Enhanced Segmentation of Deformed Waste Objects in Cluttered Environments

Muhammad Ali[a], Omar Alsuwaidi and Salman Khan[b]
Muhammad Bin Zayed University of Artificial Intelligence, U.A.E.
{firstname.lastname}@mbzuai.ac.ae

Keywords: Segmentation, Multi Scale Feed Forward Network(MSFN), Dilation, FreezeConnect, Weighted Learning.

Abstract: Recycling is a crucial process for mitigating environmental pollution; however, due to inefficiencies in waste sorting, a significant portion of recyclable waste is being underutilized. The complexity and disorganization of waste streams make it challenging to efficiently separate recyclable materials. Identifying recyclable items in cluttered environments requires the recognition of highly deformable objects by computer vision systems. To this end, we propose a computer vision-based approach capable of efficiently separating recyclable materials from waste, even in disorganized settings, by recognizing highly deformable objects. We extend an existing large-scale CNN-based model, the InternImage, by introducing Mutli-scale networks and combining cross-entropy and dice loss for improved segmentation. Our focus is on enhancing the segmentation of the ZeroWaste-f dataset, an industrial-grade dataset for waste detection and segmentation. We further propose a unique Mutli-scale feed-forward network configuration and integrate it with the InternImage architecture to effectively model Multi-scale information on the challenging ZeroWaste-f dataset for both waste detection and segmentation tasks. This improvement is further enhanced by introducing a novel Freezeconnect module which helps to counteract neuron co-adaptation during training by redistributing the learning (gradient signal) across the network. We compare our model with existing state-of-the-art baseline methods on ZeroWaste-f and TrashCAN datasets to demonstrate the effectiveness of our method.

1 INTRODUCTION

Despite employing a combination of machinery and manual labor, Materials Recovery Facilities(MRFs) currently suffer from unsatisfactory recycling rates and profit margins (Gundupalli et al., 2017). For instance, in 2018, the United States recycled less than 35% of its recyclable waste. In addition to efficiency concerns, worker safety is a paramount issue in manual waste sorting, as workers are routinely exposed to hazardous and unsanitary items, such as knives and medical needles. Moreover, the highly cluttered nature of waste streams poses a substantial obstacle in achieving automated waste detection, which entails identifying waste objects for removal from conveyor belts (Bashkirova et al., 2021).

Current progress in object classification and segmentation offers promising opportunities to enhance the efficiency and safety of the recycling process. Traditional CNN-based approaches in semantic segmentation primarily utilize local, short-range structures within images to understand their semantics, as noted in studies (Chen et al., 2017). However, their fixed geometric structures inherently restrict them to capturing only short-range context. To address this limitation, various methods like dilated convolution (Wu et al., 2019; Wu et al., 2019) and channel/spatial attention models (Huang et al., 2022) have been explored. The introduction of vision transformers (Dosovitskiy et al., 2020), which leverage self-attention mechanisms, has significantly transformed the landscape of semantic segmentation. However, these models require additional effort to achieve better performance when segmenting objects in a cluttered background. Therefore, it's important to address the challenges posed by background clutter, which can hinder the model's ability to focus on individual objects. On top of it various scale diversity also presents difficulty in capturing small objects. In light of these challenges, our research is focused on improving waste segmentation in complex environments, aiming to advance the recovery of recyclable materials from the current rate of 35% (Kaza et al., 2018). We developed a novel mixed-

[a] https://orcid.org/0000-0001-9320-2282
[b] https://orcid.org/0000-0002-9502-1749

