WasteRefine: Boundary-Aware Semantic Segmentation of Waste Materials Using a DINOv2 Backbone with Multi-Scale Feature Fusion Decoder

by

Talha Islam Khan
22101255
Trisha Das
22101422
Md. Ahnaf Iqbal
22101706
Farhan Tawseef
22201328

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

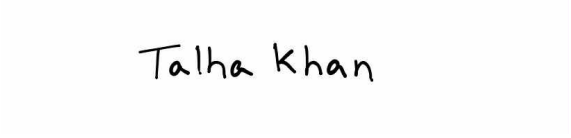
Department of Computer Science and Engineering
Brac University
April 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/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:



Talha Islam Khan

22101255

Trisha Das

22101422

Md. Ahnaf Iqbal

22101706

Farhan Tawseef

22201328

Approval

The thesis titled "WasteRefine: Boundary-Aware Semantic Segmentation of Waste Materials Using a DINOv2 Backbone with Multi-Scale Feature Fusion Decoder " submitted by

1. Talha Islam Khan (22101255)
2. Trisha Das (22101422)
3. Md. Ahnaf Iqbal (22101706)
4. Farhan Tawseef (22201328)

of Spring, 2026 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 12, 2026.

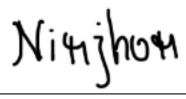
Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



NIRJHOR DATTA

Lecturer
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)

Dr. Sadia Hamid Kazi

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

Abstract

The rapid increase in world waste production needs smart, data-driven frameworks for efficient material identification and sustainable resource management. Intelligent recycling systems and waste materials spontaneous segmentation often lack behind due to scarcity of proper annotated datasets, visual ambiguities and severe class imbalancement of rare objects. The research aims to propose WasteRefine, utilizing DINOv2 Vision Transformer backbone with boundary aware semantic segmentation and multi scale feature fusion decoder for waste materials. To capture and accumulate the global context, an advanced dense predictive transformer is used consisting top-down fusion of features, Pyramid Pooling Module, Squeeze and Excitation channel attention and boundary composition component, for the proper identification of cluttered, deformed and visually ambiguous waste objects. The paper also introduces WasteRefine dataset consisting of 2,213 annotated images across four different categories: paper, soft plastic, rigid plastic and metal, marking it as the first waste semantic segmentation dataset from Bangladesh which contains visuals across various regions and annotated precisely. The proposed framework is rigorously evaluated on three different dataset WasteRefine, ZeroWaste-F and SpectralWaste (RGB) and assessed across notable published baselines. The ViT-B achieved 96.64 $\pm$ 0.16% mIoU on WasteRefine dataset, 61.94 $\pm$ 0.84% mIoU on extremely class imbalanced and deformed ZeroWaste-F dataset and 70.73 $\pm$ 0.10% FG mIoU on SpectralWaste beating all the published reports. Competitive results of the ViT-S variant with only 25.16M parameters demonstrated efficient parameter count without severe performance degradation.

Keywords: Semantic Segmentation, Self-Supervised Learning, Convolutional Neural Networks, Environmental Sustainability, Waste Classification, Vision Transformer, DINOv2, Boundary supervision.

Acknowledgement

At first we would like to express our appreciation and gratitude towards the Almighty, the main source of strength, courage and wisdom for guiding us throughout the journey. We express our heartfelt appreciation to our Supervisor, Dr. Golam Rabiul Alam, PhD for his commitment, guidance, unwavering support, dedication and constant encouragement. We also extend our sincere gratitude and appreciation towards our Co-Supervisor Nirjhor Datta, for his expert advice, continuous assistance and encouragement for betterment. With their support and assistance this research has been successfully conducted. Moreover, we acknowledge all the individuals and platforms who contributed directly or indirectly.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The rapid increase in global waste has become a major threat to the environment and public health which calls for immediate recycle and disposal management strategies. If the current trend of municipal solid waste (MSW) generation keeps continuing, there is an expectation of an increase from 2.3 tonnes in 2023 to 3.8 billion tonnes by 2050 [24]. With a growing population and rapid urbanization, there has been a shift in consumption patterns leading to the amount of waste generated per capita increase significantly. The traditional disposal method of landfilling waste has become harder to maintain due to the scarcity of land and environmental deterioration. For instance, landfills contribute to a large amount of methane emissions in the atmosphere which plays a major role in climate change. This human-caused greenhouse gas contributes almost 20% of the global methane gas emission [5]. Thus, poor waste management causes extensive ecological damage to both land and sea risking biodiversity, human-welfare and climate stability. To minimize these effects and transition to a global circular economy, the sustainability and recyclability of materials are very crucial in the long run. To address these challenges, efficient waste recycling and sorting methods have become a key strategy for reducing dependency on fossil fuels, tackling landfill scarcity and reducing greenhouse gas emissions for environmental sustainability. The ever-growing WTE market was valued at approximately USD 42.7 billion in 2025 and is expected to reach between USD 70 billion and USD 80 billion by the early 2030s [30]. This growth rate of approximately 9% to 11% is not accidental, rather a result of energy security, government policies, increased urbanization and technological advancements. To overcome the dependency on landfill waste dumping and manual waste sorting a smart and efficient automated waste sorting procedure is vital for environmental sustainability. Despite the rise of recycling facilities in many industrial sectors, the complicated nature of the various waste such as: occlusion, translucent and deformed nature make the facilities open to numerous challenges such as: proper classification and segmentation dealing with real world waste streams. Context aware information from different classes at a balanced rate is often lacking in assessing the dataset in this domain, which hinders the performance of the predicted sorting and leads to performance degradation. As a result, reliability, efficiency and robustness should be ensured while establishing the modern waste facilities.

While addressing the above mentioned limitations, the research proposes a unified framework for semantic segmentation utilizing a foundation model DINOv2 processed at the backbone. It also proposes a custom built decoder which addresses the drawbacks like clutteredness, deformness and translucent nature of the waste through dense prediction, boundary aware sharpening. For benchmarking of the system the model is tested on three different dataset ZeroWaste-f, SpectralWaste and our newly created WasteRefine dataset of 2,213 annotated images. The validation process across different dataset properly ensures robust handling of different waste categories to provide a scalable solution for the waste management system.

1.2 Rational of the Study or Motivation

Waste sorting leading to recycling reusability facilitates alternative approaches rather than landfilling. Despite providing a noble alternative approach, the environmental sustainability and systems performance are obstructed due to management of diversified municipal solid waste. Primitive approach heavily relies on labour for manual waste separation procedures which results in hindrance of precision and sorting speed. Inefficient manual sorting and waste of resources in waste management facilities causes instability resulting in operational stability disruption, increment of emissions and reduction of energy output. A metamorphic solution is offered due to the latest advancement of machine learning framework in visual data oriented waste sorting. Early sorting of waste is tremendously boosted due to usage of machine learning framework, for example accuracy overtaking 95% [31]. The motive of the research is to put forward a comprehensive and combined approach which not only classifies waste type with excellent precision but also assigns pixel level classes known as semantic segmentation for accurate localization and sorting. Efficiency enhancement, environmental outcomes, stability and operational adjustments are the changes waste management plants can expect using the predictive outcome. The paper seeks for a new data driven methodology that is responsible for optimized resource recovery through integration of automated waste type classification and predictive analysis, proposing a newly created dataset capable of semantic segmentation by addressing the class imbalancement, precise annotation of waste, various categories of waste types selection etc for better learning of the framework which the domain lack significantly. Finally, the research extensively supports the transition to a stable economy and contributing to UN Sustainable Development Goals. SDGs 7, 9, 11, 12 and limited offerings to SDGs 3, 6 and 13 which are the major SDGs this study aligns to.

1.3 Problem Statement

In order to address the evolving global waste issue and the lack of sustainable waste management, classification of waste and efficient pixel level marking for automated sorting and localization is very crucial. On the other hand, because of the absence of unified intelligent systems, operational shortcomings and computational cost, the process's efficiency is severely limited. The immediate requirement to reduce landfills, mitigate emissions and improve efficiency from wastes shape the significance of tackling this issue. A major problem that is present in the cur-

rent waste sorting methods is the disconnected structure of different methodologies, where the deformed, translucent and cluttered waste often appear misclassified, boundary unaware and failure to identify smaller objects. Additionally there is lack of proper dataset which ensures balanced objects across different classes, for example zerowaste-f dataset have metal 50 times lesser than the dominant class cardboard [7]. As a result the models cannot learn the textures and structures of every class properly. Lastly the traditional supervised model is traced with uncertainty in terms of results due to its local context awareness. This means, in any research until now, the mentioned problems have not been tackled within a single, unified framework and dataset. This restricts the full potential of the waste management plants to recover waste material and ecological sustainability as real-time, data-driven decision making processes are inhibited because of the dataset.

Subsequently, efficient models that undergo global contextual understanding and pattern recognition using transformers are yet to be deployed in this domain that can address these complicated problems. So the main challenge that this study aims to overcome is to develop a unified, integrated framework that can efficiently predict waste objects semantic results from an image based dataset that will enable for more efficient and ecologically sustainable waste management operations.

1.4 Objective

To overcome the underlying huge crisis of global waste and systemic failure in recycling facilities the research extensively focused on enhancing automatic robustness of sorting systems and accuracy. Paradigm shift from traditional local context aware supervision approach to global context aware self supervised framework, the research focused on underlying problems such as: information or data lackings in deformed waste systems and object centric precise prediction. The research objectives are as follows:

- **Creation of new dataset named WasteRefine:** Designing a domain specific dataset entitled WasteRefine with new waste images and proper annotation for semantic segmentation is the primary goal the research focused on. 2213 annotated images were captured and precisely annotated keeping five different waste types such as: background, rigid plastic, soft plastic, metal and paper/cardboard. Validation of the model against two other dataset like ZeroWaste-f and SpectralWaste along with WasteRefine provides a credible benchmark.
- **Develop a Segmentation Framework:** Development of precise semantic segmentation model for correct prediction utilizing a self distillation based foundation model DINOv2 backbone is termed as one of the primary work of the research. The framework is designed for semantic segmentation based task performance which is not only capable of waste classifications but also pixel level delineation. For proper identification of overlapping, deformed and occlusion based objects in the industrial setup the framework is essential.
- **Design a Decoder with Boundary Awareness and Multi Scale Feature Fusion:** To ensure correct pixel level object identification in an extremely

cluttered environment the custom decoder is designed which incorporates advanced dense prediction and boundary enhancement module for rectification of output features from vision transformer backbone. By utilizing global semantic context and high frequency edges the component unleashes its true potential for resolving class confusion.

- **Enhance Generalization and Model Robustness:** Robustness across various waste objects in different dataset is ensured through models architectural component selection and optimization of hyperparameters. The validation of the research across three different dataset properties advocates the models robust performance across variable dataset.
- **Mitigate Scarcity of Data through Advanced Feature Extraction:** Utilization of self supervised learning models signal on limited data is the solution we employed to tackle the problem of data scarcity. By leveraging the benefits of DINOv2 encoder such as: frozen backbone features and patch level objectives, precise accuracy of segmentation is ensured with little data annotated images, reducing manual annotation costing and labour.

1.5 Methodology in Brief

The methodology proposed in the paper is a dedicated pipeline consisting of eight stages and divided into two segments Data and Model Pipeline. A detailed methodology is illustrated through Figure 1.1.

The data pipeline consists of initial 4 stages which includes data collection, preprocessing, augmentation and Sampling and Splitting. Creation of newly captured and annotated dataset WasteRefine and collection of other two datasets for benchmarking ZeroWaste-F and SpectralWaste is done in data collection. Validation of masks, filtration of duplicate images to eliminate the similar image, image resizing to 518×518 or 256×256 pixels according to model compatibility and dataset image quality and normalization to RGB pallets are the notable preprocessing steps employed. Augmentation is also a crucial step for training protocol and the technique, process used for augmentation requires rigorous hyper parameter tuning for each model and dataset. The WasteRefine and ZeroWaste-F underwent similar augmentation adjustments such as: vertical and horizontal flips, CoarseDropout, rotation etc. On the other hand SpectralWaste dataset with very thin and rare class presence underwent more sophisticated augmentation changes. In order to address the severe class imbalance of two collected dataset the paper utilized a weighted random sampler with pixel awareness and standard splitting of 70/15/15 is utilized for WasteRefine dataset and for other two dataset the specific splitted images is used for benchmarking.

The model pipeline consists of final 4 stages which includes model architecture, boundary branch, training strategy and evaluation and results. The model architecture consists of DINOv2 Vision Transformer with two variants Base 90.61M and Small 25.16M at the encoder. The decoder extracts features from intermediate layers through an Advanced dense predictive Transformer, which adds Pyramid Pooling Module, Smooth Convolution layers, Squeeze and Excitation attention with the top

down fusion of features. A dedicated boundary composition component utilizing binary cross entropy, for the proper identification of cluttered, deformed and visually ambiguous waste objects. The training strategies utilized are divided into two steps, warmup with frozen backbone for extraction of fine features to stabilize the neck and full fine tuning for precise training with CosineAnnealing, Exponential Moving Average and a combination of dice,focal, auxiliary and edge loss. Finally the last stage is dedicated for rigorous evaluation with reproducible averaged random seeds, the results showcase mIoU, FG mIoU, Dice, Precision and Pixel Accuracy across all benchmarked datasets.

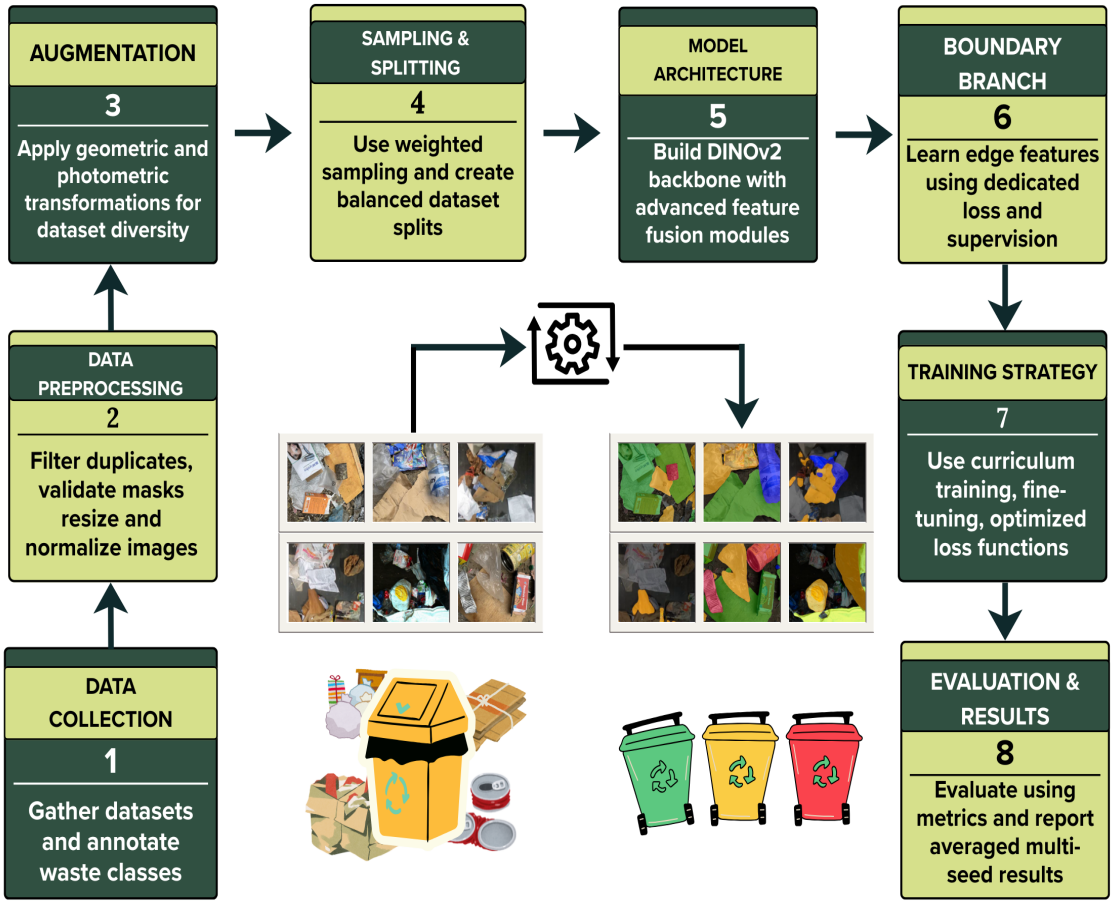


Figure 1.1: Methodology pipeline of the WasteRefine framework.

1.6 Scopes and Challenges

As the previous underlying results and quality of images on semantic segmentation waste domains dataset is limited, the research scope is primarily focused on improving the results on the existing work and proposing a new balanced dataset with proper class alignment. The research is confined to contributing in complicated cluttered environments which can be used further on other domains too. The number of problems traced is superior to the scopes. The existing qualitative dataset limitation can be identified as the main challenge, the class heavy imbalancement is another challenge while working on the existing dataset such as: Metal is 50 times lesser than dominant object cardboard in ZeroWaste-f dataset [7], the deformed and reflective nature of the waste often leads to confusion among other classes. Difficulty tracing the boundary of translucent soft plastic is also another problem [26]. Adaptability and general feature identification gap between general learned feature and deformed object shows variability in results too [11]. Custom decoder designing along with self distillation based global context awareness backbone is utilized by the research to handle the overwhelming challenges.

Chapter 2

Literature Review

2.1 Preliminaries

As the waste volume increases rapidly, automated waste management facilities have become extremely crucial for the world to deal with increasing waste materials and the urge for sustainable energy. Automated precise waste sorting with manual labor intervention is the next paradigm shift the industry awaits. Due to the information deficiency, deformed, cluttered and translucent objects in the waste field the modern recycling system efficiency, correct identification and object classification is hindered in the field of Computer vision. Transient shift from traditional supervised learning approach to DINOv2 foundation model of self supervised learning framework for mitigation of the above problems is the intent of the research. Sticking to semantic segmentation as our core approach in this research paper is to identify pixel level differences to classify distinct objects such as: plastic, metal, paper etc for correct sorting. In order to assure the precise segmentation between classes, the decoder design focused on refining boundaries of objects and semantic cues amplification for separation of objects from another in cluttered and complex environments.

2.1.1 Key Concepts and Theories

The fundamental concept beneath the research is to overcome the deficit that is present in the field of complicated cluttered, deformed waste objects and optimization of the systems. Extraction of spatial visual features without any manual intervention and rigorous annotation is done through utilization of self supervised backbone. The implementation of DINOv2 for extraction of primary features, detection of patch level information, partial and semantic information is primarily used for correct identification of waste material in extreme cluttered areas. Some underlying classes are present in the waste domain due to its limitation trace such as: metal and due to its underlying material learning it often misclassifies with other objects like plastic. The core focus of the research was to mitigate the information bridge error margin by taking extensive measures beyond dense prediction. Identification of high frequency edge details through utilization of boundary enhancements and sharpening techniques as well as enhancement of feature analysis is integrated along encoder decoder design. Unification of the above methods with the semantic segmentation framework for generation of correct mask generation can increase efficiency, correctness in autonomous sorting systems.

2.1.2 Models and Methodology

The core foundation of the research is methodically multifaceted which ought to identify specific gaps and solve using state of the art mechanisms. Initial phase focuses on categorization and segmentation for feature extraction with correctness of waste which revolves around transfer learning techniques, rooting self distillation technique backbone DINOv2. The architecture is focused on DINOv2, a self supervised approach foundation model that is capable of extracting visual representations and robust texture without any labeled and supervised training. The foundation model, pretrained on a huge curated dataset, gathers spatial and semantic understanding from global context with two different types of crops that are more superior than local context driven supervised CNN based techniques. The global context aware features are then passed through a modulated custom decoder which consists of capturing fine grained details, spatial information through dense predictive algorithms and a boundary enhanced module for rectification of output to segmentation heads.

The core motive behind the integration was to effectively fuse local edges with global aware context to reduce the classification error as well as semantic value in cluttered and deformed environments. Involvement of feature sharpening and refinement block for extreme scale variation and deformation tracing in the waste recovery facilities. The research has a further possibility to extend the methodology for better optimization of overlapping and thin materials through using refinements techniques like HQ-SAM by utilization of output tokens through semantic embedding and high resolution features in the early stage. The proposed methodology is benchmarked with three different dataset on deformed and cluttered waste fields such as: ZeroWaste-f, SpectralWaste and our own created dataset. A robust and scalable solution for pixel level analysis, boundary aware cues and global context identification moves beyond the traditional methods in the waste domain, solving major problems like misclassification, low texture identification, boundary badly marked and efficiency related to inference.

2.2 Review of Existing Research

Ali et al. [26] proposed a renowned framework of semantic segmentation named COSNet (Cluttered Objects semantic Segmentation Network) for classifying waste in difficult and cluttered scenes. Plastic having translucent properties, unclear boundaries of different materials due to deformed nature makes the performance less effective in real world sorting facilities are the notable research gap the authors described. Feature Sharpening Block (FSB) and Boundary Enhancement Module (BEM) are the two prominent additions to the architecture along with encoder and decoder systems. The suppression of the background noise through highlighting high frequency details is done through Boundary Enhancement Module and contextual features at multi-scale are captured through Feature Sharpening Block to enrich boundary details by utilizing various convolutions. These are among the primary novelty factors because it distinguishes between deformed waste materials from background by focusing on the surface cues. The paper utilized three different dataset for benchmarking such as: ZeroWaste-f, SpectralWaste and ADE20K. It

achieved the SOTA results on ZeroWaste-f till its publication with mIoU 56.67% a remarkable 1.8% gain over Fanet and it achieved mIoU 69.96% on SpectralWaste a remarkable 2.1% gain till its publication. It outperformed the transformer based solution to the problem and is often regarded as the best model under consideration due to requiring less annotated data and capability of identifying translucent objects fine details. More dependencies on the annotated data for the supervised approach is also addressed as a limitation of the paper which the author tends to overcome in the future work along with extending the framework for segmentation of transparent waste in real-world environments.

Another interesting paper by Ali et al. [16] demonstrated a network specifically designed for backbone for enhancement of segmentation in cluttered, complex and deformed scenarios named FANet (Feature Amplification Network). The failure of CNN and vision transformer to identify context in semantic level and fine details for extremely difficult and variance environments which consists of translucent, occluded and deformed materials are the research gap identified by the authors. A prominent block named Adaptive Feature Enhancement (AFE) which functions to capture multi staged features are the novel contributions. The above mentioned block consists of two components parallely functional with Spatial Context Module (SCM) through utilization of 7 x 7 kernel for increasing the field of receptivity along with Feature Refinement Module (FRM). Feature Refinement Module is well equipped for enhancing contrast and sharpening details through adaptation of high frequency object boundaries features and blob region in low frequency areas for the refinement of partial features. The extremely cluttered and deformed dataset termed as industrial grade named ZeroWaste-f was utilized to evaluate the effectiveness of the model. Notable gain of almost 2% in terms of mIoU with 55.89% was noted by the paper with 91.41% pixel accuracy, which was the best performing results on this dataset when published. Reduction of parameters, higher efficiency in terms of results, tremendous robustness in real-world cluttered environments are the notable architectural wins for the model when comparing to attention based architectures. Detection of precise boundaries for cluttered scenes, misclassification of highly translucent objects are the mentioned limitations by the paper. Enchantment of models for better prediction and results are the work that tend to work in future.

