] D. Kumar et al., "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.

[34] Z. Liu, R. Zhang, H. Zhong, and Y. Sun, "Yolov8 for fire and smoke recognition algorithm integrated with the convolutional block attention module," *Open Journal of Applied Sciences*, vol. 14, pp. 159–170, Jan. 2024. DOI: 10.4236/ojapps.2024.141012

[35] P. Wang et al., "Leaf segmentation using modified yolov8-seg models," *Life*, vol. 14, p. 780, Jun. 2024. DOI: 10.3390/life14060780

[36] Z. Fang, Y. Wu, Y. Wang, and J. Zhong, "Infusion monitoring and real-time prediction based on yolov8 with p2 enhancement," pp. 46–53, 2025. DOI: 10.1109/ICACI65340.2025.11096252

[37] *Simple Open-Vocabulary Object Detection with Vision Transformers — doi.org*, https://doi.org/10.48550/arXiv.2205.06230, [Accessed 09-10-2025].

[38] *Waste-Detection-and-Classification-Labeled-Dataset*. [Online]. Available: https://www.kaggle.com/datasets/yeaminul/waste-detection-and-classification-labeled-dataset/data