Ali, M., Alsuwaidi, O. and Khan, S.
Enhanced Segmentation of Deformed Waste Objects in Cluttered Environments.
DOI: 10.5220/0012424900003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 570-581
ISBN: 978-989-758-684-2; ISSN: 2184-4313

scale block configuration specifically designed for the effective segmentation of waste in cluttered environments. To benchmark our model effectively against relevant datasets, we selected the ZeroWaste-f dataset. The choice is motivated by the limitations of existing open-source datasets, which often comprise limited data or are produced in controlled, clean settings. Such datasets do not adequately capture the complexities and variabilities encountered in real-world waste sorting environments. While there are existing datasets like TrashCan 1.0 (Hong et al., 2020), Trash-ICRA19 (Fulton et al., 2020; Hong et al., 2020), and TACO (Proença and Simões, 2020), used for segmentation tasks, they fall short in representing the challenging and cluttered conditions typical in waste management scenarios. The ZeroWaste-f dataset (Bashkirova et al., 2021), with its more realistic and diverse samples, offers a more suitable benchmark for evaluating our model's performance in practical applications Our main contributions can be summarized as following.

- Our newly designed basic block integrates sophisticated elements, including Layer Normalization (LN), Multi-Scale Feedforward Network (MSFN), and GELU activation. This combination has proven to be highly effective in various visual processing tasks, showcasing the strength of our approach in advancing visual task performance.
- Subsequently, we introduce a multi-scale feedforward network to design a novel convolutional backbone network. By processing inputs at multiple scales, the network can effectively segment smaller, finer details as well as larger objects, crucial for accurate waste segmentation in complex scenes.
- We further introduce an innovative freeze-connect trick to retain learning. It ensures the robustness of the model by preserving the integrity of neural activity and redistributing the gradient signal, which is vital for learning intricate patterns in cluttered environments.
- We additionally introduce dilation the stem layer of the model. In cluttered waste environments, dilated convolutions help in capturing broader context, allowing the model to better understand spatial relationships between waste objects and their surroundings
- Moreover, we effectively learn better representations by employing weighted learning with cross-entropy (Mao et al., 2023) and Dice loss (Sudre et al., 2017). These combinations facilitate better approximations for CNNs with extended-range dependencies and adaptive spatial aggregation using an improved 3x3 DCN operator.

By combining these techniques, the model is equipped to handle the diverse challenges presented by waste segmentation in intricate scenes including variability in object size and shape to the prevention of overfitting and the need for adaptability to irregular forms. Finally, we evaluate our proposed model on pertinent vision tasks, such as semantic segmentation as well as detection and got improved mIoU and accuracy as given in the results section against Zero Waste and TrashCAN dataset.

2 RELATED WORK

In recent years, state-of-the-art performance in recognizing general object classes from natural scene images has been achieved by semantic segmentation models, such as Mask R-CNN (He et al., 2017) and DeepLabv3 (Chen et al., 2017). These models effectively localize objects within images while simultaneously generating high-quality segmentation masks, facilitating efficient interaction between robots and target objects ("Minaee et al., 2021). However, these models' data-hungry nature necessitates large volumes of annotated data for training, which can be both challenging and costly in specialized application scenarios. Waste recycling annotation, in particular, demands expert labelers and incurs even higher costs due to the complexity of the environments involved.

Semi-supervised segmentation methods have been proposed to address the limitations of fully-supervised methods by concurrently learning from both annotated and unannotated images. For example, ReCo (Liu et al., 2021a) introduces an unsupervised loss to optimize intra-class and inter-class variance of pixels, resulting in improved segmentation performance. On the other hand, weakly-supervised methods utilize more readily available annotations, such as image-level tags, and employ Class Activation Maps (CAM) (Jung and Oh, 2021) to identify the most discriminative regions. These regions are then used as pixel-level supervision for segmentation networks. PuzzleCAM (Jo and Yu, 2021) improves the quality of CAMs without adding layers by ensuring consistency between partial and full features. Although advanced segmentation models have been developed, their application to cluttered real-world scenarios presents challenges, such as domain shift and poor generalization. Also To address the transformer-based architecture's issue of high computational complexity and memory requirement (Dosovitskiy et al., 2021), several methods have been proposed. PVT