The article by Bashkirova et al. [7] presented ZeroWaste, the very first industrial graded dataset for opening the scope of research in extreme cluttered situations and deformed object segmentation residing in recycling systems. Due to lack of sorting efficiency less than 35% of recyclable waste gets processed, an exceptional research gap dissected by the article. To mitigate the margin of underpressed waste and insufficient high quality dataset for realworld deformed, occluded and translucent waste ZeroWaste dataset was primarily proposed. Variations of data collection from Materials Recovery Facility was developed to facilitate semi, weakly and fully supervised learning. The fully supervised learning dataset named ZeroWaste-f contains 4503 heavy cluttered waste images of five classes including background such as: metal, cardboard, rigid plastic, soft plastic and background. Many state of the art models were selected for waste object detection and semantic segmentation to establish baseline such as: TridentNet, RetinaNet and Mask R-CNN for object detection and DeepLabV3+ for segmentation. The results of the experiments showed

interesting outcomes such as: mean average precision (AP) of 24.2 with TridentNet for object detection and testing mean intersection of union mIoU of 52.13 with DeepLabV3+ was tracked as the best model. The detailed analysis revealed many crucial challenges of the dataset such as: metal is termed as extremely rare, almost 50 times lesser than cardboard. Misclassification of plastic, high cost of experienced annotation hands and state of the model underperformed in occluded and deformed nature are the notable findings from the research paper. The future work addresses the mitigation class imbalance, diversification of dataset through real and synthetic measures and efficient model proposal for proper management of deformed and cluttered waste.

The first dataset containing multimodal images from a waste sorting plant incorporating synchronized hyperspectral imaging HSI and RGB images was proposed by Casao et al. [17] The research paper identifies a crucial limitation bound to RGB based images where it lacks to differentiate visual features that are contained in cluttered, deformed and translucent waste facilities. A dual camera was set above the waste facility to capture high quality RGB images along with 224 spectral bands followed by label transfer algorithms. Elimination of the urge of complicated partial calibration from annotated RGB to HSI images was handled through this technique. The dataset contained 852 images of each format with 7 different classes including background such as: film, basket, card, video tape, filament and bag. For establishment of benchmarking MiniNet-v2, SegFormer-B0 and CMX-B0 were used focusing on higher inference time in real world implementations. CMX-B0 was declared the best model due to its hybrid approach transformer decoder, capturing and rectification of noisy signals from sensors. Unimodal results mean intersection over union mIoU 48.4 whereas the HSI performed at 54.3, showcased a tremendous improvement on results because of HSI. Moreover the hybrid fusion mechanism reached the SOTA performance of 58.2 mIoU. The integration of HSI images is computational expensive despite the high quality results from the hybrid approach. The future work highlights the mitigation of low performance of some classes such as: card and filament along with balancing the classes properly. It also highlights the potential interest of using unsupervised learning techniques due to having almost seven thousand unlabelled multimodal images.

A renowned self supervised learning method named DINO, applying self distillation method to train the encoder without any usage of labels incorporating Vision Transformers (ViT) was introduced in the paper by Caron et al. [4] The power of ViT when trained under limited supervision is still not as powerful as standard CNN methods compared to trained under full supervision with a bigger labeled dataset is addressed as the research gap. A groundbreaking student teacher concept was developed as the backbone of the network with the usage of different parameters. Exponential moving average was utilized to update students' weight when it sequentially matches the output of the teacher. Without any form of ground truth or mask of the semantic segmentation object the model learns the explicit feature through a new self supervised learning framework, which can be found as the primary novelty. A top-1 accuracy of 78.3% was achieved on ImageNet through DINO training on ViT base through frozen features with a simple k-NN classifier outperforms previous best self supervised learning techniques. Due to highly variable global context, spatial feature extraction, image retrieval and fine grained object processing, DINO

is a widely used model for semantic segmentation. The future work proposed by the paper indicated scaling the signals of the self supervision to larger dataset for extracting more complicated visual representation.

The extension of the research of a self distillation method DINO is proposed in the research by Oquab et al. [13] named DINOv2. Trained on an extremely big and curated 142 million images dataset, DINOv2 is a family of the self distillation foundation models. The revolutionary nature of natural language processing traced through big scale pretraining is the research gap correctly identified in the computer vision field. Due to lack of well equipped backbone, adapting to large scale pretraining, spatial feature identification among diverse tasks, depth perception and semantic segmentation without any fine-tuning domain related is hard. The framework is upscaled using a vision Transformer architecture through a self supervised approach combining the objective of patch level iBOT and image level DINO. The initial novelty lies under the automated sorting of the data or image from uncurated web sources to eliminate low quality images and images with distinct properties. The novelty also lies due to the stable training techniques such as: stochastic depth and FlashAttention used to train 1.1 billion parameters based models like ViT-g. While benchmarking on ImageNet-1k dataset it performed 86.5% accuracy and depth estimation on NYUv2 outperforming all the previous SOTA self supervised models and weakly supervised models like OpenCLIP with frozen features. Due to its vast capabilities of semantic features and partial features, object boundaries identification and texture mapping is often correctly identified. Data filtration in the pretraining phase and high computational pre training are the notable limitations of the paper. Visual reasoning enhancement through multimodal data integration and scalable model size ensurement are the future work proposed by the model.

Chen et al. [3] demonstrated a simple yet new framework for self supervised contrastive learning named SimCLR which uses unlabelled data for pretraining. Reduction of feature quality compared to supervised training, and complexity of implementation traced from previous methods of self supervised learning was overcome through this paper. Through the usage of latent space, maximization of the agreement based on contrastive loss between various augmented aspects of the same data, the methodology evolves. Empowerment of the system is through using larger batch sizes, projection head of non linear distribution and different view of augmented reality. Using Resnet backbone along with non linear projection head tremendously helps the quality of the features is the primary novelty from this model. The results benefit from longer training periods and utilization of large models in comparison with supervised learning. SimCLR with frozen feature and linear classification achieved top-1 accuracy of 76.5% on ImageNet matching with supervised learning model ResNet-50 during the time of publication. The paper also gives a broader knowledge about the augmentation technique and the impact of the data augmentation on the network to differentiate between objects precisely. Heavy reliance on negative samples and extreme bug batch size for example: 8192 along with extensive training procedures are termed as the limitation of the paper. Exploration of more suitable augmentation techniques and domain specific adoptions are the future work proposed by the authors.

The article by Ke et al. [11] Proposed High-Quality Segment Anything Model HQ-

SAM as an extension to SAM which produces masks with high precision. Despite SAM's powerful foundation models prominent features such as zero shot but it often fails to distinguish objects with complicated components, structures with cluttered boundaries, thus producing blobby or coarse mask in high quality environments. Introduction of global local feature block of fusion and high quality yet lightweight token for output while keeping the SAM's original parameter frozen is the methodology innovation the paper came with. Through utilization of high resolution features in early stage and semantic features for the refinement of mask quality without the urge to retrain the massive backbone is termed as the primary novel work. Benchmarking on the HQ-Seg-44K dataset consisting of 44k accurate masks demonstrated SOTA result by outperforming SAM in terms of spatial feature preservation and boundary mapping across variation of tasks. Reliance on the points or boxes which is given as the prompt by users or autonomous systems is termed as the limitation of the model. The extension of the high quality refinement block for the segmentation of video and cloud tasks based on 3D point is promoted for the future work by the authors.

The paper by Jaspers et al. [28] aims to make a contribution to the theme of surgical computer vision by offering a new self-supervised learning (SSL) foundation model named SurgeNetXL. This is to address the significant research gap in terms of small-scale training dataset availability and a lack of standard benchmarks within the medical domain. The main contribution of the paper lies in pre-training the model on the largest available dataset within the domain which consists of more than 4.7 million video frames from 23 distinct procedures by utilizing the DINO distillation technique. From a methodological perspective, a variety of models were evaluated in terms of semantic segmentation, phase recognition and Critical View of Safety (CVS) classification. This reveals that the CAFormer hybrid model was deemed more appropriate owing to the smaller size of parameters which enabled a balance between local and global contextual learning, as well as the need for increased speed within such applications. This paper revealed that SurgeNetXL performs best in terms of accuracy by outperforming all existing surgical foundation models by 2.4%, 9.0% and 12.6% on each of the tasks respectively. Also outperforming ImageNet-based approaches by 14.4%, 4.0%, and 1.6% on each of the tasks respectively. However, limitations of the research include the use of unoptimized, standardized hyperparameters to isolate pre-training effects, as well as a strict constraint to frame-based rather than video-based analysis due to the massive computational requirements, which could be overcome in future research by using adaptive training approaches and incorporating other video-based SSL approaches, such as V-JEPA and DINOv2, to capture the significant temporal dynamics of surgical procedures.

This paper by Akilan et al. [25] seeks to fill the research gap by discussing the theme of self-supervised learning (SSL) for image segmentation. The research focuses on the application of SSL for image segmentation and the limitations that have characterized the field so far. The novelty of the research is the categorization of more than 150 recent publications on the theme into predictive, generative, and contrastive pretext methodologies, in addition to a curated review of the relevant datasets for image segmentation. The methodology of the research entails the evaluation of different self-supervised learning models and the conclusion that the best models for image segmentation are the dense contrastive models and the

multi-scale encoder-decoder models due to their ability to effectively combine the global context with the pixel-level representation that is strictly necessary for the task of image segmentation. The results show that the accuracy of the SSL image segmentation models is extremely competitive and comparable to that of the supervised models. The multi-scale encoder-decoder model with Pixel-wise Contrastive Loss (PCL) achieved an impressive accuracy of 88.1% on the Kvasir-SEG dataset, whereas the best SOTA models achieved an accuracy of only 91.3% on urban scene datasets such as SYNTHIA. The limitations of the research include the high computational and memory complexity of the models for high-resolution images and the difficulty of effectively handling complex high-level relationships without the use of labels. The future focus of the research is to incorporate the use of semantic priors and improve the few-shot and zero-shot learning ability of the models and domain adaptation for handling shifts in datasets and the need for real-time and interactive image segmentation for applications such as autonomous driving and robotics.

The theme of automated medical image segmentation is explored in this research by Liu et al. [20], which attempts to bridge the gap of computationally expensive transformer-based vision foundation models. This gap was caused by a combination of high quadratic complexity and a large number of parameters. A novel architecture was developed, Swin-UMamba†, which was a first in adapting a Mamba-based vision foundation model to image segmentation. In terms of methodology, a U-shaped structure was used, consisting of Visual State Space (VSS) blocks with 2D-selective-scan for linear complexity long range dependency modeling. A self-supervised adaptation scheme was also implemented to bridge the domain gap from natural images used in ImageNet pretraining to medical images. This was the most suitable domain for Swin-UMamba† to be used in, as it balanced global contextual learning exceptionally well with exceptional efficiency in terms of computation. This was achieved by drastically cutting down network parameters to 28M and FLOPs to 18.9G, providing a broader effective receptive field compared to CNNs or ViTs. This meant achieving state-of-the-art accuracy, beating seven of the best models with impressive results. This included a 0.7767 DSC on AbdomenMRI images, a 0.6931 DSC on Endoscopy images, and a 0.6219 F1 score on Microscopy images. However, limitations of this research included the need for a lot of computation and data required in the self-supervised adaptation stage. Future work will be based on creating Mamba-based foundation models pre-trained on large-scale medical images.

This paper by Qi et al. [22] focuses on the theme of semi-supervised semantic segmentation for real-world waste classification, thereby bridging the research gap that exists due to the lack of pixel-level annotated datasets and the inability of existing models to effectively handle intricately cluttered and unpredictable waste environments. The novelty of the study is the incorporation of the NUNI-Waste dataset with a continuous non-uniform data augmentation strategy that simulates natural lighting and object polymorphism without the loss of image data and an adaptive weighted loss function with dynamic masking for handling extreme class imbalance. The methodology for the study includes the use of a U-Net baseline with a ResNet-50 backbone that is trained with consistency regularization on the Zerowaste dataset. The U-Net structure is the best choice for the study due to the effectiveness of the encoder-decoder structure in restoring spatial information and the ability to perform well with smaller datasets by capturing global and local

features with higher precision than other models like the DeepLabv3+ approach. The proposed approach achieves a mean Intersection over Union (mIoU) accuracy of 55.37%, thereby indicating a notable improvement of 3.74% compared to the baseline Zerowaste approach. Although the study was successful in achieving the objectives and the proposed approach was effective in the context of the experiment and the theme chosen for the study, the study also indicates some limitations and drawbacks of the study and the approach proposed for the experiment. These limitations and drawbacks of the study and the approach proposed for the experiment need to be addressed by the future study that builds upon the present study and the approach proposed for the experiment by introducing dynamic weighting and testing the approach for robustness against physical adversarial attacks.

This study proposed by Ranftl et al. [6] has the theme of progressing the dense prediction tasks in computer vision that include monocular depth estimation and semantic segmentation tasks. It is done by the utilization of transformer models. The authors find that there is a serious research gap in that fully-convolutional networks downsample images stepwise which inevitably leads to the irreversible loss of feature resolution and granularity. This is very important in comparison with getting pixel-level predictions. The novelty of this publication is that it introduces the Dense Prediction Transformer (DPT) which is a variant of the convolutional backbone which uses a Vision Transformer (ViT). In the encoder-decoder model, a ViT encoder is used as the encoder to take image patches as inputs in the form of single tokens. The input tokens of various levels are later recreated into representation of pictures and progressively able to unite with a convolutional decoder. This design is the most appropriate because it does not involve explicit downsampling. This ensures a high-resolution representation at every stage and a global receptive field that results in finer-grained, globally consistent predictions. Subsequently, the DPT architecture leads to an impressive 49.02% mIoU accuracy on the ADE20K semantic segmentation dataset and a relative 28% performance improvement on depth estimation. The major shortcoming is that transformers require huge amounts of training data, and larger models do slightly worse on smaller datasets. Although the work in the future is not stated clearly, it suggests that future work may be conducted in the investigation of the data-efficient pretraining and replicating the architectural size on a larger scale dataset that can maximize its potential.

In this paper by Ali et al. [15], the author is committed to the issue of computer vision and waste detection and segmentation to enhance the recycling processes. The gap in the research that is noted by the authors is that the existing models and data sets do not consider the critical level of clutter and disorder of real-world waste streams that has strongly deformed objects. The new features of this work are the improvement of the InternImage CNN model with a new Multi-Scale Feed-Forward Network (MSFN), dilated convolution and composite loss function (cross-entropy and Dice) as well as a special Freezeconnect regularization module. The improved CNN model is most appropriate since the MSFN and dilated convolution effectively balance the trade off between local fineness and global contextuality dependencies which makes it very effective in identifying multi-scale, deformable objects in disordered backgrounds. Consequently, the model attained the state of art with a mIoU of 53.95% and 91.30% pixel accuracy on the Zero Waste-f test set. The only weakness is that less deformed objects such as cardboard also have less significant

improvements with the use of MSFN. It can be said that in the future, this exact model will be modified in order to trace the waste flow and streamline the processing paths.

This study presented by Islam and Alam [14] examines the theme of waste detection and classification through computer vision that will be utilized to automate the separation of garbage to enhance recycling process. One of the research gaps identified by the authors is that the current waste datasets and existing frameworks do not illustrate the contextual real-world estimates. These real-world wastes are considerably challenged by overlapping garbage items and their inference expenses are excessively high to implement edges in real life. The main novelty of this paper is that it introduces the concept of GACO, 2,200 contextual garbage dump images with 135,541 individual annotations that are totally human labeled along with a well-developed real-time object detector. The given methodology follows a single-stage neural network architecture based on YOLOv4. This particular model is the most suitable due to the fact that the CSPDarknet53 backbone cuts down the computation of the networks by 20 percent that allows the lightning-fast inference rate 55-58 ms needed to launch edge deployments in real-time. As a result, the framework achieves a reliable accuracy, specifically a mean Average Precision (mAP) of 66.08% at an IoU threshold of 0.35. However, limitations include poorer detection accuracy for glass and paper waste, alongside a restricted number of training examples. Future work will focus on enriching the dataset to handle wider environmental variations and optimizing the model further.

This article by Chen and Zhang [18] focuses on improving semantic segmentation and network durability with transformer networks. The authors find one research gap which is that convolutional networks can be not trained to be resistant to corrupted images and transformer models cannot be trained with low computational costs and lose their context when downsized. CESegNet, which presents a hierarchical simplified transformer encoder, two separate Context-Enhancement Modules (CEM) and a Contextual Fusion Decoder (CFD) is the novelty of this paper. Spatial-reduction attention is utilized to reduce complexity and the local context is preserved with the use of overlapping patch embedding and stable feedforward networks. The model is the most suitable model as it has been able to balance the global receptive field of the transformer with the control of the growth of parameters which leads to the highly stable segmentation in complicated surroundings. Consequently, CESegNet provides semantic parsing of the state-of-the-art and zero-shot robustness of a strong type. To be more precise, the CESegNet-Large model on the Cityscapes validation set has an impressive 82.21% mIoU and on the ADE20K set, the model has 48.77% mIoU. One of the weaknesses that were observed in the study was the constraints of hardware video memory as they limited ablation experiment training to 40,000 iterations. Although the explicit future work is not described in great detail, the authors emphasize that the great robustness of the network will contribute substantially to its practicality and flexibility.

Nahiduzzaman et al. [31] presented a deep learning framework which claimed to identify and work on some notable and critical points which therefore lacked in waste classification systems such as: Computational inefficiency, lack of real-time applicability and limited class diversity. A three-stage classification model was pro-

posed by the authors which combined a lightweight Parallel Depth-wise Separable Convolutional Neural Network (DP-CNN) paired with an Ensemble Extreme Learning Machine (En-ELM) classifier. Firstly a binary classifier was introduced to distinguish between biodegradable and non-biodegradable objects from the waste images. Secondly the waste images were distinguished between nine categories such as: glass,metal,medical waste etc from the above two classifiers and finally divided into thirty-six granular classes such as: foods, newspapers, beverage cans etc using the newly developed TriCascade WasteImage dataset consisting of 35,264 annotated images. The three staged classification model DP-CNN-En-ELM achieved robust performance: 98.77% AUC and 96% accuracy for binary classification, 98.57% AUC and 91% accuracy for nine-class categorization and 98.68% AUC with 85.25% accuracy in thirty-six-class fine-grained classification. En-ELM classifier demonstrated minimized misclassification in complex waste categories compared to PI-ELM and L1-RELM. Explainable AI (XAI) tools such as: SHAP, guided saliency mapping and GradCAM ensures transparency in decision-making and adopting real-world industrial applications. A hardware prototype showcases the usability in augmented real-time environments which validates the use case of the system in low resourceful environments. The author contributed significantly by upholding a high accuracy with low parameter model yet dealing with increased class complexity and computational constraints. The study can significantly contribute to waste sorting systems in low resource ensuring to fulfill environmental sustainability.

In their study, Maheri et al. [12] demonstrated the CO adsorption capacity of Biomass Waste Derived Porous Carbon (BWDPC) which is termed as one of the vital and prominent material and storage applications for carbon capture and its issue of predicting accurately. The performance differs from variable factors such as: elemental composition, environmental changes and surface area. Due to the changes of the properties in BWDPC, the need for an advanced predictive model is required resulting in advanced ML and DL techniques deployment. Rigorous preprocessing techniques are followed which includes scaling via standard normalization and imputation of missing values using Gradient Boosting Regressor (GBR). The data were augmented artificially by pointwise feature division on two datasets BWDPC-38 and BWDPC-46, which contains 28 and 36 features. Different models were evaluated such as: MLP, LSTM, GBR and CNN which assessed 7-fold cross-validation and grid search for hyperparameter tuning. CNN captured complex patterns through three hidden layers, for fast learning it used ReLU and Adam optimizer with minimal learning rate of 0.0001 for accurate and stable learning. The results validated CNN as best performing model with R^2 score of 85.73% $\pm$ 1.36 on BWDPC-38 which outperformed MGBR R^2 score (83.15% $\pm$ 2.8). Feature analysis results revealed temperature and carbon to pressure ratio as the most influential factor for CO adsorption. The limitation was also put forward by the author which claims LSTM's non-sequencials data limitedness. Future work involves refinement of feature selection, exploration of ensemble learning and intervention of dataset variation for better generalisation. An adaptive and strong machine learning framework is presented in the study which not only improves CO capturing efficiency but also takes a step ahead to data driven methodology to tackle climate change.

Otgonbayar and Mazzotti [21], proposed a mathematical model to assess the system's power and heat output from the impact caused by integrated post-combustion

capture facilities which consisted of energy and mass across the steam cycle of WTE plants. The integrated WTE-carbon capture and storage(CCS) system can be analysed by taking account of three main objectives: the heat supply, electricity output and the capture efficiency. The heat produced from municipal solid waste (MSW) can be captured by CCS plants. Direct Heat recovery and Indirect Heat recovery systems for District heating are stated. In terms of power-based scenario results, WTE plants delivered production of heat on a smaller scale but 37% efficiency reduction is stated because 99% of heat is available for capturing. Recovery of waste heat in both high and low temperatures can be achieved in CCS plants. A high efficiency reduction from 42% to 36% was pointed out while analyzing COP between 4 and 8. Reboiler temperature, district heating, configuration and technology used for capturing combinedly gives predictions about quality as well as amount of heat recovery. The model's data was compared with KVA Hagenholz energy output data and errors detected were, Energy in waste 1.4%, HP stream energy 0.6% and electricity out was 8.1%. The model forecasted good results in comparison with the data of real plants.

According to Lin [19], an advanced garbage classification model that integrates image recognition and AI techniques was presented. The model aims to fix problems with manual waste storing, which is mostly slow and inefficient especially in terms of complex environmental contexts. Previous methods struggled with extracting image features properly, limited adaptability to different lighting, hidden objects and many types of wastes. The author introduces a hybrid deep learning architecture that cascades an improved PCANet with SDenseNet to overcome the challenges. The enhanced PCANet uses unsupervised learning using PCA-based convolutional kernels, to extract rich multi-scale features it is optimized through region clustering and entropy maximization. SDenseNet is used for high level feature learning, dense connectivity, skip connections to reuse features, transition modules, balancing computational efficiency especially for waste image classification.The architecture is developed through a layered learning strategy and standard back propagation, with standard training methods to improve accuracy. By experimental evaluations, it is demonstrated that the proposed model outperformed traditional CNNs, achieving 96.36% sensitivity, 96.58% specificity, and 94.25% accuracy, with an AUC of 0.9884—better than models like VGG-16 and Wide ResNet in all metrics. Furthermore, it maintained competitive training time (54.23 minutes for 156 epochs) and real-time classification speed (11.25ms per image). Notably, it also performed well in real-world settings like beaches and greenbelts. The actual contribution of this article is its innovative combination of layered learning and effective feature merging, which has achieved high classification accuracy keeping computational demands low. However, the author struggled with classifying similar looking waste types and scaling highly noisy data. Future prospect of the research intends to use variability of data and develop adaptive training methods to improve its flexibility and to enhance generalizability.

Ahmed et al. [9] demonstrated a deep learning-based framework for classifying recyclable waste materials which points out the inefficiencies in traditional waste management systems. The study enveloped the pressing environmental and its operational challenges compared with manual and automatic operational challenges. With the aim of solving this, multiple algorithms MobileNetV2, DenseNet169, CNN

and ResNet50V2 were evaluated and contributed to enhance model performance. This study builds a model following the DL approach. Progression was made by addition of 11 hidden layers, one Flattened type, two Dropout types, three Conv2D types, three Dense types and MaxPooling2D. Accuracy measurement of 87.22% including error rate 0.517 was achieved. The accuracy swiped to 88.5% and an error rate of 0.324 after tuning of hyperparameter and cross validation. The ResNet50V2 algorithm outperformed other models in every aspect with F1-score 98.38%, Precision 98.35%, Accuracy 98.95% and Recall 98.38%. This data represents a significant improvement over the traditional machine learning approaches. 70% of the dataset was used for training and 30% of the dataset was used for training and testing. Compared to prior work, the study fills a gap in the integration of high-performing, lightweight CNN models for real time recycling applications. Limitations such as reliance on a relatively small and imbalanced dataset, which may affect real-world generalization. Future developments in dataset expansion, integrating multimodal data, and deploying the model in smart waste bins.

A notable provocation in managing urban waste, accurately classifying waste to recyclable and organic categories was addressed in the article by Chauhan et al. [10] by introducing the research work and introduced a Deep Neural Network based learning system called LADS (Learning-based Approach with Deep Neural Networks for Smart Systems). Smart cities waste classification techniques are termed as inefficient and scarce because of using primitive models and not being able to sufficiently fulfill the need accordingly. A 12 layered CNN architecture which includes three layered MaxPool2D, three dense layers and six layered Conv2D for training up the model which included Keras and TensorFlow for preprocessing steps such as: normalization, augmentation of data and resizing of image. Initial categorization into recyclable and organic waste was done using a dataset consisting of 25,077 images. Heavy Weighted models like VGG16,ResNet34,AlexNet etc were easily beaten by the proposed model with accuracy 94.53%, precision 88.7% and recall 94.37% whereas ResNet34's showed accuracy 93.5% and precision 93.2%. The proposed model got heavily trained up using dataset of real world Images and synthetic for extensive verification and validation for the models robustness. A scalable, accurate and low cost model was deployed using real-time data fusioned into deep learning for urban environments; was the gap the research paper focused on. Classifications accuracy enhanced, pipeline based efficient preprocessing on heterogeneous dataset and CNN architectures customisation was the notable contribution through the research. Nonetheless due to having similarities in images recyclable waste was often misinterpreted as organic waste can be termed as one of the limitations of the model. Exploration of waste categorisation, classification methods enhancement and integration of real time data with large datasets with IoT are the future prospects recommended by the authors.

A deep learning-based predictive framework, DLHHV-MSW, was newly formulated by Ishaque and Florence [27] to address the challenge of accurately estimating the Higher Heating Value (HHV) of MSW which is a critical parameter for optimizing Waste-to-Energy (WTE) processes. Traditional HHV determination methods are often costly, time-consuming, and environmentally harmful. After preprocessing steps like min-max scaling and mean imputation for missing values, their methodology uses a Deep Belief Network (DBN), comprising multiple Restricted Boltzmann Ma-

chines, to predict HHV from MSW elemental composition (ash, carbon, hydrogen, nitrogen, oxygen, sulfur, and water). A important innovation is the hyperparameter optimization of the DBN architecture implementing Oppositional Cat Swarm Optimization (OCSO) and this extends standard Cat Swarm Optimization with oppositional-based learning for enhanced convergence and research of the solution space.On a benchmark dataset, DLHHV-MSW demonstrated superior performance, obtaining a Correlation Coefficient (CC) of 0.996, a RMSE of 2.342 MJ/kg, and a Mean Absolute Deviation (MAD) of 0.136 MJ/kg. The research found that the RMSE was 16-25% lower and the MAD was upto 50% better than SVM and RF models after just 6.1 seconds of training time. The study addresses a gap by creating a better, automated, and computationally efficient alternative to traditional calorimetric methods and static ML models for HHV prediction that are already available. The synergistic DBN-OCSO integration which leads to an enhanced predictive accuracy and robustness across diverse regional datasets are the important contributions of this paper. The authors address real-world implementation in WTE facilities, adaptation for additional waste streams, and the integration of further regularization methods to handle high-dimensional data as areas for further research.

In their work, Madhavi et al. [29] proposed SwinConvNeXt. This is a novel deep learning architecture which acknowledges the challenges of real-time garbage image classification, particularly the limitations of existing models in accuracy, computational efficiency, and handling visually similar objects of differing sizes. An enhanced Swin Transformer was used as a method which uses hierarchical feature extraction that is a shifting window mechanism, and global multi-head self-attention with ReLU activation that helps to capture global image features. This is paired with an improved ConvNeXt block, optimized with depth-wise separable convolutions, normalized layers, and ReLU activation for efficient local feature extraction along with a spatial attention mechanism to emphasize relevant image regions. Images were preprocessed by resizing to 224x224 resolution and the model was trained for five epochs using Adam optimization. SwinConvNeXt achieved 98.97% accuracy, 98.42% precision, and 98.61% recall on the public Garbage Classification dataset. The research gap is filled by suggesting a lightweight, computationally efficient solution that outperforms previous methods in classifying often small or overlapping garbage items, a common problem in existing models.Major contributions are the synergistic fusion of enhanced global and local feature extraction capabilities within a lightweight design resulting in superior performance and reduced computational cost compared to state-of-the-art models. The authors did not explicitly state limitations of their current model but proposed future directions that are: investigating real-time object detection and tracking, multi-modal sensor integration (e.g., RGB-D, hyperspectral imaging), utilizing 3D images, and researching dynamic model adaptation for diverse real-world scenarios.

An outstanding effort towards enhancing waste sorting and recycling efficiency was given by Sayem et al. [32], who acknowledged the problem of inadequate performance of existing deep learning models due to limited trash image categories in datasets and the complexity of real-world waste. They proposed a robust deep learning-based approach for both classification and object detection, using a comprehensive dataset of 10,406 images across 28 distinct recyclable waste categories. Augmentation, resizing (224x224 for classification, 640x640 for detection), and nor-

malization were used for data preprocessing. The researchers presented 2S DenseViT for classification which is a unique dual-stream network that combines pre-trained DenseNet-201 and MaxViT feature extractors with relevant features that are propagated to a final classifier. They also incorporated the GELAN-E (Generalized Efficient Layer Aggregation Network) model for object detection. The 2S DenseViT model has a classification accuracy of 83.11%, precision of 83.33%, and an F1-score of 83.05%. The GELAN-E model obtained a mean Average Precision (mAP50) of 63% for object detection which is better than several YOLO variants. This paper addresses a research gap by showcasing models trained on a complex, multi-class dataset representing industrial conditions that the previous researchers could not do due to using simpler datasets. Main contributions validate the novel 2S DenseViT architecture and the superior performance of GELAN-E in waste detection. The authors conceded that their 2S DenseViT struggled with visually complex categories like "bottle-multicolor" and proposed future work to include investigating advanced detection techniques, expanding the dataset, and optimizing models for computational efficiency and specific bottle types.

Malik et al. [8] addressed the challenges encountered during managing Municipal Solid Waste(MSW) and proposed a deep learning model for automated waste classification. The article focused on a specifically utilized CNN following the EfficientNet-B0 architecture, which uses compound scaling to further classify wastes into various labels such as: plastic, metal, and paper. Initially, the model was pre-trained on the broad ImageNEt dataset and then refined by using liter specific images to boost its classification efficiency. The new method had about 81.2% accuracy which is almost on par with the complex EfficientNet-B3 model but the key highlight is that it uses far fewer parameters which leads to a 4x decrease in computational demand (in FLOPS). This demonstrates that the model is highly suitable for a resource-tight environment for its low computing cost. Past studies denoted that CNN architecture was highly demanding on computational cost and when scaled down, it couldn't perform accurately. The authors have successfully addressed this research gap by introducing this newer model which can maintain high accuracy and be computationally inexpensive at the same time. Despite the progress, the authors pointed out that the model's accuracy heavily depends on the range and diversity of the local training data making it region-specific. They suggested future search should include data augmentation and building hybrid models using EfficientNet architectures and vision transformers so that it would perform better in a wider variety of geographical areas.

2.3 Summary of Key Findings

The aggregated research papers have developed an increase in ML, DL and SSL deployment throughout waste management systems, from waste type classification and metric evaluation in waste to the optimization of WTE processes. Numerous DL architectures such as: CNNs (VGG, MobileNet, custom architectures like Swin-ConvNeXt and ResNet) are used in various studies for achieving high scale accuracy in classification or categorization of waste type and identification through image recognition. Various Self supervised learning methods such as: SimCLR, DINOV2, SAM along with different decoders like DPT, UNET, DeepLab etc are also used

in numerous segmentation tasks for various precise identification of objects which can further be applied in the waste domain. Waste recycling rate and autonomous sorting techniques can be boosted with the foundational step to categorize waste efficiently for environmental sustainability. For improving models robustness, authors consistently demanded for diversified large datasets and exploration methodology such as: data augmentation and transfer learning.

Apart from basic ML algorithms, advanced algorithms implementation such as: Vision transformer, self supervised learning in the waste domain field seems to be limited. Advanced AI, ML, DL, SSL etc have the potential to be implemented on automated sorting in facilities for efficiency enhancement, robust identification. Some studies pointed to the implementation of DL methods in the waste classification domain, ML adoption for semantic segmentation and other methodology of Self supervised learning for segmentation and classification utilizing different methods. ML and DL implementation on waste type categorization and nutrient metric evaluation in the waste domain is already in advancement but facilities have tremendous scope of implementing SSL for recovering of undetermined waste, optimization of processes, modelling predictiveness and advanced control strategies. The complex nature of the different waste and variability of data types from waste plants make the model implementation complicated.

Deep learning, Machine learning and Self supervised learning have immense possibilities of growth in this deformed and cluttered field to improve efficiency, sustainability and intelligent waste and water management overcoming the research challenges such as: generalization of model, quality and data availability on a larger scale. The tabular representation of research paper workings is presented in Table 2.1.

Table 2.1: Summary of Reviewed Literature

Segment	Prime Focus	ML/DL Methodologies	Key Insights	References
Waste Segmentation Datasets	Dataset creation for real-world waste segmentation under clutter and deformability	Instance + semantic segmentation, RGB-NIR fusion, semi-supervised learning	Highlights challenges of occlusion and non-rigid waste; spectral and SSL methods reduce annotation dependency and improve segmentation accuracy	[7], [17]
Cluttered Scene Segmentation	Accurate segmentation in complex, occluded environments	CNN encoder-decoder, attention modules, deformable convolutions, Transformers	Boundary-aware learning and feature amplification significantly improve segmentation performance in dense scenes	[15], [16], [18], [22], [26]
Vision Foundation Models	Learning generalizable visual representations via self-supervision	DINO, DINOv2, SimCLR, ViT, Mamba-based models	SSL enables strong transfer to segmentation tasks; large-scale pretraining and domain adaptation yield state-of-the-art results	[3], [4], [13], [20], [25], [28]
Segmentation Architectures	Advanced architectures for dense prediction and universal segmentation	Vision Transformers (DPT), SAM, HQ-SAM	Transformer-based models outperform CNNs; SAM variants enhance boundary precision and generalization	[6], [11]
Waste Classification & Detection	Automated waste recognition for sorting and recycling	CNNs, Transfer Learning, YOLO, Swin-ConvNeXt hybrids	High classification accuracy (>90%) achievable; hybrid and multi-task models improve real-time performance	[8], [9], [14], [19], [29], [31], [32]
Waste Management & MSW Modelling	AI-driven waste management and environmental impact assessment	DNN, IoT integration, regression models (RF, XGBoost, ANN)	AI improves collection efficiency, environmental assessment, and waste-to-energy planning	[10], [12], [21], [27]

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

In order to implement an effective waste segmentation system, requirement of robust computational setup is required in order to process feature maps at high dimension involving Vision Transformer architecture at backbone. The research purpose served its model development and training by utilizing the CPU of Intel Core i7-14700K along with NVIDIA GeForce RTX 4080 Super (16GB VRAM) GPU. The primary and secondary memory used in the research environment is 64GB of DDR5 RAM and 1TB SSD. Vision Transformer DINOv2 encoder [13] employed in our backbone and custom decoder with advanced dense prediction and boundary context aware components usually are memory intensive so a high performing setup is needed in the first hand. On the other hand the model development utilized PyTorch framework along with CUDA for proper utilization of GPU acceleration on the software end. The features of DINOv2 are imported through HuggingFace library and the model development code was scratchly built and written for combining the backbone features with semantic gap for cluttered scenes. The training process advocated a batch size of 2 along with gradient accumulation steps of 16 which effectively created a batch size of 32, making it compatible with DINOv2 to get better performance. The training protocol was developed after trial and error for better convergence without making the CUDA memory overflowing. Table 3.1. portrays the training, hardware and the software components used for model development.

Table 3.1: Hardware and Software Configuration for Model Development

Category	Component	Specification
Hardware		
	Processor (CPU)	Intel Core i7-14700K (20 Cores, up to 5.6 GHz)
	Graphics Card (GPU)	NVIDIA GeForce RTX 4080 Super (16 GB VRAM)
	System Memory (RAM)	64 GB DDR5
	Storage	1 TB SSD (NVMe)
Software		
	Operating System	Windows 11
	Deep Learning Framework	PyTorch (with CUDA GPU Acceleration)
	Backbone Library	HuggingFace Transformers (DINOv2)
	Augmentation Library	Albumentations
	Dataset Annotation Tool	Roboflow
	Programming Language	Python 3.10
	Development Environment	Visual Studio Code
Training Configuration		
	Effective Batch Size	32 (Physical Batch 2 × Gradient Accumulation Steps 16)
	Mixed Precision (AMP)	FP16 forward/backward + FP32 master weights
	GPU Memory Strategy	Gradient Accumulation + AMP to fit 16 GB VRAM

3.2 Societal Impact

Reduction of intensive manual labour, contamination of the hazardous chemical materials and sorting timespan is the beneficiary part from the model deployment. Improved recycling material time and labour make the deployment moving towards a circular economy where uncontrolled urbanization regions like Bangladesh can be benefited. Excessive dumping of waste in landfill makes the air more polluted and chemicals prevails through wastewater streams in highly populated areas. The system is designed to mitigate environmental hazardous factors and focused on public health through removal of manual interventions. A safe industrial system is also ensured through the full deployment of the system.

3.3 Environmental Impact

The system assures proper automated sorting of the waste materials which then further be recycled or reused without keeping it unused. This method serves the greatest interest of the environment due to balancing the ecological environment, recycling and recovering underpressed material due to human sorting negligence. The system is primarily focused on five different classes but can also be subdivided into other classes like SpectralWaste, it shows the optimized nature of the model. Reduction of methane and toxic leachate can be ensured through sorting it precisely. Finally the proposal of a newly created dataset totally from dumpward, streets and dustbin in the context of Bangladesh critically addresses the dataset deficit in Asian countries. The model performance across different geographical regions and waste types makes it more efficient for deployment in any other geographical areas.

3.4 Ethical Issues

The research focused on ethical constraints such as: data collection, privacy, representation and transparency. The newly created dataset named WasteRefine's picture collection and annotation is done at firsthand from different regions such as: roadsides, landfills, municipal dustbins, dumpyard and domestic areas. Filtration and sorting of the images is strictly done for proper ensurement of no personally identifiable information (PII), for example no human faces, privacy documents, personal belongings is included in the proposed dataset. The research also acknowledges the ethical progress of technological use while development of models such as: proper acknowledgment to the other resources utilization etc. The data acquisition from other sources for the research purpose is properly cited and acknowledged according to the specific regulatory requirements. The benchmarking process of the model across different dataset aligns with the specific data and splitting order to abide by guidelines.

3.5 Standards

The data standard named COCO (Common Objects in Context) is adhered to for the proposal of WasteRefine dataset [33]. Precise polygon mask annotation for semantic segmentation is used rather than any bounded box alignment. The labeling of the images was extensively done by two individuals and further validation of the annotation was done by another two individuals to ensure no error margin. For annotation of the dataset to generate ground truth mask Roboflow [23] was utilized, it ensured seamless and precise polygon drawing for complicated, deformed, cluttered, occluded and translucent objects at higher details. The dataset images imported in COCO format were further converted to semantic segmentation ground truth which is standard for the work. To conclude the COCO and semantic segmentation ground truth mask were both ensured in the dataset for reusability.

3.6 Project Management Plan

The research was extensively done over a period of one year dividing into three stages. Initial four months were utilized for literature analysis across the waste domain and identifying the issues regarding this domain for proper model development. The field work for collection of dataset involving site exploration, identification of waste collection area and dumpward observation was also done at the initial stage. The next four months were structured for dataset development and annotation purposes. It also included the finding of other related dataset for proper benchmarking. Alongside, other models assessment and methodological development were also explored throughout this period. The final four months were utilized for model development purposes. The custom decoder design with rigorous trial and error of each and every individual component and different training strategies among different dataset was also employed for high precision performance. The writing of the research paper was developed in all the three stages and refinement of the writeup was also ensured at the end of the work to ensure proper research objectives as planned.

3.7 Risk Management

Hardware memory limitations are among the first risks that the research needed to tackle. VRAM of the 16 GB NVIDIA GeForce RTX 4080 Super had limited memory capacity compared to the need for training large transformer models. Usage of gradient accumulation steps was deployed to mitigate this condition without hurting the performance due to batch size reduction. The training process advocated a batch size of 2 along with gradient accumulation steps of 16 which effectively created a batch size of 32, making it compatible with DINOv2 to get better performance. The second risk was about heavy class imbalancement of different datasets that the research used for benchmarking. ZeroWaste-f dataset has metal 50 times lesser than the dominant class cardboard [7]. As a result the models cannot learn the textures and structures of every class properly. To manage the proposed risk, training strategies were employed for specific dataset distinctively and a proper class balanced dataset was proposed through WasteRefine. For mitigation of overfitting issues in smaller datasets trained on vision transformers the researchers employed various augmentation strategies according to the need of different datasets. To further assess the robustness of the model across different varieties of waste, ZeroWaste-f, SpectralWaste and WasteRefine dataset is used. It proved generalization capabilities of the architecture in reliable as well as challenging scenarios of industry.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

The chapter gives the full methodology of the WasteRefine segmentation framework, including the design rationale, architectural details, data collection and preprocessing pipeline, and the training strategies that were used for each dataset in this work. The ultimate goal of the study is to create an effective, generalizable semantic segmentation system that will be able to correctly label waste materials at the pixel level in various mixed waste real-world data. We achieve this by adopting a new segmentation architecture that is built on the DINOv2 Vision Transformer architecture, but adds to it an additional multi-scale decoder, including Pyramid Pooling, Squeeze-and-Excitation channel attention, and comprehensive boundary supervision branch. The entire methodology follows four primary steps, which are outlined in this chapter: dataset collection and annotation, data preprocessing and augmentation, model design and specification, and implementation and training.

The general outline is based on a guided semantic segmentation pattern. The model is trained on an input RGB image to produce a fine pixel-level class map that identifies each pixel with one of the categories of waste material from the previously defined set of categories or as the background class. Training and evaluation are performed using three different datasets: the dataset WasteRefine that was collected and labeled as a novel development of this paper, the publicly available ZeroWaste-F benchmark [7] and the SpectralWaste dataset [17] its RGB type. Each of the two datasets presents its own problem such as the imbalance between classes, object size, visual confusion and the lack of samples. This brings about dataset-specific training and augmentation conditions without altering the overall model architecture. Reproductibility by thorough validation is the fundamental principle of the methodology. Each experiment is conducted using three random seeds per dataset and the findings are reported as the average and standard deviation of those seeds as considered in the deep learning research. The architecture consists of two versions of the backbone DINOv2 ViT-Small (ViT-S) and DINOv2 ViT-Base (ViT-B) that can be compared directly to the performance in terms of model complexity. One of the most important results of the research is that the ViT-S version which has about 25.16M parameters and outperforms significantly larger baseline models. This proves the effectiveness of the decoder architecture proposed in the study. Figure 4.1 depicts a detailed methodology used in this study.

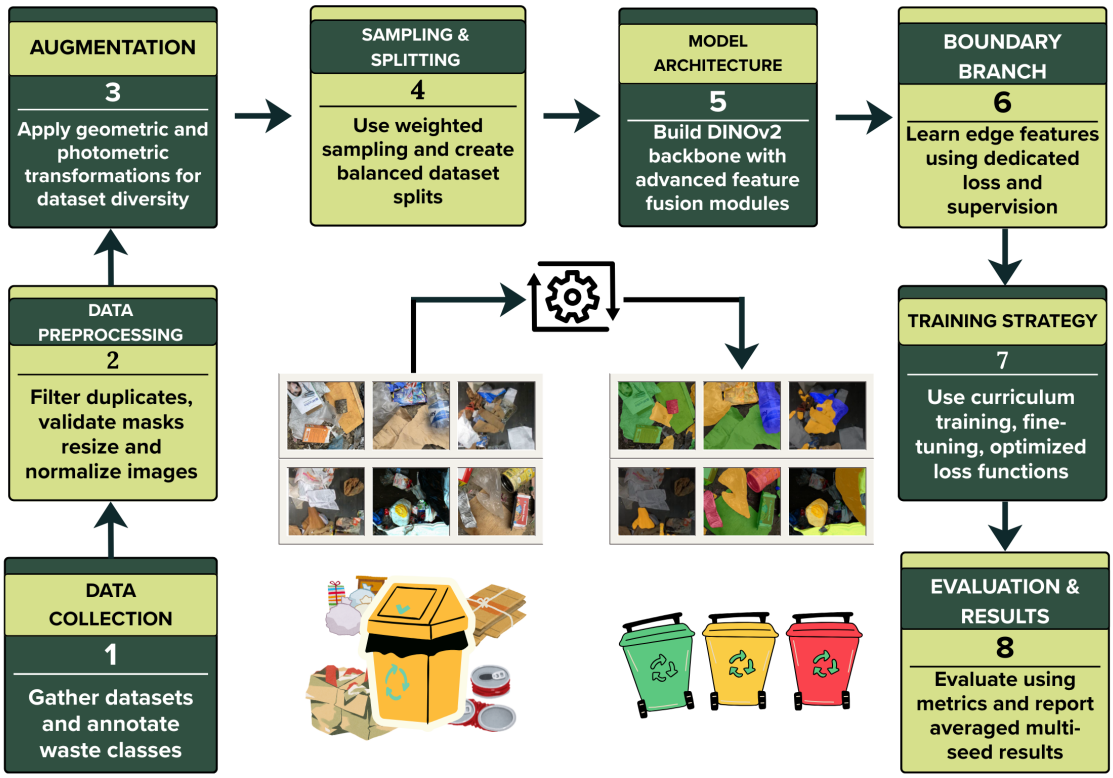


Figure 4.1: End-to-end methodology pipeline of the WasteRefine framework.