(Wang et al., 2021; Bello et al., 2020) and Linformer (Wang et al., 2020) perform global attention on the downsampled key and value maps, respectively. DAT (Pathak et al., 2015) employs deformable attention to sparsely sample information from value maps, while HaloNet (Vaswani et al., 2021) and Swin transformer (Liu et al., 2021b) rely on local attention mechanisms and utilize haloing and shift operations for information exchange among neighboring local areas. These methods aim to improve the computational efficiency of global attention in transformer-based models, enabling their application in more challenging downstream tasks. After the availability of large-scale datasets and computational resources, convolutional neural networks (CNNs) became mainstream for visual recognition. Several deeper and more effective neural network architectures have been proposed since AlexNet (Krizhevsky et al., 2012), including VGG (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), ResNet (He et al., 2016), ResNeXt (Xie et al., 2017), and EfficientNet (Tan and Le, 2019). Modern CNNs (Zhai et al., 2022) have shown promising performance on vision tasks by adopting advanced designs and logic of transformers such as incorporating the self-attention mechanism. They discover better components through macro/micro designs and introduce improved convolutions with long-range dependencies along with adaptive spatial aggregation techniques. In our study, we build upon the findings of Wenhai Wang et al. (Wang et al., 2022b), which demonstrated the effectiveness of large-scale CNN-based models. convolution v2 (DCNv2) (Zhu et al., 2019) we provide experimental results using our ZeroWaste-f dataset. While the previous research emphasize scale diversity (Zamir et al., 2021) in image restoration, our unique contribution introduces a novel multiscale module for segmentation enhancement. Incorporating a multi-scale feed-forward network is paramount to addressing the specific challenges posed by our dataset, which involves highly deformed shapes in cluttered environments. By integrating this network architecture into our model, we anticipate achieving improved results and better adaptability to the complexities presented by the data.

3 METHODOLOGY

Our methodology is inspired by the current state-of-the-art vision model InternImage (Wang et al., 2023; Zamir et al., 2021) using UperNet(Xiao et al., 2018), which is a versatile model used for various computer vision tasks. Our goal is to tailor it towards segmentation tasks specifically and evaluate its performance on our dataset while benchmarking against the original model. Initially, we adjust the model's hyperparameters accordingly to optimize results, while also introducing a Mutli-scale feed-forward network (FFN) that enhances the model's ability to handle Mutli-scale object settings. We also incorporate a relatively large dilated convolution kernel in the stem layer to capture more spatially aware representations across multiple channels and aggregate information from an expanded receptive field, generating rich feature maps.

Furthermore, we employ a combination of two different loss functions, cross-entropy (CE) and Dice (Sudre et al., 2017), to drive the model's training process. This implementation propagates a stronger gradient signal throughout the giant network, improving the learning process. Our integration of the MSFN, dilated layers, and combined loss function constitutes substantial and novel contributions that significantly enhance the segmentation model's performance. We demonstrated these enhancements through comprehensive evaluations, showcasing their effectiveness. We firmly believe that our work represents a notable advancement in the field of image segmentation, offering valuable insights and improved results for various applications. In Sections 3.3, 3.2 and 3.4 we provide further elaboration on the model's main building blocks and modifications, which have been tailored specifically for segmentation tasks in cluttered environments, such as those commonly encountered with waste objects.

3.1 Basic Block

Our proposed block incorporates advanced components such as Layer Normalization (LN), Multi-Scale Feedforward Network (MSFN), and GELU activation, it demonstrates effectiveness across a range of visual tasks (Wang et al., 2021). Additionally, it utilizes separable 3x3 depth-wise convolutional kernels followed by a linear projection to estimate the sampling offsets and modulation scales of deformable convolutions. We use the idea of separable convolutions to predict the sampling offsets and modulation scales of deformable convolutions. We show our main block as a Multi-Scale block where core operator is DCNv3 and then input features are passed through LN and then pass through Multi-scale feed-forward network as illustrated in Figure 1. Design and the composition of this block as in Figure 1 will be described in sections 3.2 to 3.5.