4.2 Preliminary Design or Model Specification

4.2.1 Model Design

The suggested segmentation architecture is a custom one that consists of a frozen or fine-tuned DINOv2 Vision Transformer backbone with a new multi-scale decoder. The design is aimed at resolving the challenges of waste material segmentation that were found in real-life waste streams, such as high intra-class visual similarity of specific waste types, extreme class imbalance of certain waste streams, the occurrence of thin and vague objects like filament threads and video tape strips. Also, the necessity to achieve fine-grained boundary accuracy in order to distinguish between adjacent and overlapping waste pieces in order to prevent misclassification. The entire architecture is shown in Figure 4.2.

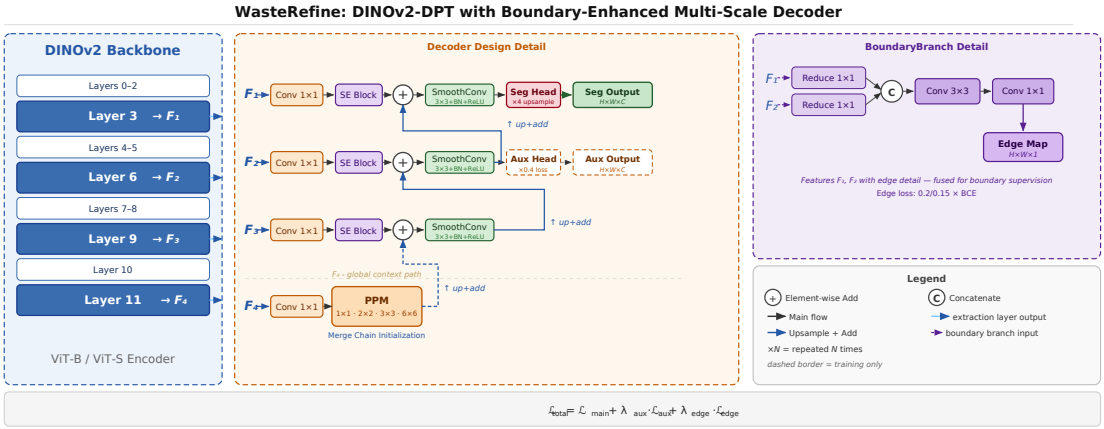


Figure 4.2: Architecture of the proposed WasteRefine DINOv2-DPT model with Boundary-Enhanced Multi-Scale Decoder.

Backbone: The proposed model is based on DINOv2 Vision Transformer [13] which is developed by Meta AI and pretrained on 142 million images but using self-supervised learning through DINO v2 objective. Rather than convolutional networks that include ResNet and EfficientNet, DINOv2 was chosen because of a number of reasons. To begin with, self-attention in Vision Transformers attracts long-range spatial-dependency across the image as a whole which allows the model to make inferences about the world beyond the image. For example, a thin filament thread that is not only detected on its own but also in context with the conveyor belt backdrop as a global context. Second, the large-scale pretraining of DINOv2 can generate feature representations that can generalise effectively to more specialised problems like industrial waste where there are limited labelled examples. Third, ViT patch-based processing produces spatially structured feature maps whose resolution is $H/14 \times W/14$, and dense prediction tasks [6] can be directly applied to them.

The features are obtained at four layers of the transformer 3, 6, 9, 11 which are associated with increased levels of high-level feature representation. The features of layer 3 (F1) are the ones that have the best spatial detail and are most sensitive to local textures and edges. Patterns taken in layer 6 (F2) are intermediate in nature. The high level region semantics are found in layer 9 features (F3). The deepest extracted features (F4) in layer 11 have the richest semantic content with the least positional precision. They use two variants of the backbone. One is ViT-Small (ViT-S) that has 384-dimensional feature embeddings and 22.05M backbone parameters, while the other one is ViT-Base (ViT-B) that has 768-dimensional feature embeddings and 86.58M backbone parameters. The backbone is frozen during the warm up training stage. In later steps, the backbone is further refined with a multiplier of learning rate of 0.1 with the decoder so that there is stable adaptation of the features but there is no severe forgetting of the pretrained forms.

Conv 1×1 Projections: An independent 1×1 convolutional projection layers exists where every resultant feature map from F1 to F4 is passed through it first. The purpose of these projections is to downsample the features of the backbone (384 when using ViT-S or 768 when using ViT-B) into a new 256-dimensional representation on which the decoder operates. With the 1×1 kernel, each location in space is projected

separately so that the projection does not mix spatial information between pixels in neighbouring locations. This ensures the spatial fidelity of the original features is preserved. This decoder architecture is what also makes the decoder architecture backbone independent. The same AdvancedDPTNeck is going to execute the same code on both the ViT-S and ViT-B backbone but the only difference is the projection weights.

Pyramid Pooling Module (PPM): Pyramid Pooling Module was first used in the PSPNet architecture [1] and is only used on the final feature map F4 after projection. The PPM spatial pools the feature map with four scales of 1×1, 2×2, 3×3, and 6×6 global average pooling and concatenates the pooled representations back to original pixel resolution. Then a bottleneck convolution is applied. This operation offers explicit multi-scale contextual awareness to the model at the semantic level. The 1×1 pooling of the model captures a single global representation of the entire image. The 2×2 and 3×3 captures more broad regional distributions while the 6×6 captures more finer sub-region context. Application of PPM is only at F4 due to semantically deepest feature representation that provides most meaning to the global context. Using PPM on the shallower layers like F1 or F2 would blur the small positional information that these layers are specifically stored to give.

Top-Down Feature Fusion (AdvancedDPTNeck): The entire decoder is based on the top-down feature pyramid fusion approach that is built on the inspiration of the Dense Prediction Transformer (DPT) decoder introduced by Ranftl et al [6]. Firstly, the F4 representation with PPM addition, the decoder upsamples and adds features sequentially at the bottom to the top of the feature hierarchy in the following order: F4 $\rightarrow$ F3 $\rightarrow$ F2 $\rightarrow$ F1. Each time a fusion step is made, the representation of the deeper level is upsampled and is then appended element-wise to the projected feature of the current level. This design carries semantic context at a global scale into the positional specific shallow layers originating in the deep layers in order to let the model know *what* an object is (F4) and at the same time know precisely *where* its boundaries are (F1 and F2). The addition operation is parameter free and computationally efficient as compared to the concatenation based fusion which would raise the channel dimensionality with every step.

Squeeze-and-Excitation (SE) Blocks: A Squeeze-and-Excitation block [2] is used on every step to the entire feature map after the completion of Conv 1×1 Projections. The SE block performs channel-wise attention. Firstly, it average-pools the spatial dimensions across the entire feature map to get a channel descriptor vector and then inputting this into two fully connected layers of an activation of Sigmoid and a non-linearity of ReLU. Then the resultant weights channel-wise are used to rescale the feature map. The idea of adding SE blocks before each fusion point is to focus on the most relevant channels and inhibit noisy ones in order to achieve an accurate prediction. It is used to remove redundant or conflicting information after each projections from distinct backbone feature which further combines with the feature signals in addition component, resulting in better upsampled deeper representation and projected shallow representation. This is especially crucial during waste segmentation, as the visual data of class differentiation (colour, texture, shape) significantly differs between types of waste.

SmoothConv: SmoothConv operation is applied to the features coming from element wise addition component. SmoothConv is a 3×3 convolutional layer sequentially followed by Batch Normalization layer then a ReLU layer and a 2D Dropout with rate of 0.1. The convolution 3×3 positionally blends the data of neighbouring pixels which results in the representation of more coherent features with fewer edges before the next fusion or the final prediction. Normalisation distribution of features per batch will prevent training issues and fasten convergence during the batch normalization training. A regulariser utilized in the dropout of this step minimises the probability of overfitting smaller data sets like WasteRefine and SpectralWaste. The reason behind the use of the 3×3 kernel is intentional. This is the smallest size of the kernel that supports the use of spatial neighbourhood information (8 pixels surrounding the pixel) yet does not result in an excessive spatial blurring introduced by larger kernels (5×5 or 7×7) and in turn destroys precision of boundaries.

BoundaryBranch: The BoundaryBranch is a special prediction head that is geared towards resolving the difficulty of narrow and unclear object boundaries. One intended use is for the waste segmentation especially in classes like filament (a thread-like and contains within one to two pixels) and video tape (narrow strips of objects). The branch receives as input the two shallowest projected feature maps, F1 and F2. It captures the finest spatial-resolution and the most delicate spatial-resolution. They are reduced to lower dimensional representations separately using a 1×1 convolutional layer, and then the two representations are concatenated and processed using a 3×3 convolutional layer with Batch Normalization and ReLU. Lastly, a 1×1 convolutional layer that yields a single channel logit map. This map is expanded to the full resolution of the image and is trained with a binary cross-entropy (BCE) loss against ground truth edge maps depending on the gradient of the edges of the boundaries of the segmentation masks. This map is expanded to the full resolution of the image and is trained with a binary cross-entropy (BCE) loss against ground truth edge maps. This is done depending on the gradient of the edges of the boundaries of the segmentation masks.The major design difference between our BoundaryBranch and the Boundary Enhancement Module (BEM) of COSNet is that the boundary signal is clear specification.

COSNet [26] BEM is introduced as an intermediate layer which guides to construct features implicitly based on a max-pooling and subtraction mechanism to provide additional details to the internal feature representation. But it does not generate explicit edge prediction. In contrast, the explicit binary edge prediction map is a separate model output of our BoundaryBranch, with a loss of its own. A direct oversight like this is a commitment to the backbone properties and shallow projection in order to actively encode the gradient sensitive information. This results in more powerful gradient signals and specifically to the acquisition of gradient sensitive representations. This difference can be especially useful with those thin-object classes in which automatic boundary information is not available.

Segmentation Head and Auxiliary Head: The main segmentation head uses as input the F1-level SmoothConv output which is the most refined feature representation of the decoder. It is a combination of F4-based global semantic context and F-based fine spatial detail and applies a 3×3 convolutional layer. Then Batch Normalization, ReLU, Dropout (0.2) and a 1×1 output convolution with C channels

are followed, where C is the number of classes of the dataset. Bilinear interpolation is then used in upsampling the feature map 4 times to resample the prediction back to the original input size. It is the major segmentation output $H \times W \times C$. During training, an auxiliary segmentation head has the same architecture by utilizing F2-level SmoothConv output that generates a secondary prediction. The auxiliary head is completely disposed of in the inference phase. This is because it injects a direct gradient signal into the middle levels of the decoder which addresses the vanishing gradient problem. Otherwise, it would happen when the training signal is required to pass through all the depth of the top-down fusion pathway. The auxiliary head is associated with the training loss of 0.4 where the main head is not affected by the weight of the auxiliary head.

4.2.2 Model Training Curriculum

The training is conducted in a two-step curriculum manner to make sure that the pretrained DINOv2 backbone is not destabilised by an important gradient update at the first training stage. Also, ensuring it can be optimised fully at the final training stage.

Stage 1 Warmup with Frozen Backbone: In the first step, all the DINOv2 backbone parameters are fixed. Only the decoder components, the projections, PPM, SE blocks, SmoothConv layers, BoundaryBranch and segmentation heads are trained here. Without the learning rate, the training of the dataset-dependent number of epochs (5 epochs in WasteRefine and ZeroWaste-F, 10 epochs in SpectralWaste) is repeated. The stage would allow the decoder to accumulate normal semantic gradient flow, and then the backbone will be handed over to backpropagation. In case the backbone was poorly adapted according to the first epoch, the randomly initialised decoder weights would bring about massive, noisy gradients that would pollute the already very informative pretrained representations. One method of making sure the decoder becomes trained when the backbone is frozen at warmup would be to make sure that the decoder has learned to make effective use of the fixed backbone features before beginning to engage in full fine-tuning. This ensures the establishment of a stable baseline before engagement in full tuning.

Stage 2 Fine-Tuning with EMA and Cosine Annealing: During the second phase, all the backbone and decoder parameters are jointly trained in AdamW optimiser. The backbone multiplier of the learning rate is positioned at 0.1 instead of letting the backbone over-update in comparison to the decoder. This ensures the actual learning rate of the backbone is one-tenth of the learning rate of the decoder. The learning rate is set to 1×10^{-4} for all datasets. The decay of the weight is normalised to 1×10^{-4}. Stage 2 is based on the CosineAnnealingLR scheduler that declines the learning rate of the initial base value to zero during the rest of the epochs. This helps the model not to fluctuate around the local minimum.

Stage 2 has Exponential Moving Average (EMA) of model weights with a decay factor of 0.999. The EMA model keeps a smooth version of the weights of the parameters by averaging the parameters values over the training steps. The evaluation EMA weights are then evaluated at the end of training instead of the raw optimised weights. This is done since the EMA weights are more likely to make

better-generalising and more stable predictions by averaging out the noise caused by stochastic gradient updates. It is especially useful when the dataset is limited in sample size like WasteRefine (2,213 images) and SpectralWaste (852 images).

All experiments are compiled to run on Automatic Mixed Precision (AMP) and both forward and backward passes are run in FP16 arithmetic where possible and in FP32 arithmetic where any critical numerical operation must be performed. This not only stores memory within the GPUs but also speeds up training without losing any significant accuracy. When gradient accumulation is applied 16 times, a batch size of 32 is practically achieved which partially compensates the physical memory constraints of the GPU memory used by WasteRefine and ZeroWaste-F. This accepts high-resolution 518×518 input images. All of the training hyperparameters of the three datasets are summarised in Table 4.1.

Table 4.1: Training hyperparameters for all three datasets.

Hyperparameter	**WasteRefine**	**ZeroWaste-F**	**SpectralWaste**
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Backbone LR Multiplier	0.1×	0.1×	0.1×
Weight Decay	1×10^{-4}	1×10^{-4}	1×10^{-4}
Epochs	100	50	300/500
Early Stopping Patience	15	20	35
Warmup Epochs (frozen backbone)	5	5	10
Batch Size	2	2	2
Gradient Accumulation Steps	16	16	16
Effective Batch Size	32	32	32
AMP (Mixed Precision)	Enabled	Enabled	Enabled
EMA Decay	0.999	0.999	0.999
LR Scheduler	CosineAnnealing	CosineAnnealing	CosineAnnealing
Copy-Paste Probability	0.30	0.25	Disabled
Edge Loss Weight (w_{edge})	0.20	0.20	0.15
Main Loss	Dice + Focal	Dice + Focal	Dice + Focal
Aux Loss Weight	0.40	0.40	0.40
Image Size	518×518	518×518	256×256
Seeds (reproducibility)	0, 42, 1337	0, 42, 123	0, 42, 1337

4.3 Data Collection

4.3.1 Data Acquisition

Three datasets are employed in this work. Each of them is linked to another part of the experiment analysis. The first one is WasteRefine, a new dataset that was collected and labelled during this research study. The second and third are publicly available dataset ZeroWaste-F and SpectralWaste (RGB version). This allows us to compare the proposed method with the state-of-the-art algorithms used in the past in standardised conditions. The three datasets taken altogether may be interpreted as an assortment of waste types, imaging circumstances, and volume of datasets. This allows a comprehensive assessment of the overall generalisability of the suggested model.

WasteRefine Dataset: The new data gathered and annotated as part of this study is referred to as WasteRefine. It contains 2,213 images with four categories of foreground waste materials Soft Plastic, Rigid Plastic, Paper, and Metal and a background category. These photographs were taken in real life situations that replicated the type of waste stream experienced in recycling plants. The support of the Roboflow platform resulted in the ability to organize and annotate this data along with annotating even more precise polygons, versioning datasets and managing their quality in a grouped way. In this data set, 10,860 instances of the object exist along with soft plastic being the most common 29.8% and metal being the rarest 16.9 percent respectively. A moderation of slight class imbalance that constructs the training plan, is present. The dataset is split into training (70%, 1549 images), validation (15%, 332 images) and test (15%, 332 images) sets in each of which the distribution of the classes is the same. Some qualitative visuals are showcased in Figure 4.3

Figure 4.3: Qualitative Images of newly created WasteRefine Dataset.

ZeroWaste-F Dataset: ZeroWaste-F [7] is an open-source semantic segmentation merit of waste products classification which is presented along the ZeroWaste project. It has 4,503 annotated pixel-based pictures in four foreground categories Rigid Plastic, Soft Plastic, Metal, Cardboard and background. The dataset is the largest one among the three that have been employed in this work and it has 27,744 annotated object instances. Nonetheless, it is extremely class-imbalanced. Cardboard takes 66.8% out of the total number of objects whereas Metal takes 1.4%. This makes after-experimental prediction of the minority classes a challenging issue. The official train/validation/test split which consists of 3,002 training images (≈67%), 572 validation images (≈13%), and 929 test images (≈20%) are used here

too. Preprocessing is done to resize the images to 518×518 pixels which is the native input resolution of DINOv2.

SpectralWaste Dataset (RGB Variant): The spectralwaste [17] is a waste segmentation dataset which was initially developed as a hyperspectral imager research data. However, there is an RGB version that is utilized in this study to permit a just comparison with RGB-based baseline models such as COSNet [26] and FANet [16]. RGB variant has 852 images of 7 classes composed of background, film, basket, cardboard, video tape, filament and bag. It contains the least data set in this study and has the worst class imbalance of the three with bag class taking 46.3 percent of all the cases and filament (5.4%) and cardboard (3.3%) are very rare. The dataset is of approximately 60/20/20 train/validation/test split containing 514 training images, 167 validation images, and 171 test images. Another unique difficulty of SpectralWaste is the existence of video tape and filament classes which are visually ambiguous and have very few pixels per occurrence. Hence, they can be hard to divide into accurate segments. Images in RGB mode are all 256×256 pixels and this is kept throughout training so that they are consistent with the official benchmark evaluation protocol. The total number of dataset images and splits are summarized in Table 4.2. The splits of the data set by the class are described in Table 4.3 and Figures 4.4 and 4.5.

Table 4.2: Image and object counts with train/validation/test splits for all three datasets.

Dataset	Total Images	Total Objects	Train	Validation	Test	Class (incl. BG)
WasteRefine	2,213	10,860	1,549 (70%)	332 (15%)	332 (15%)	5
ZeroWaste-F	4,503	27,744	3,002 (≈67%)	572 (≈13%)	929 (≈20%)	5
SpectralWaste	852	2,059	514 (≈60%)	167 (≈20%)	171 (≈20%)	7

Table 4.3: Per-class object instance distribution across training, validation, and test splits. The SpectralWaste per-split object counts are not reported in the original dataset documentation.

Dataset Class	Train Objects	Val Objects	Test Objects	Total	% of Total
WasteRefine Dataset					
Soft Plastic	2,266	464	507	3,237	29.8%
Rigid Plastic	2,137	478	460	3,075	28.3%
Paper	1,881	426	401	2,708	24.9%
Metal	1,283	280	272	1,835	16.9%
ZeroWaste-F Dataset					
Cardboard	12,940	2,167	3,428	18,535	66.8%
Soft Plastic	4,862	855	1,236	6,953	25.1%
Rigid Plastic	1,160	305	315	1,780	6.4%
Metal	263	60	63	386	1.4%

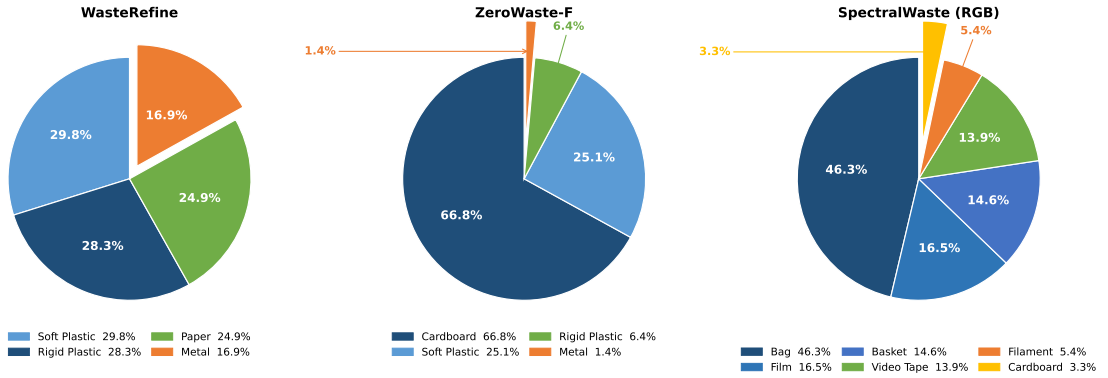


Figure 4.4: Per-class object distribution across the three datasets. Exploded slices indicate the rarest class per dataset, highlighting severe class imbalance.

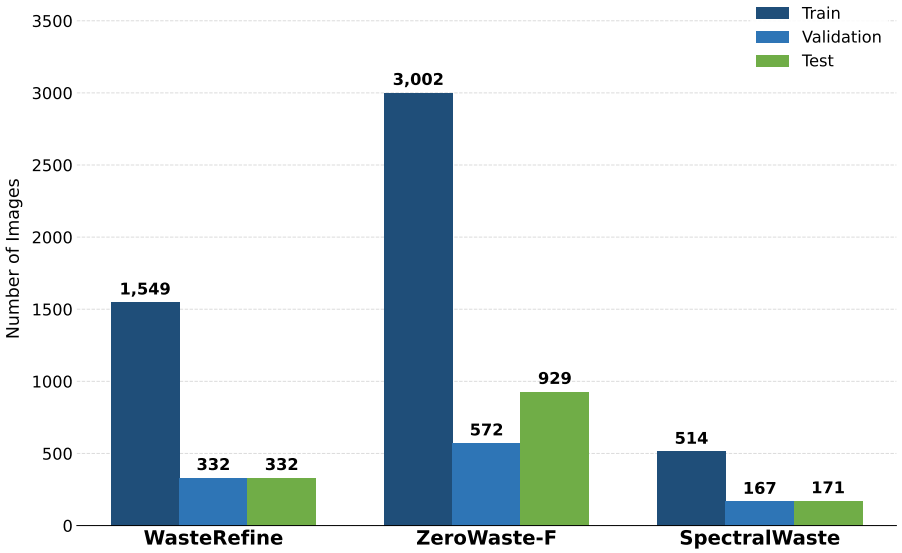


Figure 4.5: Train, validation, and test image counts per dataset. SpectralWaste contains the fewest training samples (514).

4.3.2 Data Cleaning

The process of data cleaning was applied selectively based on the type of data and its source. In the WasteRefine dataset, the most critical cleaning concern was the near-duplicate/visually close images that could lead to a data leakage between the training and validation split or a deliberate attempt to optimize training performance because of the availability of near-identical images. This was addressed with the similarity detecting potential of Roboflow which finds and labels high visual similarity imagery with the help of perceptual hash. It removed a few near-duplicates like 17 images of the data after the automated screening before splitting the data to create the final 2,213 images.

In the case of ZeroWaste-F and SpectralWaste with quality checked benchmark datasets that are described by their authors, no further data cleanup was needed besides format verification. It was ensured that all image-mask pairs were properly

aligned and that all mask pixel values fell within the appropriate range of class IDs and that corrupt files were not found in the archives downloaded. This SpectralWaste dataset also needed a mask loading step to support both palette-indexed PNG format and RGB-encoded mask formats since not all subsets of the SpectralWaste dataset used the same mask encoding. In order to address this discrepancy a powerful dual-format mask loader was applied to resolve the issue.

4.3.3 Data Transformation and Augmentation

It is important to note that data augmentation is an imperative part of the training pipeline, especially when dealing with datasets having a small number of samples, in this case, the WasteRefine (2,213 images) and the SpectralWaste (852 images). Augmentation is done to artificially enhance the diversity of the training set by flipping images and masking pixels of those images randomly. It also complex the model to learn viewpoint, lighting and partial visual obstruction invariant features. Any augmentation is done with the help of the Albumentations library that should make sure that both picture and the mask are subjected to the same transformations.

There are two different augmentation strategies that are based on the visual features and classes distributions of datasets as they differ. WasteRefine and ZeroWaste-F have an augmentation pipeline that is designed to be used in general-purpose waste segmentation. On the other hand, SpectralWaste uses a conservative pipeline to ensure that it does not destroy the fine spatial structure of thin-object classes, such as filament and video tape.

WasteRefine and ZeroWaste-F Augmentation Pipeline: WasteRefine and ZeroWaste-F use bilinear interpolation and constant padding of the boundaries. It also has geometric distortions, horizontal and vertical inversion (probability with 0.5) and random rotation ($\pm 45^\circ$ range). These geometric augmentations are based on the reality of variability of the real world with the imaging conditions of waste and may be measured at a variety of angles to the conveyor belts or bins. The Photometric augmentations incorporate ColorJitter with brightness (maximum of 0.2), contrast and saturation (maximum of 0.2) and hue (maximum of 0.1) changes (enacted with probability 0.4) in order to generate sensor noise caused by low-quality imaging hardware. CoarseDropout randomly conceals 1 to 8 rectangular areas of 4 to 16 pixels across (probability 0.2) to recreate effects of partial concealment of waste items as one might see on a bin or a conveyor belt when multiple objects are placed in the same bin or on the same conveyor. Another method used during training is Copy-Paste augmentation that produces new training examples by pasting an instance of a rare class on other images, with a probability of 0.30 in WasteRefine and 0.25 in ZeroWaste. It is specifically focused on rare classes such as Metal to reduce the impact of class imbalances.

SpectralWaste Augmentation Pipeline: The SpectralWaste augmentation techniques is purposely more conservative than the one employed with WasteRefine and ZeroWaste-F. The most significant change is that the range in rotation is reduced by half to $\pm 10^\circ$ (down from $\pm 45^\circ$) and the application possibility of colorJitter has been reduced to 0.12 in brightness, contrast and saturation. The uncertainty of Gaussian noise has been lowered by half to 0.05. Both CoarseDropout and Copy-

Paste augmentation isn't applied here in SpectralWaste. The argument behind this conservative approach is that the filament classes and video tape classes are made up of structures that are just a few pixels wide. These thin structures can drift to border regions, padded off or may be cropped off through ruthless rotation of 45°. The course dropout of 4-16 pixels can be whole rectangles of pixels covering the tiny set of the filament pixels that can be present in an image may provide a bad and inaccurate training signal to the model in that class.

Just as is true, in case of aggressive colour jitter on a dataset where certain classes are differentiated by colour (e.g. the green of filament, the pink-mauve of video tape), where some classes are differentiated by colour can run the risk of making class-discriminative colour cues uninformative. In the case of few and thin classes, the conservative augmentation method trades effectiveness of the dataset against the integrity of the training signal. The entire augmentation specification for each dataset is summarised in Table 4.4.

Table 4.4: Augmentation strategies for training splits. WasteRefine and ZeroWaste-F share an identical augmentation pipeline. SpectralWaste uses different technique to preserve thin-object class integrity.

Augmentation	**WasteRefine & ZeroWaste-F**	**SpectralWaste**
Input Resize	518×518	256×256
Horizontal Flip	$p = 0.5$	$p = 0.5$
Vertical Flip	$p = 0.5$	$p = 0.5$
Rotation	$\pm 45°$, $p = 0.5$	$\pm 10°$, $p = 0.3$ (reduced)
Colour Jitter (B/C/S)	0.2/0.2/0.2, $p = 0.4$	0.12/0.12/0.12, $p = 0.35$
Hue Jitter	0.1, $p = 0.4$	0.06, $p = 0.35$
Gaussian Noise	$p = 0.1$	$p = 0.05$ (minimal)
CoarseDropout	1–8 holes, 4–16 px, $p = 0.2$	Not applied
Copy-Paste Augmentation	Yes (prob. varies per dataset)	Not applied
Normalisation Mean	(0.485, 0.456, 0.406)	(0.485, 0.456, 0.406)
Normalisation Std	(0.229, 0.224, 0.225)	(0.229, 0.224, 0.225)

4.3.4 Data Splitting and Sampling

The dataset is splitted in 3 parts: train, validation and test. In WasteRefine, it is 70/15/15. With ZeroWaste-F and SpectralWaste, formal splits of the public benchmark are used to permit a fair and directly comparable performance to be compared with published results. Whichever is the case, the test set does not play any role whatsoever in the training. Hyperparameter selection process and can be reported only when stating the resulting performance.

To overcome the issue of class imbalance in training, a Pixel-Aware Weighted Random Sampler is used on all training dataloaders. Each of the training images is given a sampling weight by the sampler based on the representation of the rare and important classes in that image. In each image, the mask is scanned to estimate the fraction of pixels occupied by each of the classes and a weighted average of the scores of class importance depending on the square root of the pixel fraction occupied by the class in each image is calculated. The square root transformation is applied to put a curb on the prevalence of the most prevalent class (background) and to

reward proportionately images that comprise the rare classes of the foreground. In the case of SpectralWaste, presence bonuses are additional to images with pixels of filament or video tape. This is done so that such rarer classes are given high sampling priority, rather than the high relative sampling priority that they would have been awarded by their pixel fraction alone. The Weighted Random Sampler applies the replacement mechanism to sample images at every training epoch. This ensures biasness toward the images of the rare-class and against the images of the common-class, but with no data modification or synthetic samples.

4.3.5 Data Reduction

The explicit reduction of data samples was not applied to this study because the sample sizes in all three datasets were already low. The limitation on the training examples would have additionally limited the generalisation of the model to the variety of the appearance of waste in the real world. But in the case of the development of the WasteRefine dataset, visually similar and near-duplicated images were filtered using the Roboflow platform and about 17 images were removed from the initial set. This filtering was done to guarantee the quality of data, not for any computation efficiency reason. The remaining images were not downsampled on purpose since even the biggest dataset (ZeroWaste-F with 4,503 images) is small by deep learning standards.

Image resizing is one of the types of implicit data reductions that was used. WasteRefine and ZeroWaste-F have all images preprocessed to 518×518 pixels which is the native input size suggested by the DINOv2 backbone which operates 14×14 pixel patches and produces a 37×37 positional feature grid at this resolution. SpectralWaste images are stored in the original resolution of 256×256 pixels that is retained so as not to be unfair to the original benchmark evaluation protocol. Although the scaling of bigger images decreases the positional resolution of the original image, it had to be done in order to achieve uniform batch scaling and memory management used by the GPU. However, the final resolutions are not so small as to lose enough detail to use them in dense prediction. Instead of taking away data, the research lays a lot of stress on data augmentation as the chief process of spreading the scope of effective datasets and overfitting, as outlined in Section 4.3.3.

4.3.6 Summary of Preprocessed Data

Table 4.5 represents a detailed summary of the preprocessed dataset layout. It contains the final image counts per split, input resolution, number of classes, and sampling strategy employed during training.

Table 4.5: Summary of preprocessed data configuration for all three datasets.

Dataset	**Train**	**Val**	**Test**	**Resolution**	**Classes**	**Sampler**
WasteRefine (Ours)	1,549	332	332	518^2	5	Weighted Random
ZeroWaste-F	3,002	572	929	518^2	5	Weighted Random
SpectralWaste (RGB)	514	167	171	256^2	7	Weighted Random

4.4 Implementation of Selected Design

This section implements the proposed WasteRefine architecture that is applied to the aforementioned three datasets. Even though the model architecture is predetermined, any dataset should be trained with a selected training strategy that takes into consideration both its data distribution and class imbalance profile and segmentation challenging. The experiments are executed with PyTorch and loaded with HuggingFace Transformers library to load the backbones of DINOv2 and Albumentations to augment them. The training proceeds through a single-gpu windows based platform and all the codes are written to permit the ViT-S as well as ViT-B variants of the backbone on the single configuration interface.

4.4.1 Loss Function

A comprehensive and composite depiction of Focal and Dice loss is employed for training purpose of the segmentation model addressing class imbalancement and improving segmentation accuracy. The overall segmentation loss is presented below:

$$L_{\text{main/aux}} = 0.5\, L_{\text{Dice}} + 0.5\, L_{\text{Focal}} \tag{4.1}$$

The Dice loss is presented as:

$$L_{\text{Dice}} = 1 - \frac{2\sum(P \cdot G) + \epsilon}{\sum P + \sum G + \epsilon} \tag{4.2}$$

where P denotes the predicted probability map, G denotes the ground-truth mask, and $\epsilon = 10^{-6}$ is a smoothing constant for ensuring numerical stability.

The Focal loss is defined as:

$$L_{\text{Focal}} = \alpha(1 - p_t)^{\gamma} \cdot CE \tag{4.3}$$

where p_t is the true class predicted probability, $CE = -\log(p_t)$ is the cross-entropy loss, $\alpha = 0.25$ is the class balancing factor, and $\gamma = 2.0$ is the focusing parameter that reduces the contribution of easy samples.

The Binary Cross-Entropy (BCE) loss is for L_{edge} which is defined as:

$$L_{\text{BCE}} = -\left[y \log(p_t) + (1 - y) \log(1 - p_t)\right] \tag{4.4}$$

where p_t is the true class predicted probability and y denotes ground truth label.

Auxiliary supervision and edge-aware loss were introduced for improvement of boundary precision and stability of learning. The overall training loss is defined as:

$$L_{\text{total}} = L_{\text{main}} + \lambda_{\text{aux}} L_{\text{aux}} + \lambda_{\text{edge}} L_{\text{edge}} \tag{4.5}$$

where $\lambda_{\text{aux}} = 0.4$ contributes to auxiliary supervision, and $\lambda_{\text{edge}} = 0.2/0.15$ regulates the influence of the boundary-aware loss. The edge loss is implemented using binary cross-entropy between predicted and ground-truth boundaries.

4.4.2 Optimization Strategy

AdamW optimizer along with regularization technique weight decay is utilized for optimization of model. Following the memory constraints issue of GPU, employment of gradient accumulation with batch size is done for generic effective batch size:

$$\text{Effective Batch Size} = \text{Batch Size} \times \text{Accumulation Steps} \tag{4.6}$$

Batch size of 2 and 16 accumulation steps were utilized to ensure effective batch size of 32.

Weight decay:

$$L_{\text{reg}} = L + \lambda \|\theta\|^2 \tag{4.7}$$

where $\lambda = 10^{-4}$ denotes weight decay coefficient and θ denotes model parameters.

Backbone learning rate:

$$LR_{\text{backbone}} = \eta \cdot LR_{\text{base}} \tag{4.8}$$

where $LR_{\text{base}} = 10^{-4}$ and $\eta = 0.1$.

Cosine Annealing:

$$LR_t = LR_{\min} + \frac{1}{2}(LR_{\max} - LR_{\min}) \left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{4.9}$$

where $LR_{\min} = 10^{-6}$, $LR_{\max} = 10^{-4}$, t denotes current epoch and T denotes total epochs.

To improve training stability and generalization, Exponential Moving Average (EMA) of model parameters was applied:

$$\theta_{\text{EMA}} = \beta\theta_{\text{EMA}} + (1 - \beta)\theta \tag{4.10}$$

where $\beta = 0.999$ is refered to decay coefficient, θ denotes the current model parameters, and θ_{EMA} represents the exponentially averaged parameters.

4.4.3 Training Strategy and Selection

All three datasets are trained at a single code level and the behaviour specific to the dataset is entirely managed in configuration files. The general plan is to use a two-stage curriculum as outlined in Section 4.2.2 where the backbone is frozen during the warmup stage and then the end-to-end fine-tuning. The total loss in training goes as:

$$L_{\text{total}} = L_{\text{main}} + \lambda_{\text{aux}} L_{\text{aux}} + \lambda_{\text{edge}} L_{\text{edge}} \tag{4.11}$$

where the primary segmentation output loss is the sum of the Dice loss and Focal loss. The auxiliary segmentation output loss is the same sum of loss with a weight of $\lambda_{\text{aux}} = 0.4$ and the edge loss is the Binary Cross-Entropy loss on the BoundaryBranch edge map output where $\lambda_{\text{edge}} = 0.2/0.15$ is used as a dataset specific weight. The

Dice loss directly takes part in the optimization of the intersection between predicted and the ground truth segmentation masks which can be quite useful when using imbalanced datasets since it balances all the classes by their frequency. The Focal loss gives less significance to the easily classified background pixels and concentrates training on hard and rare class instances. The edge loss gives explicit gradient signals which are aimed at the precision of the boundaries.

WasteRefine Training

As the original contribution to this work, WasteRefine can be the most sensitive to hyperaparameter tuning. The training will terminate on 100 epochs or 15 epochs for early stopping if there is no improvement in mIoU score. The warming up phase is 5 epochs. The edge loss weight is set to 0.20. To offset the low size of the dataset (2,213 images vs 4,503), the probability of augmentation with Copypasting will be 0.30, a slight bit higher than ZeroWaste-F. The largest target of the Copy-Paste is the most uncommon form of metal at 16.9%, which is oversampled at the weighted sampler. The edge loss weight is set to 0.20.

Despite the relatively small size of the dataset, the WasteRefine model produces exceptionally good performance which can be observed in Table 4.6 and Figure 4.6. The ViT-B model obtains mean mIoU of $96.64 \pm 0.16\%$ and $96.26 \pm 0.07\%$ for three seeds. The very low values of the standard deviations across seeds (0.16% and 0.07% respectively) show that the training procedure is very stable and reproducible for this set of data. The high performance achieved is not due to favourably chosen random initialisation values, but is a real capability of the model. Notably, the ViT-S model with less total parameters (25.16M) achieves 96.26%, which is higher than all evaluated baselines including DINOv2-Base with Simple DPT (94.31%, 88.25M). It also beats ResNet-34 + UNet (92.37%, 24.44M) and SegFormer-B2 (72.92%, 27.35M). This demonstrates that the efficiency of the custom decoder enables strong performance with the smaller backbone.

ZeroWaste-F Training

The biggest data set used in this paper is ZeroWaste-F, consisting of 4,503 images. However, it also has the highest imbalance of classes where the Metal class makes up only 1.4% of all annotated samples. The training is set to 50 epochs and early stopping patience of 20. The warming up phase is 5 epochs. The likelihood of Copy-Paste is 0.25 and the weighted sampler will up-weight intensively with images that contain Metal instances. It has a 518×518 image resolution which is set to match the recommended DINOv2 input size.

ViT-B model attains a mean mIoU of $61.94 \pm 0.84\%$ with seed 0, 42, as well as seed 123. The ViT-S model achieves $59.75 \pm 0.75\%$. Both models outperform every published baseline on this benchmark. The highest previous score was COSNet [26] at 56.67% (37M parameters) which means that our ViT-B is 5.27 percentage points better. Most importantly, Our ViT-S is 3.08 percentage points better and uses 32 per cent of the parameters that COSNet used. This proves that the suggested decoder can be used in sync with the global attention mechanism of the DINOv2 backbone that significantly improves the state of the art on ZeroWaste-F. The increased vari-

ance between seeds (±0.84% for ViT-B, ±0.75% for ViT-S) than WasteRefine is due to the increased difficulty of ZeroWaste-F, especially due to the total lack of Metal objects.

SpectralWaste (RGB Variant) Training

The most difficult dataset in this paper is that of SpectralWaste because it has a small size (852 images). Not only that, it has imbalanced classes and contains thin-object classes (filament and video tape) that are visually ambiguous. The training is also set at a maximum of 300 or 500 epochs, much higher than the other datasets. The early stopping patience has been set at 35 epochs, indicating slower and less predictable convergence behaviour with learning on small samples with extreme class imbalance. Warmup is also increased to 10 epochs to give the decoder more time to stabilise before the backbone fine-tuning is commenced.

The copy-paste augmentation is disabled on SpectralWaste which is explained in Section 4.3.3 because it has the risk of ruining the structural integrity of the thin-object classes. The weighted sampler is used to give filament and video tape image high presence bonuses. The edge loss weight is also scaled down to 0.15 (compared to 0.20 in other datasets) to prevent an over-penalisation of prediction mistakes on the boundary. It is done for the extremely rare filament class, where ground truth edge maps are very sparse and noisy. The assessed metric is Foreground mIoU (FG mIoU) where the background is not included in the mean calculation. This ensures following the SpectralWaste benchmark reporting standard as published before [17].

ViT-B model has a mean FG mIoU of 70.73±0.10% on seeds 0, 42, 1337. This is the highest reported value on the SpectralWaste RGB benchmark. This exceeds COSNet [26] of the previous best result of 69.96% FG mIoU, DINOv2-Base of 69.71% FG mIoU, FANet [16] 67.83% FG mIoU, SegFormer-B0 48.4% FG mIoU and MiniNet-v2 at 44.5%. With 25.16M parameters, the ViT-S model attains 68.74 ± 0.32% FG mIoU, which is 32 times smaller than the 37M of COSNet, yet it does not exceed 1.22 percentage points of COSNet. It is interesting to note that the ViT-B standard deviation of ±0.10% is lowest of all the three datasets. This indicates that although the dataset is challenging to train, the model may be trained in a highly consistent and reproducible manner. The video tape class is the hardest of all seeds and both backbone variants. As it has a very thin structure and resembles some plastic film materials in appearance, this results in achieving about 36-40% IoU.

Table 4.6: Reproducibility results across three random seeds per dataset.

Dataset	Model	Seed A	Seed B	Seed C	Mean ± Std (%)
WasteRefine (seeds: 0, 42, 1337) - mIoU					
WasteRefine	ViT-B	96.80	96.69	96.43	96.64 ± 0.16
WasteRefine	ViT-S	96.36	96.19	96.23	96.26 ± 0.07
SpectralWaste (seeds: 0, 42, 1337) - FG mIoU					
SpectralWaste	ViT-B	70.86	70.69	70.64	70.73 ± 0.10
SpectralWaste	ViT-S	68.29	69.00	68.94	68.74 ± 0.32
ZeroWaste-F (seeds: 0, 42, 123) - mIoU					
ZeroWaste-F	ViT-B	61.97	62.95	60.89	61.94 ± 0.84
ZeroWaste-F	ViT-S	60.33	60.19	58.74	59.75 ± 0.75

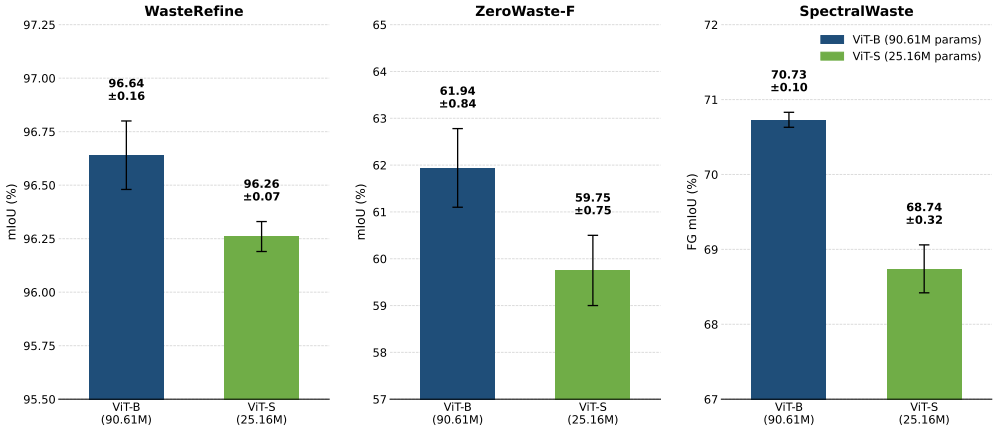


Figure 4.6: Mean performance with ± standard deviation error bars across three random seeds for ViT-B and ViT-S on all three datasets.

Chapter 5

Result Analysis

5.1 Performance Evaluation

This chapter provides a detailed analysis of the experimental outcomes obtained through the proposed WasteRefine framework for image segmentation. In this context, the experimental evaluation of the proposed method is conducted on the three datasets. Additionally, the evaluation of the proposed method is conducted through a series of experiments, namely per-class evaluation, multi-metric evaluation, design iteration history, statistical reproducibility evaluation, and benchmarking. In this context, it is important to note that the results reported in this chapter correspond to the benchmarking experiment unless otherwise mentioned. Moreover, the per-class evaluation of the proposed method is reported for the seed 42, which is considered the primary seed for evaluation purposes. Furthermore, the mean ± standard deviation of the proposed method is reported for the overall reproducibility evaluation.

5.1.1 Evaluation Metrics and Mathematical Formulation

Model performance is assessed using several metrics, each targeting a different aspect of segmentation quality. No single metric tells the whole story in multi-class problems; each one exposes different failure patterns. The metrics used throughout this work are defined below.

Intersection over Union (IoU): The per-class IoU, also known as the Jaccard Index, measures the overlap between the predicted segmentation mask and the ground truth mask for each individual class. It is defined as:

$$\mathrm{IoU}_c = \frac{|\mathrm{TP}_c|}{|\mathrm{TP}_c| + |\mathrm{FP}_c| + |\mathrm{FN}_c|} \tag{5.1}$$

Here TP_c, FP_c, and FN_c are the true positives, false positives, and false negatives for class c. IoU runs from 0 (no overlap) to 1 (perfect match) and penalises over-segmentation and under-segmentation equally.

Mean Intersection over Union (mIoU): The mean IoU is computed as the arithmetic mean of per-class IoU values across all C classes including background:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c \tag{5.2}$$

mIoU serves as the main benchmark metric for the WasteRefine and ZeroWaste-F evaluations.

Foreground mIoU (FG mIoU): For the SpectralWaste dataset similar to the original benchmark reporting convention, FG mIoU is used. This means that the background class will be excluded from the mean calculation.

$$\text{FG mIoU} = \frac{1}{C-1} \sum_{c=1}^{C-1} \text{IoU}_c \tag{5.3}$$

FG mIoU is more sensitive to the foreground classes. This measure of mIoU is recommended when the background class dominates the pixel count. This reporting of FG mIoU will prevent the overall score from being artificially inflated.

Dice Coefficient: The Dice coefficient (also known as the F1 score for binary segmentation) measures the harmonic mean of precision and recall at the pixel level for each class:

$$\text{Dice}_c = \frac{2 \times |\text{TP}_c|}{2 \times |\text{TP}_c| + |\text{FP}_c| + |\text{FN}_c|} \tag{5.4}$$

Dice is more sensitive to the overlap quality than IoU and tends to award higher scores for partially correct predictions. The macro Dice averages per-class Dice scores across all classes equally.

Precision: Pixel-level precision measures the proportion of predicted positive pixels that are truly positive for each class:

$$\text{Precision}_c = \frac{|\text{TP}_c|}{|\text{TP}_c| + |\text{FP}_c|} \tag{5.5}$$

High precision indicates that when the model predicts a class, it is usually correct. Low precision indicates over-prediction or hallucination of a class. Macro precision averages per-class precision across all classes.

Pixel Accuracy (Macro): Macro pixel accuracy reports the per-class pixel accuracy averaged equally across all classes:

$$\text{PixAcc}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{5.6}$$

This is equivalent to macro recall and treats each class equally regardless of its pixel count, making it more informative than global pixel accuracy for imbalanced datasets.

Pixel Accuracy (Micro): Micro pixel accuracy is defined as the ratio of correctly classified pixels to the total number of pixels in the entire dataset, irrespective of classes:

$$\text{PixAcc}_{\text{micro}} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FP}_c + \text{FN}_c)} \tag{5.7}$$

Due to the micro-pixel accuracy being controlled by the most frequent class, it remains high even when infrequent classes are not considered. It is mainly included for completeness and to allow cross-study comparison.

5.1.2 Segmentation Result Analysis

The following subsections gives an analysis of each class as well as overall segmentation performance on all three datasets for both ViT-B and ViT-S variants. The results reported in this section are for evaluation on a held-out test set using a seed value of 42.

WasteRefine Dataset

The WasteRefine dataset is the contribution of this paper, with a total of 2,213 images spread across four classes of foreground waste and a background class. Both models performed well on the task with ViT-B achieving 96.69% mIoU and ViT-S achieving 96.19%. This is much better than all the baselines. Performance of all classes are in Table 5.1. There are also some visual comparisons in Figure 5.1 and 5.2.

Table 5.1: Per-class segmentation results on the WasteRefine test set (seed 42).

Class	**ViT-B (90.61M)**				**ViT-S (25.16M)**			
	IoU	**Dice**	**Prec.**	**PixAcc**	**IoU**	**Dice**	**Prec.**	**PixAcc**
Background	0.9805	0.9902	0.9915	0.9888	0.9784	0.9891	0.9916	0.9865
Metal	0.9641	0.9817	0.9822	0.9812	0.9589	0.9790	0.9770	0.9811
Paper	0.9695	0.9845	0.9823	0.9867	0.9650	0.9822	0.9789	0.9855
Rigid Plastic	0.9529	0.9759	0.9702	0.9817	0.9461	0.9723	0.9671	0.9775
Soft Plastic	0.9675	0.9835	0.9840	0.9830	0.9609	0.9801	0.9774	0.9828
Overall (macro)	0.9669	0.9832	0.9820	0.9843	0.9619	0.9805	0.9784	0.9846
Pixel Acc (micro)	—	—	—	0.9866	—	—	—	0.9866

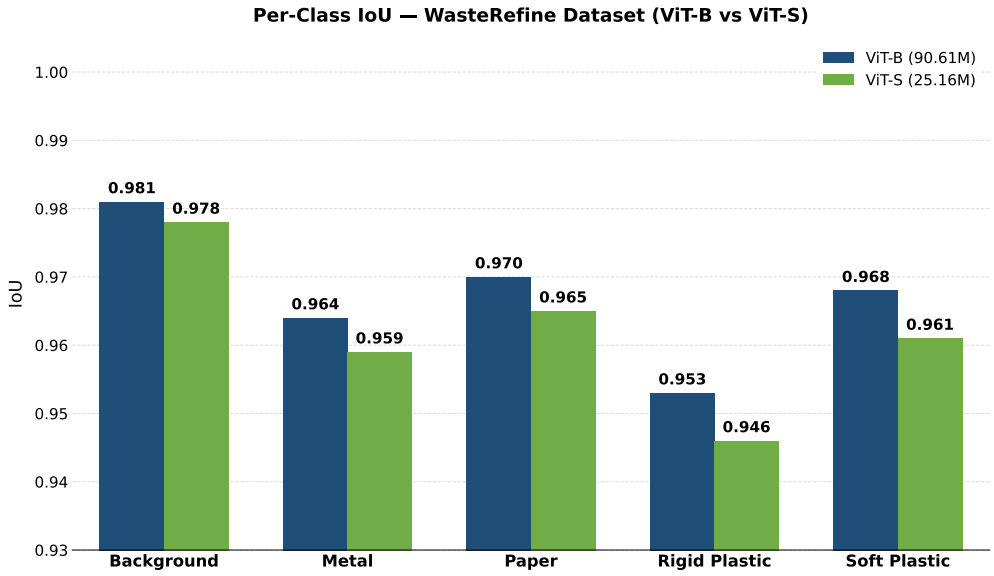


Figure 5.1: Per-class IoU comparison between ViT-B and ViT-S on the WasteRefine dataset.

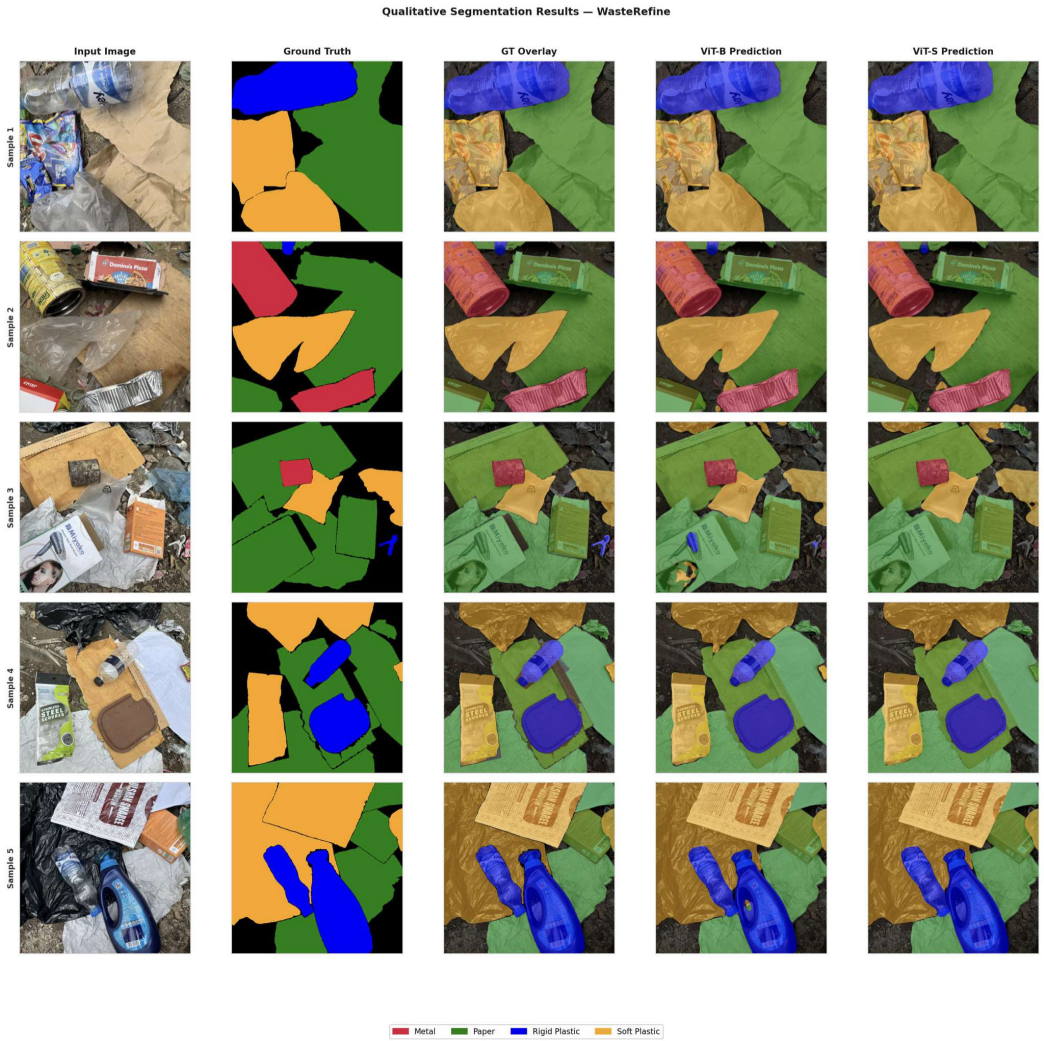


Figure 5.2: Qualitative segmentation outputs on the WasteRefine dataset.

Performance was remarkably uniform across all five classes. Even the hardest class, Rigid Plastic reached 0.9529 IoU on ViT-B and 0.9461 on ViT-S which are still well above any baseline. This consistency likely reflects WasteRefine's relatively balanced class distribution as Metal, which is the rarest class at 16.9%, appears frequently enough that the model receives a solid learning signal for it. Metal's IoU of 0.9641 on ViT-B is noteworthy, especially as it is the least occurring class. This is the result of the Pixel Aware Weighted Sampler + Copy-Paste augmentation strategy used at a probability of 0.30. The difference between ViT-B and ViT-S models was small throughout - a maximum difference of 0.0052 on Metal and 0.0068 on Rigid Plastic. This suggests that the custom decoder is able to glean sufficient discriminative information from the lighter model. Micro pixel accuracy was 0.9866 for all models. This suggests that pixel-level classification does not depend on the size of the model.

Rigid Plastic has the lowest IoU among all variants. Rigid Plastic and Soft Plastic have visual overlap since they may be mistaken for each other as they are both transparent containers or bottles. Soft Plastic achieves a higher IoU value (0.9675 ViT-B) than Rigid Plastic (0.9529 ViT-B) under similar conditions. This indicates that DINOv2 excels at capturing distinguishable Soft Plastic textures probably because of their crumpled nature.

ZeroWaste-F Dataset

The ZeroWaste-F dataset presented the biggest challenge with respect to class imbalance. Metal has just 1.4% of the total instances which is roughly 50 times fewer than the instances of Cardboard. ViT-B achieved 62.95% mIoU overall and ViT-S reached 60.19%. Breakdowns of each class are in Table 5.2 and the qualitative comparisons are in Figure 5.3 and 5.4.

Table 5.2: Per-class segmentation results on the ZeroWaste-F test set (seed 42).

Class	**ViT-B (90.61M)**				**ViT-S (25.16M)**			
	IoU	**Dice**	**Prec.**	**PixAcc**	**IoU**	**Dice**	**Prec.**	**PixAcc**
Background	0.9212	0.9590	0.9592	0.9588	0.9139	0.9550	0.9582	0.9519
Rigid Plastic	0.4928	0.6603	0.7178	0.6112	0.4870	0.6550	0.6436	0.6668
Soft Plastic	0.6249	0.7691	0.7516	0.7875	0.6069	0.7553	0.7378	0.7737
Metal	0.4150	0.5866	0.6120	0.5632	0.3499	0.5185	0.5112	0.5259
Cardboard	0.6935	0.8190	0.8481	0.7919	0.6516	0.7890	0.7922	0.7859
Overall (macro)	0.6295	0.7588	0.7777	0.7425	0.6019	0.7346	0.7286	0.7408
Pixel Acc (micro)	—	—	—	0.9267	—	—	—	0.9196

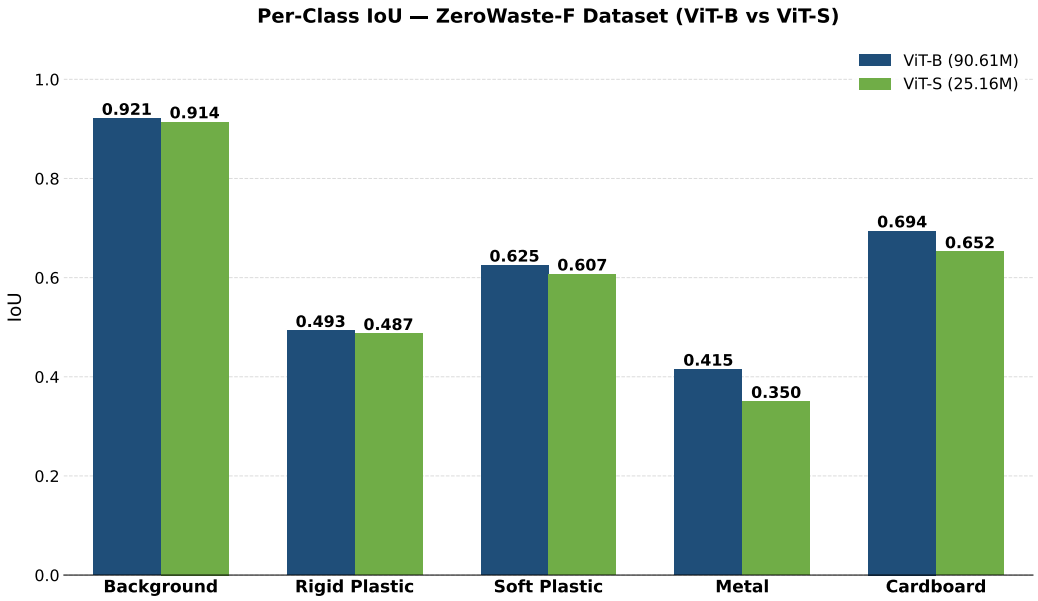


Figure 5.3: Per-class IoU comparison between ViT-B and ViT-S on the ZeroWaste-F dataset.

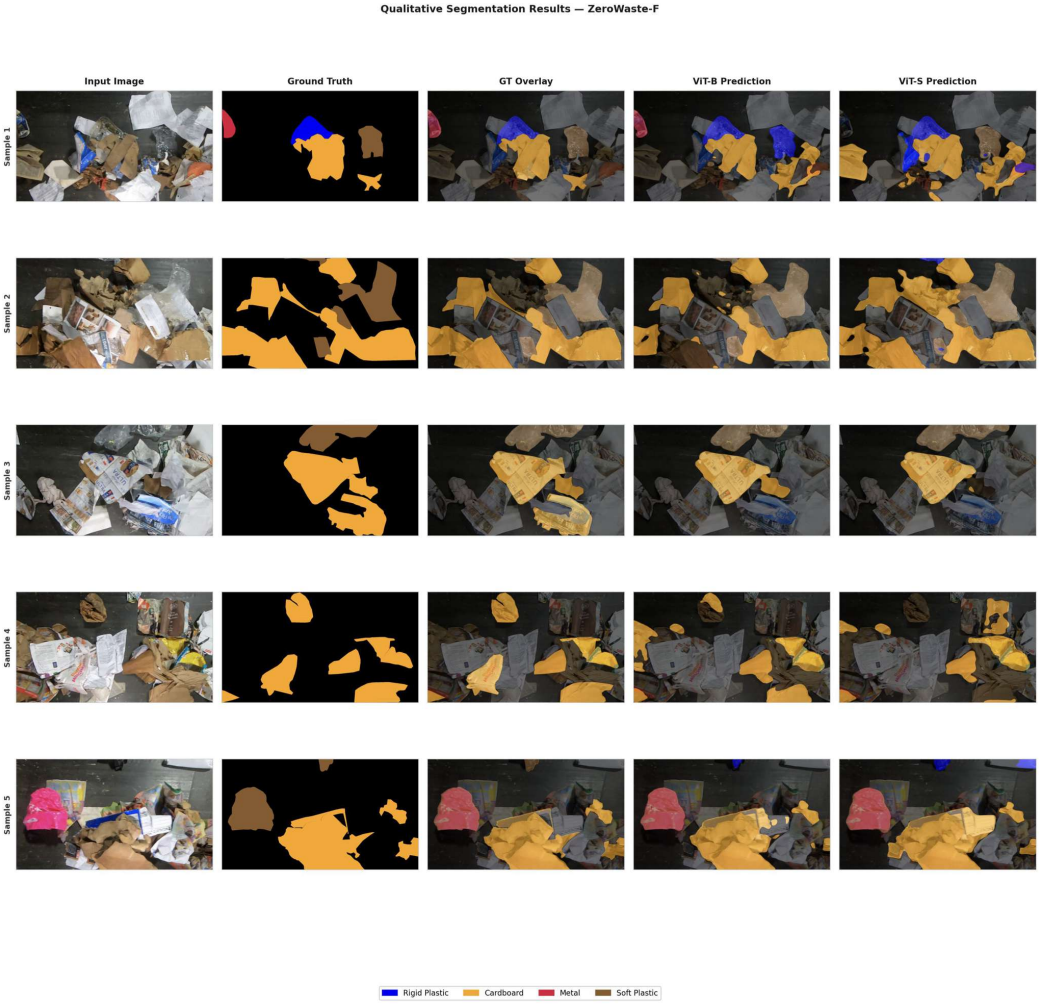


Figure 5.4: Qualitative segmentation outputs on the ZeroWaste-F dataset.

Performance tracked class frequency closely. Cardboard (66.8% of instances) came out on top at 0.6935 IoU on ViT-B and 0.6516 on ViT-S. Soft Plastic (25.1%) followed at 0.6249 and 0.6069 and Rigid Plastic (6.4%) reached 0.4928 and 0.4870. Metal is the rarest class at 1.4% and scored lowest. 0.4150 on ViT-B and just 0.3499 on ViT-S. That 0.0651 gap between backbones is the largest ViT-B advantage found across any class and dataset in the entire study. The reason why ViT-B performs better than ViT-S on Metal is the dimensionality of the feature space which is 768 compared to 384 for ViT-S. Metal is a genuinely hard class to recognize because crumpled foil, tin cans, bottle caps, etc., can look quite similar to each other and the ability to tell them apart from plastic or background is simply not as good with the 384 space of ViT-S. In summary, when Metal detection is a priority such as in ferrous metal recovery pipelines then ViT-B is the way to go even if it means more parameters.

The background IoU of 0.9212 stands in sharp contrast to the foreground scores - expected given ZeroWaste-F's severe imbalance. The micro pixel accuracy of 0.9267 is elevated by background dominance, while the macro pixel accuracy of 0.7425 more honestly reflects the model's performance across all classes. Both ViT-B and ViT-S comfortably surpass all published baselines on this dataset, ViT-B by 5.27 percentage points over COSNet (56.67%) and ViT-S by 3.08 percentage points, establishing new state-of-the-art results on ZeroWaste-F.

SpectralWaste Dataset

SpectralWaste presents the most complex per-class analysis due to the presence of two structurally challenging classes, Video Tape and Filament which consist of thin elongated structures that occupy very few pixels per instance. The per-class results for both ViT-B and ViT-S are presented in Table 5.3 and Figure 5.5 and qualitative images in Figure 5.6. Performance is reported using FG mIoU to maintain consistency with the original benchmark evaluation protocol.

Table 5.3: Per-class segmentation results on the SpectralWaste test set (seed 42). FG mIoU excludes the background class.

Class	**ViT-B (90.61M)**				**ViT-S (25.16M)**			
	IoU	**Dice**	**Prec.**	**PixAcc**	**IoU**	**Dice**	**Prec.**	**PixAcc**
Background	0.9385	0.9683	0.9644	0.9722	0.9304	0.9640	0.9668	0.9612
Film	0.7953	0.8860	0.8894	0.8825	0.7888	0.8819	0.8890	0.8750
Basket	0.8211	0.9017	0.9223	0.8821	0.8167	0.8991	0.9045	0.8938
Cardboard	0.8502	0.9190	0.9411	0.8980	0.8261	0.9047	0.9596	0.8559
Video Tape	0.3804	0.5511	0.5659	0.5371	0.3655	0.5353	0.5228	0.5484
Filament	0.6773	0.8076	0.8029	0.8123	0.6397	0.7803	0.7456	0.8183
Bag	0.7171	0.8353	0.8564	0.8151	0.7035	0.8259	0.7939	0.8606
Overall (macro)	0.7400	0.8384	0.8489	0.8284	0.7244	0.8273	0.8260	0.8305
FG mIoU only	0.7069	0.8167	0.8297	0.8045	0.6900	0.8045	0.8026	0.8087
Pixel Acc (micro)	—	—	—	0.9460	—	—	—	0.9403

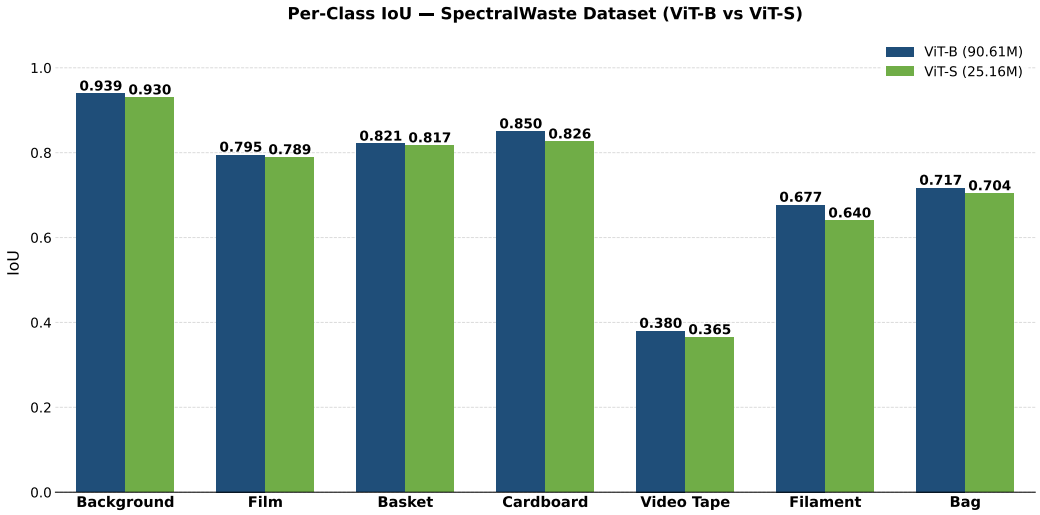


Figure 5.5: Per-class IoU comparison between ViT-B and ViT-S on the SpectralWaste dataset.

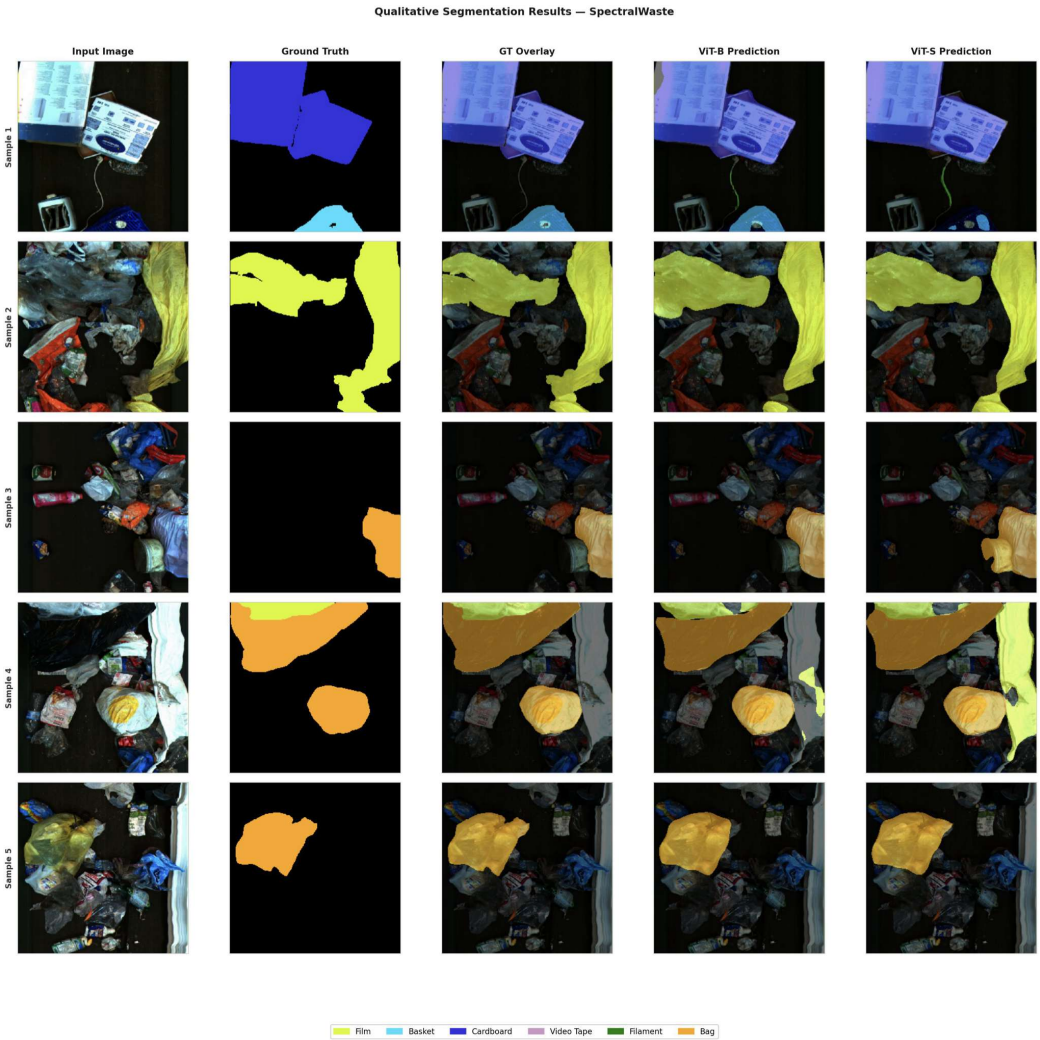


Figure 5.6: Qualitative segmentation outputs on the SpectralWaste (RGB) dataset.

Cardboard achieves the highest foreground IoU at 0.8502 (ViT-B), followed by Basket at 0.8211, Film at 0.7953, and Bag at 0.7171. These classes have relatively distinct visual signatures. The texture of Cardboard is a brown/beige corrugated pattern, Basket has a rigid structured pattern and Film has a reflective surface which the global attention mechanism of DINOv2 can successfully identify. The model Filament attains a fair IoU score of 0.6773, even in extreme scarcity (5.4%), which proves the effectiveness of BoundaryBranch and its BCE edge loss, as it focuses on retaining spatial details in the fine features F1 and F2 which can be lost in the aggregation process.

Video Tape stood out as the most failed case in this study with a mere 0.3804 IoU on ViT-B and 0.3655 on ViT-S. The class is represented by thin pink mauve strips (RGB 202, 152, 195) which may easily be mistaken for plastic film in different lighting conditions. Precision on ViT-B is 0.5659 which means if the model does flag some Video Tape pixels, it is most likely to be true. The real problem is with the recall where the model is missing large chunks of this class. The fact that there are only 287 instances in 514 training images combined with the visual ambiguity means there is a ceiling to what any model can achieve in this class.

Overall, ViT-B reached 70.69% FG mIoU and ViT-S reached 68.74%. Both clear every published baseline except COSNet (69.96%), which ViT-S trails by 1.22 points; ViT-B edges past COSNet by 0.73 points. Figure 5.7 shows a radar chart comparing all five metrics across both backbones and all three datasets.

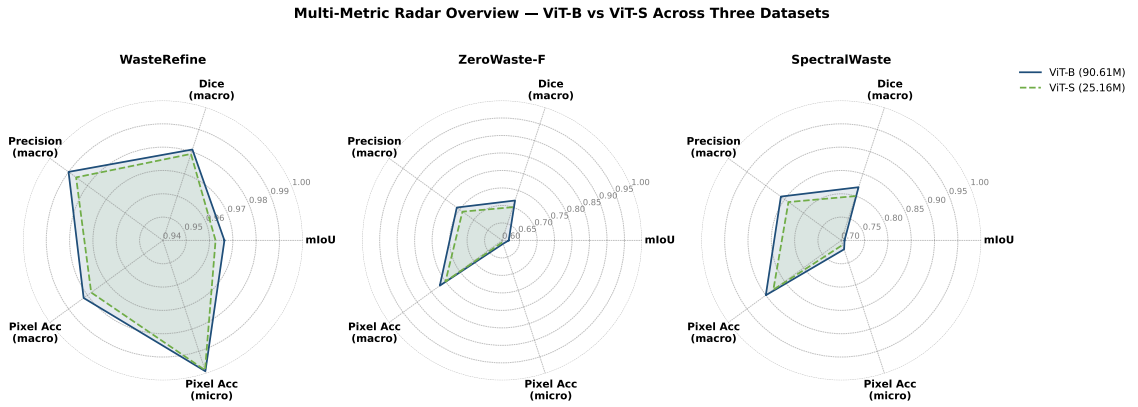


Figure 5.7: Multi-metric radar chart comparing ViT-B (solid) and ViT-S (dashed) across all three datasets.

5.2 Analysis of Design Solutions

Before arriving at the final WasteRefine architecture, sixteen experimental configurations were tested systematically. All experiments in this phase used the DINOv2-Base encoder trained on ZeroWaste-F only, which was the primary testbed throughout development, partly because its extreme class imbalance made it the most demanding case, and partly because it was the dataset the research was initially built around. The complete design iteration results are summarised in Table 5.4 and visualised in Figure 5.8.

Table 5.4: Complete design iteration history on ZeroWaste-F. The final proposed model achieves 61.94% mIoU, surpassing the Simple DPT baseline by 1.94 percentage points.

Group	**Configuration**	**Key Modification**	**mIoU (%)**
1	DINOv2-Base + Simple DPT Decoder	Standard encoder-decoder. Frozen DINOv2 backbone with a linear DPT neck.	60.00
2	DINOv2-Base + DAPT + DPT Neck	Encoder fine-tuned on unlabelled waste images prior to supervised training.	56.00
	DINOv2-Base + DAPT + GroupNorm + Pyramid Rescaling	GroupNorm added for batch size 1 stability; pyramid rescaling for multi-scale context.	56.00
	DINOv2-Base + DAPT + Heavy Loss Weighting	1.5× penalty applied to Metal and Soft Plastic classes within the loss function.	55.00
	DINOv2-Base + DAPT + EMA Stabilisation	Exponential Moving Average applied to stabilise DAPT training at batch size 1.	56.19
	DINOv2-Base + DAPT + Mosaic Augmentation	Mosaic stitching of four images combined with domain-adapted encoder weights.	56.92
3	DINOv2-Base + Progressive LR Disparity	10× learning rate gap between backbone and decoder to preserve pre-trained features.	60.00
	DINOv2-Base + Deep Supervision	Auxiliary segmentation heads added at intermediate decoder stages.	60.00
	DINOv2-Base + 53× Metal Class Weighting	Extreme pixel-level loss reweighting applied to the rarest class (Metal).	56.00
	DINOv2-Base + Full Data-Centric Augmentation	Aggressive oversampling combined with high-probability Copy-Paste augmentation.	49.00
4	DINOv2-Base + Deformable Neck + Tversky Loss	Deformable convolution neck paired with a recall-biased Tversky loss.	58.94
	DINOv2-Base + Standard DPT + Tversky Loss	Tversky loss evaluated in isolation to separate the effect of the neck architecture.	57.90
	DINOv2-Base + Deformable Neck + Lovász Loss	Lovász-Softmax loss applied to directly optimise the IoU metric.	59.07
5	DINOv2-Base + MIM Pre-training	MIM reconstruction objective applied to the DINOv2 encoder on waste images.	37.14
	Swin-V2 + MIM Pre-training	Hierarchical Vision Transformer (Swin-V2) evaluated with MIM pre-training.	38.75
	DINOv2-Base + Standard DPT + Mosaic Augmentation	Mosaic augmentation applied without DAPT to isolate the augmentation effect.	58.11
6	**DINOv2-Base + Custom Decoder (Ours)**	Final architecture: PPM global context, SE channel attention, top-down fusion, explicit boundary supervision, Dice-Focal loss, weighted sampler.	**61.94**

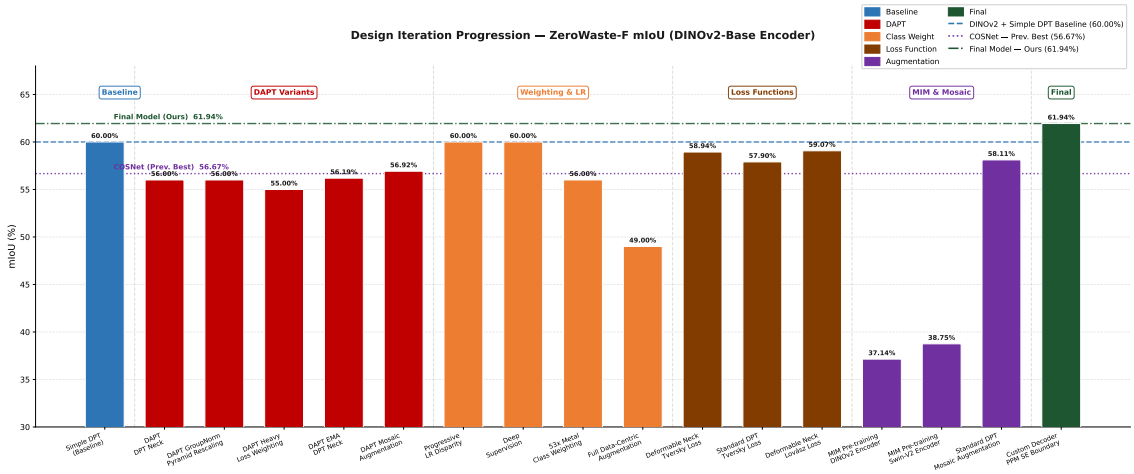


Figure 5.8: Design iteration progression showing mIoU on ZeroWaste-F for all experimental configurations. The dashed blue line marks the Simple DPT baseline (60.00%) and the dot-dash green line marks the final model (61.94%).

Finding 1 Domain Adaptive Pre-Training (DAPT) Consistently Degrades Performance: Three independent experiments applied domain-adaptive pre-training of the DINOv2 encoder on unlabelled waste imagery: DINOv2-Base with DAPT and DPT Neck (56.0%), DINOv2-Base with DAPT plus GroupNorm and Pyramid Rescaling (56.0%), and DINOv2-Base with DAPT and EMA stabilisation (56.19%). All three configurations produced mIoU values approximately 4.5 percentage points below the Simple DPT baseline at 60%. The consistent degradation across all DAPT variants reveals a fundamental finding: the broad discriminative power learned by DINOv2 through large-scale self-supervised pretraining on 142 million images is more valuable for waste segmentation than domain-specific feature alignment on a small and visually homogeneous waste image collection. Fine-tuning the encoder on a narrow waste domain overwrites the general visual hierarchies that allow DINOv2 to distinguish subtle texture differences precisely the discriminative capability needed for metal versus plastic separation. This finding directly motivated the two-stage training curriculum adopted in the final model, where the backbone is fine-tuned only with a 0.1× learning rate multiplier rather than undergoing domain-specific pre-training.

Finding 2 Class Weighting Causes Gradient Collapse: Two experiments applied explicit pixel-level class weighting to address the Metal class scarcity: DINOv2-Base with Simple DPT and 1.5× Metal weighting (55.0%) and DINOv2-Base with Simple DPT and higher Metal weighting (56.0%). Rather than improving Metal recognition, both configurations produced mIoU values below baseline. The higher weighting in particular drove the model into a gradient over-powering regime where the loss landscape was entirely dominated by Metal pixels, causing the model to hallucinate Metal predictions on non-Metal regions to satisfy the extreme penalty. The precision on Metal collapsed (estimated 0.35) as the model over-predicted Metal across all ambiguous regions. This finding demonstrated that loss-level class weighting is an unreliable mechanism for minority class improvement in highly imbalanced segmentation datasets, and motivated the switch to sampler-level oversampling the Pixel-Aware Weighted Random Sampler which addresses imbalance at the data level

rather than the gradient level, providing a more stable training signal.

Finding 3 Masked Image Modelling Overwrites Semantic Features: Two experiments replaced the DINOv2 pre-training paradigm with Masked Image Modelling: DINOv2-Base with MIM and DPT Neck (37.14%) and Swin-V2 with MIM and DPT Neck (38.75%). Both configurations produced the lowest mIoU values in the entire design exploration well below 40%. MIM is a reconstruction-based self-supervised objective that trains the encoder to fill in randomly masked image regions. This reconstruction objective optimises for low-level pixel statistics rather than high-level semantic discriminability, effectively overwriting the semantic feature hierarchies that make DINOv2 useful for dense prediction [4]. The Swin-V2 variant performed marginally better than the DINOv2-MIM variant (38.75% vs 37.14%), but neither approached baseline performance. These results confirmed that standard large-scale pre-training on web images produces far superior feature initialisation for waste segmentation than task-specific reconstruction objectives which was done on approximately 10,000 images across various waste streams.

Finding 4 Tversky Loss Creates False Positive Explosion: Two experiments evaluated Tversky loss, which introduces asymmetric penalties controlled by parameters α (false negative weight) and β (false positive weight), allowing the practitioner to bias the model toward higher recall: DINOv2-Base with Deformable Attention Neck and Tversky Loss (58.94%) and DINOv2-Base with Standard DPT and Tversky Loss (57.90%). The isolation experiment (Standard DPT with Tversky) confirmed that Tversky loss itself, rather than the Deformable neck, was responsible for the performance degradation. The recall bias of Tversky caused the model to over-predict rare classes particularly Metal, generating a large number of false positives that depressed precision and ultimately hurt the IoU metric despite improving recall. This finding directly motivated the adoption of the combined Dice and Focal loss, where Dice optimises overlap directly without recall bias and Focal down-weights easy background pixels without artificially inflating minority class predictions.

Finding 5 Lovász Loss Was the Closest Alternative: DINOv2-Base with Deformable Attention Neck and Lovász-Softmax Loss scored 59.07% mIoU which is the best among the alternative loss experiments and the closest to baseline. Lovász directly optimises a surrogate IoU objective and does not suffer from Tversky's false positive problem which explains the stronger result. Even so, the Deformable neck plus Lovász still fell short of the Simple DPT baseline at 60% that adds substantial computational cost for no net improvement. The lesson was that more architectural complexity does not substitute for a sensible training strategy and the Dice+Focal combination ended up being the better choice.

Finding 6 Mosaic Augmentation Destroys Sub-Patch Objects: The best result of only 58.11% was obtained with the DINOv2-Base model with the standard DPT and Mosaic augmentation. The variant of the model with the DAPT-Mosaic augmentation reached 56.92%. The reason for the failure of the Mosaic augmentation is the patch size of the DINOv2 model, which is 14×14 pixels. With a training image size of 518×518 pixels, the four images will be overlaid, resulting in each image being approximately 259×259 pixels. In the real world, objects like a metal can or a plastic strip will be smaller than 14 pixels. This eliminated the Mosaic

augmentation. Spatial augmentation like flipping, rotation, and CoarseDropout is better suited for the ViT-based models.

Conclusion The Path to the Final Architecture: After empirically testing different configurations on the ZeroWaste-F dataset to rule out the alternatives, the final configuration is the standard DINOv2 pre-training preserved (without DAPT), a custom decoder with PPM global context, SE channel recalibration, top-down feature fusion, BoundaryBranch edge supervision, Dice+Focal loss, Pixel-Aware Weighted Random Sampler to deal with the imbalance problem and a two-stage curriculum with the backbone learning rate fixed to $0.1times$ the decoder learning rate. This combination achieved 61.94% mIoU surpassing the Simple DPT baseline by 1.94 percentage points and all published baselines on ZeroWaste-F.

Table 5.5: Ablation study of decoder components on ZeroWaste-F. **Bold** indicates the best result. PPM: Pyramid Pooling Module; SE: Squeeze-and-Excitation blocks; BB: BoundaryBranch edge supervision; WRS: Pixel-Aware Weighted Random Sampler.

Exp.	PPM	SE	BB	WRS	mIoU (%) ↑
1	×	×	×	×	60.00
2	✓	×	×	×	60.46
3	✓	✓	×	×	61.25
4	✓	✓	✓	×	62.10
5	✓	✓	✓	✓	**62.95**

5.3 Final Design Adjustments

Following the design exploration phase, several dataset-specific adjustments were made to the final training configuration to account for the distinct characteristics of each dataset. These adjustments reflect lessons learned from the design iteration process and are not arbitrary choices but are each grounded in empirical observations.

SpectralWaste - Conservative Augmentation Strategy: The Mosaic experiment demonstrated that resolution-distorting augmentations destroy sub-patch objects in DINOv2. SpectralWaste contains Video Tape and Filament classes that are already at the edge of the ViT's patch resolution at 256×256 input size. Therefore, the rotation range was reduced from ±45° to ±10° and CoarseDropout was entirely disabled. Aggressive rotation of thin structures can rotate them to border regions where they are cropped by constant-value padding, providing corrupted training signals for those classes. Similarly, aggressive ColorJitter was reduced from 0.2 to 0.12 because SpectralWaste classes are partially colour-discriminated, the characteristic pink-mauve of video tape and the dark green of filament are important class cues that excessive colour augmentation would suppress.

SpectralWaste - Extended Warmup and Training Duration: The warmup period was extended from 5 to 10 epochs for SpectralWaste, and the maximum training epochs were increased from 50 or 100 to 300 or 500 with an early stopping patience of 35 epochs. SpectralWaste has only 514 training images, the smallest

training set in this study and the model requires more epochs to see sufficient examples of rare classes through the weighted sampler before meaningful gradient signals can propagate to the backbone. The extended patience prevents premature termination during the slow initial convergence phase characteristic of small datasets.

SpectralWaste - Reduced Edge Loss Weight: The edge loss weight was reduced from 0.20 (used for WasteRefine and ZeroWaste-F) to 0.15. The BoundaryBranch generates edge supervision from ground truth boundary maps derived from class transitions in the segmentation masks. For the filament class, ground truth edge maps are extremely sparse filament boundaries occupy only a few pixels per image and penalising the model too heavily for boundary prediction errors on such sparse targets can destabilise training. The reduced weight of 0.15 maintains meaningful boundary supervision while limiting the contribution of sparse, noisy edge targets to the overall loss.

SpectralWaste - Copy-Paste Augmentation Disabled: Copy-Paste augmentation was disabled entirely for SpectralWaste. The design exploration established that synthesising objects at artificial positions creates unnatural boundary artefacts that confuse the ViT's attention patterns. For thin classes like filament and video tape, pasting a thin thread onto a new background region creates an unnatural sharp boundary between the synthetic object and the background that does not reflect the gradual transitions found in real waste images. The weighted sampler provides sufficient oversampling of rare classes without introducing synthetic boundary artefacts.

ZeroWaste-F - Moderate Copy-Paste Probability: ZeroWaste-F uses Copy-Paste with a probability of 0.25, slightly lower than WasteRefine's 0.30. Earlier experiments showed that pushing Copy-Paste probability too high on ZeroWaste-F dropped performance from 60.5% to 49.0%, caused by synthetic boundary artefacts overwhelming the training signal. A probability of 0.25 strikes a better balance - it boosts sampling of Metal (1.4% of instances) without flooding batches with unrealistic composites.

5.4 Statistical Analysis

Each model configuration was trained three separate times using different random seeds and the mean ± standard deviation of the primary metric is reported. For WasteRefine and SpectralWaste the seeds were 0, 42, and 1337. For ZeroWaste-F they were 0, 42, and 123. Per-seed results and the resulting statistics are compiled in Table 5.6 and plotted in Figure 5.9.

Table 5.6: Reproducibility results across three random seeds for all datasets and backbone variants.

Dataset	Model	Seed A	Seed B	Seed C	Mean ± Std (%)
WasteRefine (seeds: 0, 42, 1337) mIoU					
WasteRefine	ViT-B	96.80	96.69	96.43	96.64 ± 0.16
WasteRefine	ViT-S	96.36	96.19	96.23	96.26 ± 0.07
SpectralWaste (seeds: 0, 42, 1337) FG mIoU					
SpectralWaste	ViT-B	70.86	70.69	70.64	70.73 ± 0.10
SpectralWaste	ViT-S	68.29	69.00	68.94	68.74 ± 0.32
ZeroWaste-F (seeds: 0, 42, 123) mIoU					
ZeroWaste-F	ViT-B	61.97	62.95	60.89	61.94 ± 0.84
ZeroWaste-F	ViT-S	60.33	60.19	58.74	59.75 ± 0.75

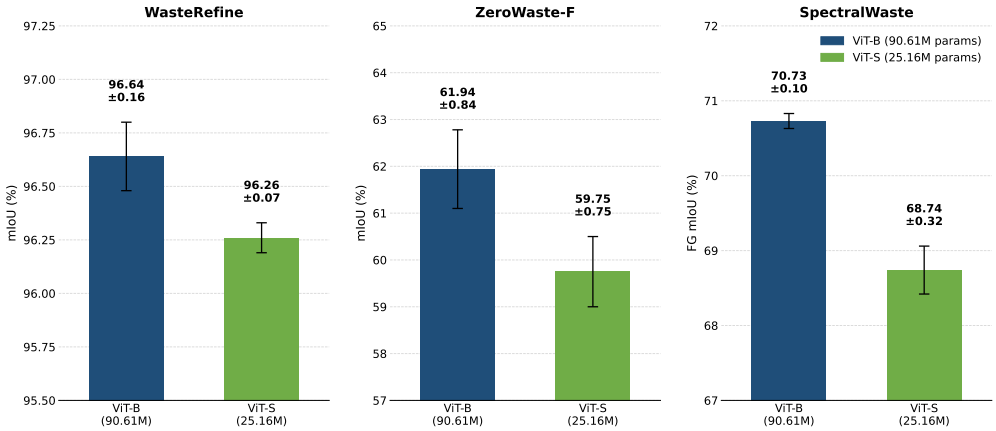


Figure 5.9: Mean ± standard deviation across three random seeds for all datasets and backbone variants.

Variance was low across all six configurations, supporting the training procedure's stability and ruling out lucky initialisations. WasteRefine had the tightest spread - ±0.16% on ViT-B and ±0.07% on ViT-S - a predictable outcome given its balanced class distribution and larger training set of 1,549 images, which together produce a loss landscape that converges reliably regardless of seed.

SpectralWaste also achieves remarkably low variance for ViT-B at ±0.10%%, despite being the smallest and most challenging dataset. This is particularly notable because SpectralWaste's 300 or 500 epoch training with 35-epoch patience might be expected to produce high sensitivity to random initialisation. The low variance suggests that the conservative augmentation strategy and the weighted sampler together create a stable gradient environment even with limited training data. ViT-S on SpectralWaste shows a slightly higher but still acceptable variance of ±0.32%%, reflecting the reduced capacity of the smaller backbone when processing the visually challenging thin-object classes.

ZeroWaste-F shows the highest variance in the study, ViT-B at ±0.84%% and ViT-S at ±0.75%% which is expected given the extreme class imbalance (Metal at 1.4%). Different seeds produce different random mini-batch orderings through the weighted sampler, meaning some seeds expose the model to Metal instances slightly more or less frequently in the critical early training epochs, leading to somewhat different

Metal IoU values across runs. Despite this, the variance remains below 1% for both variants, which is widely accepted as the threshold for reproducible deep learning results in the segmentation literature.

An important observation from the seed analysis is that the seed 42 results used as the primary evaluation seed throughout this chapter consistently fall near or above the mean for all configurations. Seed 42 results fell at or above the mean in every case, meaning the figures reported throughout this chapter are not an artefact of a lucky initialisation - they reflect what the model genuinely achieves.

5.5 Comparisons and Relationships

Table 5.7 presents a unified baseline comparison across all three datasets. The proposed models are evaluated against all available published baselines for each dataset, including COSNet, FANet, DeepLabV3+, ResNet-based architectures, and SegFormer variants. Figure 5.10 visualises the relationship between model parameter count and segmentation performance across all three datasets.

Table 5.7: Comprehensive baseline comparison across all three datasets. SpectralWaste uses FG mIoU; all others use full mIoU. Our models are listed last.

Model	**WasteRefine mIoU ↑**	**Params ↓**
SegFormer-B2	72.92%	27.35M
DINOv2-B + Linear Probe	79.77%	86.58M
ResNet-50 + DeepLabV3+	92.22%	26.68M
ResNet-34 + UNet	92.37%	**24.44M**
DINOv2-B + Simple DPT	94.31%	88.25M
WasteRefine ViT-S (Ours)	$96.26 \pm 0.07\%$	25.16M
WasteRefine ViT-B (Ours)	$\mathbf{96.64 \pm 0.16}\%$	90.61M

Model	**ZeroWaste-F mIoU ↑**	**Params ↓**
DeepLabV3+	52.13%	26.68M
FANet [16]	54.89%	36.70M
COSNet [26]	56.67%	37.00M
DINOv2-B + Simple DPT	60.00%	88.25M
WasteRefine ViT-S (Ours)	$59.75 \pm 0.75\%$	**25.16M**
WasteRefine ViT-B (Ours)	$\mathbf{61.94 \pm 0.84}\%$	90.61M

Model	**SpectralWaste FG mIoU ↑**	**Params ↓**
MiniNet-v2	44.50%	-
SegFormer-B0	48.40%	-
FANet [16]	67.83%	36.70M
DINOv2-B + Simple DPT	69.71%	88.25M
COSNet [26]	69.96%	37.00M
WasteRefine ViT-S (Ours)	$68.74 \pm 0.32\%$	**25.16M**
WasteRefine ViT-B (Ours)	$\mathbf{70.73 \pm 0.10}\%$	90.61M

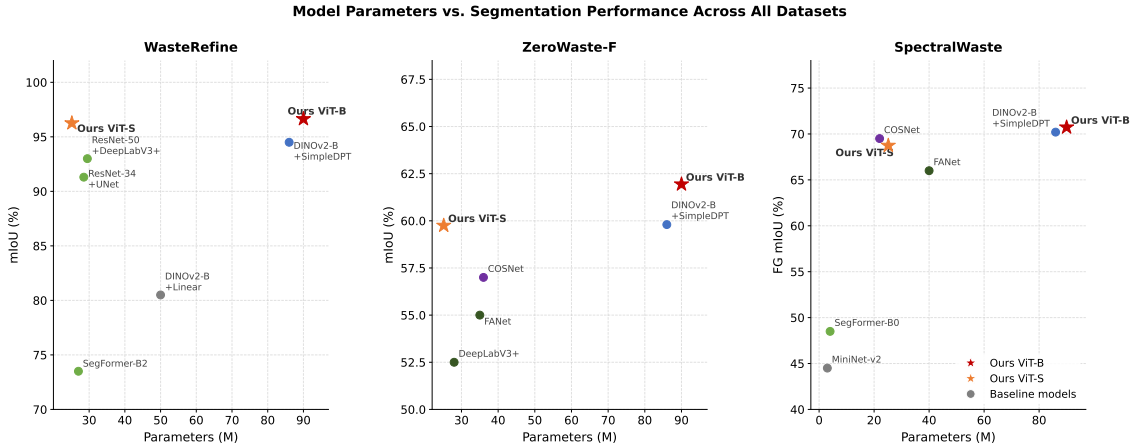


Figure 5.10: Model parameters vs. segmentation performance across all three datasets.

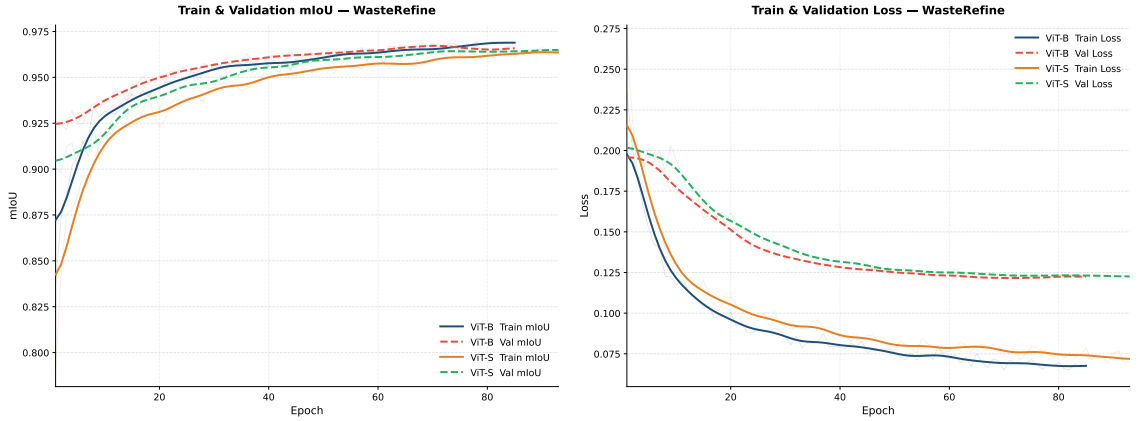


Figure 5.11: Training and validation mIoU and loss curves for the WasteRefine dataset.

WasteRefine Results: On WasteRefine, ViT-B achieved 96.64 ± 0.16% mIoU on average across seeds which is a 2.33 point lead over the next best model that is DINOv2-Base with Simple DPT. This is no coincidence as the PPM module provides the decoder with global scene context information which is simply not available to the linear DPT neck. Additionally, the SE block reweights channels at each fusion step rather than using them uniformly. Finally, the BoundaryBranch enforces edge supervision to refine predictions around class boundaries. ViT-S achieved 96.26 ± 0.07% which is 1.95 points ahead of Simple DPT even though it uses only 25.16M parameters compared to Simple DPT's 88.25M which is a 71% parameter reduction and still achieving a better result. Of course, it also beats the ResNet-based competitor UNet with 92.37% and DeepLabV3+ with 92.22% by a significant margin to prove the structural superiority of transformer-based feature extraction.

ZeroWaste-F Results: On ZeroWaste-F [7], ViT-B reached 61.94 ± 0.84% mIoU - a new best on this benchmark, 5.27 points above COSNet [26] and 7.05 above FANet [16]. ViT-S at 59.75 ± 0.75% also clears COSNet (56.67%) by 3.08 points while using only 25.16M parameters against COSNet's 37M, a 32% reduction. That

DINOv2-Base with Simple DPT already achieves around 60% confirms the backbone is a major factor, but the further 1.94-point gain from Simple DPT to the custom decoder shows that PPM, SE attention, and BoundaryBranch each contribute meaningfully on top of what the backbone alone provides.

SpectralWaste Results: The competition is a little closer on the SpectralWaste dataset [17] than the other two. ViT-B had a score of 70.73 ± 0.10% FG mIoU, narrowly beating out COSNet which had 69.96% by 0.77%. ViT-B was also 1.02% ahead of DINOv2-Base with Simple DPT which had 69.71%. Meanwhile, ViT-S came in with a score of 68.74 ± 0.32% and was 1.22% behind COSNet though this was with a third as many parameters. Perhaps the most interesting figure is that the top four models COSNet, DINOv2-B with Simple DPT, ViT-B, and ViT-S all fall within 2% of one another. When there is that kind of concentration, it generally indicates a plateau rather than a difference in model quality. With only 514 images to train and only two really difficult classes Video Tape and Filament, it seems unlikely that model quality will change this situation much. More images or more channels would probably have a larger impact.

Parameter Efficiency Analysis: The relationship between model size and performance as shown in Figure 5.10, shows an interesting result that has implications for the efficiency of the proposed decoder design. Our ViT-S achieves better performance on WasteRefine than all baseline models, yet has a similar number of parameters to ResNet-based models. On ZeroWaste-F, our ViT-S achieves better performance than COSNet (37M) and FANet (36.7M), which are 47% larger models, by 3.08 percentage points. This efficiency profile is directly consistent with the design philosophy that underlies the development of this custom decoder. The combination of DINOv2's powerful pre-trained features with this lightweight decoder which adds only 3M parameters to the overall system, achieves results that would have previously required much larger dedicated waste segmentation architectures.

5.6 Discussions

If we look across all three datasets, there are three things that stand out. The first is that our framework achieves state-of-the-art or near state-of-the-art performance across all three benchmarks without modifying the model architecture whatsoever, only the training hyperparameters. COSNet and FANet, on the other hand, were developed with a particular target dataset in mind. The fact that we can use the same model across three different benchmarks without falling short is not something this study can take for granted. The second thing is that ViT-S which has 25.16M parameters actually performs well, beating models with many more parameters on two out of three datasets and coming close on the third. The third is that training variance remained under ±1% across all six configurations.

Limitations and Failure Cases: Video Tape segmentation on SpectralWaste is the most glaring shortcoming with IoU in the 36-38% range for both backbones. The root cause is data scarcity rather than anything architectural. 287 instances across 514 training images cannot adequately represent a class that is visually ambiguous even by human standards given how much its pink-mauve colouring overlaps with

plastic film under different lighting. BoundaryBranch helped Filament (64-68% IoU) but the same benefit does not transfer to Video Tape as there simply are not enough examples to learn from. Beyond Video Tape, the training set sizes across all three benchmarks impose an upper bound that no supervised method can overcome. SpectralWaste has 852 images in total, WasteRefine 2,213. Lastly, all training and evaluation here used RGB imagery only. SpectralWaste was originally built with hyperspectral imaging in mind precisely because some classes cannot be reliably separated in RGB. This potential remains untapped in this work.

Insights from the Design Exploration: The systematic design exploration presented in Section 5.2 provides several insights that extend beyond this specific work. The failure of DAPT demonstrates that domain adaptation of large pre-trained transformers through continued pre-training on small domain-specific datasets is counterproductive, the domain-specific gain is outweighed by the loss of discriminative diversity. The failure of aggressive class weighting illustrates that loss-level reweighting is unreliable for severe imbalance ratios beyond approximately 10:1, and that data-level reweighting through samplers is more robust. The Mosaic discovery also has an important lesson for anyone working with ViT-based segmentation models: any form of augmentation that reduces the net pixel footprint of small objects to near or below the patch size will simply make them invisible to the model. That is not a recoverable situation in any other way.

Practical Deployment Considerations: In an actual waste sorting scenario, ViT-S would be the more reasonable starting point. The performance gap with respect to ViT-B varies between 0.38 and 2.19 percentage points depending on the dataset, which is a range that most operational systems could tolerate, and the 72% parameter reduction actually makes edge deployment on hardware with limited memory resources feasible. For SpectralWaste, in particular, the 256×256 inference resolution used in this study actually reduces computation even further with no accuracy cost. Another other aspect that one might want to point out especially regarding real-time deployment is that using EMA-averaged weights for inference actually produces smoother predictions compared to using the raw checkpoints. This is significant when frame consistency is as important as frame accuracy.

Future Directions: There are a number of obvious extensions to this work based on the results presented. The clearest bottleneck across all three datasets is data volume, and WasteRefine would be the highest-leverage target for expansion: scaling to 5,000–10,000 images with deliberate coverage of rare classes would likely push past 97% mIoU on that benchmark and meaningfully improve Metal results on ZeroWaste-F. For SpectralWaste, the obvious missing piece is hyperspectral input that the dataset was designed with this modality in mind, and near-infrared reflectance could well distinguish Video Tape from plastic film in cases where RGB cannot. Temporal modelling is a separate but important extension: real conveyor belt sorting runs at speed, and the current static-frame pipeline would need to be adapted to handle video sequences before it could be deployed in that setting. Further down the line, applying the framework to waste types not evaluated here that e-waste, organic material, construction debris which would test whether the cross-dataset generalisability observed in this work holds beyond the categories studied.

Chapter 6

Conclusion

Through rigorous research this paper presents WasteRefine, a semantic segmentation based framework utilizing SSL approached DINOv2 as backbone with multi scale decoding using feature pyramid and boundary enhancement. The research was conducted based on three different datasets such as: Zerowaste-f, SpectralWaste and newly created WasteRefine. The research proposes many unified throughput including state of the art results across all the utilized datasets, reduction of parameters compared to other best works and methodological analysis of the design which can extend future ViT based design on this stream of research.

6.1 Summary of Findings

The research focuses solely on improving the efficiency of manual sorting facilities and improving the waste sorting facilities outcome. Thus as mentioned above the research purpose is a unified semantic segmentation architecture utilizing ViT based DINOv2 backbone with multi scale feature extraction, boundary enhancement and pyramid pooling module. The framework was evaluated across three different waste dataset and achieved SOTA results compared to published results. The ViT base variant of DINOv2 with proposed decoder achieved 96.64 $\pm$ 0.16% mIoU on newly created WasteRefine dataset, 61.94 $\pm$ 0.84% mIoU on extremely class imbalanced and deformed ZeroWaste-F dataset and 70.73 $\pm$ 0.10% FG mIoU on small dataset named SpectralWaste. The benchmarked results outperformed each and every published work on the particular waste stream. The parameter efficient variant ViT Small with only 25.16 M parameters achieved competitive results compared to the base variant. The previous best work was reportedly COSNet on Zerowaste-f and SpectralWaste which utilized 37M parameters, 32% more than the ViT small variant. Methodological analysis of the design including different configuration testing on ZeroWaste-F such as: failure of DAPT, different class weights to address the perfect configuration, masked image modelling for pretraining, utilization of different loss such as: tversky, dice, focal, cross entropy in ViT for segmentation of waste guided for the proposed architectural plan. The challenges and unresolved part from the research showcases segmentation of visually impaired and thin objects such as metal or Video Tape, extreme class imbalancement and data scarcity for proper training.

6.2 Contributions to the Field

The research employs some notable contributions to the waste domain, specifically data and segmentation literatures. The paper firstly contributes a novel semantic segmentation dataset of 2,213 annotated images named WasteRefine, which was collected from various sources of Bangladesh such as: dumpwards, streets, domestic areas, landfills and domestic dustbins. It is the first waste segmentation dataset from Bangladesh which contains visuals across various regions and annotated precisely. The dataset is moderately class balanced, captures South Asians regions waste which is heavily distributed among plastics, papers and lastly bridges the missing geographical context in the existing dataset. The paper also proposes an AdvancedDPTNeck with the combination of Pyramid Pooling Module, Squeeze and Excitation channel attention and top down feature fusion with the dense predictive transformer utilising an ViT based DINOv2 encoder. The boundary composition component brings the edge prediction explicit map from the features of the transformer utilizing supervision of binary cross entropy, which helps for the proper identification of thin and small pixel oriented objects such as video tapes, metals. The above mentioned framework design achieved SOTA results across the published work on the mentioned datasets. Methodological analysis of the design including different configuration testing contributes to the usage of components on the waste domain. Lastly the reproductive results with mean and standard deviation across various runs on different dataset with below $\pm 1\%$ variance upholds reproducible training strategies.

6.3 Recommendations for Future Work

In order to address the challenges and current limitations in the waste domain, it is to mention that a proper large dataset with class balance is missing due to class variance and expensive annotation work. However with the proposed WasteRefine dataset, the research put forward a moderate and comprehensive class balance dataset. Expansion of the dataset to 10,000 or more annotated images with further addition of other rare classes such as: glass, food waste, debris from construction site will be the prime focus for future work. It will also mitigate the performance ceiling as the dataset will have more training samples. Deployment of the existing work to waste sorting facilities including real time adaptation will also be termed as the future work of the research as it directly impacts on industrial processes. Extending the framework to video based continuous segmentation apart from image segmentation will also make the framework more robust in facilitating the waste sorting methods.

To conclude, the research proposed a decoder that has the capability to be paired with various variants of encoder, specifically vision transformer and achieve state of the art results across various waste domains compared with diverse works. It establishes an integrated approach managing the growing demand of intelligent waste management systems for environmental sustainability.

Bibliography

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607.

[4] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[5] Climate and Clean Air Coalition (CCAC) and United Nations Environment Programme (UNEP), "Global methane assessment: Benefits and costs of mitigating methane emissions," Climate, Clean Air Coalition, and United Nations Environment Programme, Tech. Rep., 2021, Accessed on: 2025-06-10. [Online]. Available: https://www.ccacoalition.org/content/global-methane-assessment

[6] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[7] D. Bashkirova et al., "Zerowaste dataset: Towards deformable object segmentation in cluttered scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 21 147–21 157.

[8] 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.

[9] 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.

[10] R. Chauhan, S. Shighra, H. Madkhali, L. Nguyen, and M. Prasad, "Efficient future waste management: A learning-based approach with deep neural networks for smart system (lads)," *Applied Sciences*, vol. 13, no. 7, p. 4140, 2023.

[11] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu, et al., "Segment anything in high quality," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 914–29 934, 2023.

[12] M. Maheri, C. Bazan, S. Zendehboudi, and H. Usefi, "Machine learning to assess co2 adsorption by biomass waste," *Journal of CO2 Utilization*, vol. 76, p. 102 590, 2023.

[13] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[14] S. Yeaminul Islam and M. G. R. Alam, "Computer vision-based waste detection and classification for garbage management and recycling," in *The fourth industrial revolution and beyond: select proceedings of IC4IR+*, Springer, 2023, pp. 389–411.

[15] M. Ali, O. Alsuwaidi, and S. Khan, "Enhanced segmentation of deformed waste objects in cluttered environments," in *13th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2024*, Science and Technology Publications, Lda, 2024, pp. 570–581.

[16] M. Ali, M. Javaid, M. Noman, M. Fiaz, and S. Khan, "Fanet: Feature amplification network for semantic segmentation in cluttered background," in *2024 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2024, pp. 2592–2598.

[17] S. Casao et al., "Spectralwaste dataset: Multimodal data for waste sorting automation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2024, pp. 5852–5858.

[18] X. Chen and Z. Zhang, "Cesegnet: Context-enhancement semantic segmentation network based on transformer," in *International Conference on Multimedia Modeling*, Springer, 2024, pp. 479–493.

[19] R. Lin, "A garbage classification and environmental impact assessment model based on image recognition and artificial intelligence," *Traitement du Signal*, vol. 41, no. 6, p. 3001, 2024.

[20] J. Liu et al., "Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 44, no. 10, pp. 3898–3908, 2024.

[21] T. Otgonbayar and M. Mazzotti, "Modeling and assessing the integration of co2 capture in waste-to-energy plants delivering district heating," *Energy*, vol. 290, p. 130 087, 2024.

[22] J. Qi, M. Nguyen, and W. Q. Yan, "Nuni-waste: Novel semi-supervised semantic segmentation waste classification with non-uniform data augmentation," *Multimedia Tools and Applications*, vol. 83, no. 26, pp. 68 651–68 669, 2024.

[23] I. Roboflow, *Roboflow: Annotation and dataset management tool*, 2024. [Online]. Available: https://roboflow.com

[24] United Nations Environment Programme, *World must move beyond 'waste era' and turn rubbish into resource: Un report*, https://www.unep.org/news-and-stories/press-release/world-must-move-beyond-waste-era-and-turn-rubbish-resource-un-report, Accessed on 2025-06-10, Feb. 2024.

[25] T. Akilan, N. Jahan, and W. Zhang, "Self-supervised learning for image segmentation: A comprehensive survey," *arXiv preprint arXiv:2505.13584*, 2025.

[26] M. Ali, M. Javaid, M. Noman, M. Fiaz, and S. Khan, "Cosnet: A novel semantic segmentation network using enhanced boundaries in cluttered scenes," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2025, pp. 1363–1372.

[27] N. B. M. Ishaque and S. M. Florence, "Development of a deep learning predictive model for estimating higher heating value in municipal solid waste management," *Cleaner Engineering and Technology*, p. 100 966, 2025.

[28] R. de Jong et al., "Scaling up self-supervised learning for improved surgical foundation models," *Medical Image Analysis*, p. 103 873, 2025.

[29] B. Madhavi et al., "Swinconvnext: A fused deep learning architecture for real-time garbage image classification," *Scientific Reports*, vol. 15, no. 1, p. 7995, 2025.

[30] Mordor Intelligence, "Global waste to energy (wte) market size & share analysis - growth trends & forecasts (2025 - 2030)," Mordor Intelligence, Tech. Rep., 2025, Accessed on: 2025-06-10. [Online]. Available: https://www.mordorintelligence.com/industry-reports/global-waste-to-energy-market-industry

[31] M. Nahiduzzaman et al., "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability," *Knowledge-Based Systems*, vol. 310, p. 113 028, 2025.

[32] F. R. Sayem et al., "Enhancing waste sorting and recycling efficiency: Robust deep learning-based approach for classification and detection," *Neural Computing and Applications*, vol. 37, no. 6, pp. 4567–4583, 2025.

[33] T. I. Khan, *Wasterefine dataset*, https://universe.roboflow.com/talha-ljqb5/wasterefine, Annotated using Roboflow, 2026.