3.2 Multi Scale Feed Forward Network

Here, we propose deploying a Multi-Scale Feedforward Network (MSFN), which addresses the issue of single-scale convolutions used in previous implementations by combining information from multiple scales to improve the model's robustness against occlusions and viewpoint changes. The MSFN includes two Mutli-scale depth-wise convolution paths in the transmission process, taking into account the correlations of Mutli-scale feature maps. After Layer Normalization and channel dimension expansion using a 1x1 convolution, the processed input is fed into two parallel branches that employ 3x3 and 5x5 depth-wise convolutions each, enhancing the extraction of Mutli-scale local information. This feature fusion procedure within the MSFN is designed to better integrate the short-range and long-range dependencies, improving the model's overall segmentation performance. The motivation behind using the MSFN is to enhance the model's segmentation ability to capture both local details and global context, allowing the incorporation of local feature extraction and fusion at different scales. By effectively fusing features at multiple scales, the MSFN optimizes the trade-off between local and global information, leading to improved segmentation accuracy and performance on a variety of images with diverse object scales and complexities. A diagram illustrating the MSFN can be seen in Figure 1. The MSFN layer can be modeled as:

$$\begin{aligned}
X' &= f_{1\times1}(LN(X)) \\
X_A &= \sigma(f_{3\times3}(X')) \\
X_B &= \sigma(f_{5\times5}(X')) \\
\hat{X_A} &= \sigma(f_{3\times3}[X_A, X_B]) \\
\hat{X_B} &= \sigma(f_{5\times5}[X_A, X_B]) \\
X &= \sigma(f_{1\times1}[\hat{X_A}, \hat{X_B}]) + X \qquad (1)
\end{aligned}$$

where $f_{1\times1}$ is 1×1 convolution while $f_{3\times3}$ and $f_{5\times5}$ are depthwise convolution whereas sigma is the GELU activation function. Eq.(1) describes the output signal after MSFN block.

3.3 Stem Layer with Dilation

The stem layer is a critical component of the model, designed to perform preliminary feature extraction from the input image and generate pertinent feature maps for downstream tasks. The prepared input data is then processed by subsequent deformable convolutions, which help the network learn spatial transformations and more complex patterns in the images. Since segmenting a cluttered scene depends on multiple spatial cues and requires long-range dependence, we propose augmenting the stem layer by incorporating a large dilated convolutional kernel. By adding the dilated convolutional kernel to the stem layer, we achieve better performance in our semantic segmentation tasks. Our experiments empirically demonstrate that this modification helps capture global details and larger context in the input images, leading to more accurate and robust predictions.

The stem layer comprises 3x3 convolutions with a stride of two, followed by Layer Normalization (LN), essentially acting as an improved downsampling layer through the use of an adaptive spatial aggregation technique. The specific components of the stem layer are detailed in Figure 1. Overall, the stem layer consists of two regular convolutional layers with intermediate normalization and activation layers. Additionally, a dilated convolutional layer is added after the second convolutional layer to capture context information with a larger receptive field.

3.4 Cross-Entropy + Dice Loss

To further enhance the segmentation performance, we used a weighted combination of cross-entropy loss and Dice loss. Dice loss enhances object localization by measuring overlap between predicted and ground truth masks, improving fine detail capture compared to Cross-Entropy loss. The weighted loss, represented by L, is composed of the cross-entropy loss, L_{CE}, and the Dice loss, L_D. By combining both cross-entropy and Dice losses, our model achieved better results compared to using only cross-entropy or Dice loss individually. The combined loss L can be formulated as such:

$$L = L_{CE} + \alpha L_D \qquad (2)$$

To optimize our approach for aligned representation learning, we experimented with different weight combinations. After evaluating several options, we found that assigning a weight of $\alpha = 0.2$ to L_D provides us with the best results as given in Eq.(2). The α vs IoU effects are given in the Appendix in Figure 1.

3.5 Freezeconnect

In the pursuit of addressing overfitting in deep learning models, we introduce a novel regularization technique called *Freezeconnect*. This approach as in Figure 3 contrasts with conventional methods such as Dropout, where neurons are randomly dropped during the forward pass, potentially disrupting the neural activity within the network Unlike Dropout, Freezeconnect preserves the overall output signal by not dropping any neurons. Freeze connect maintains the integrity of the neural activity within the network, cru-

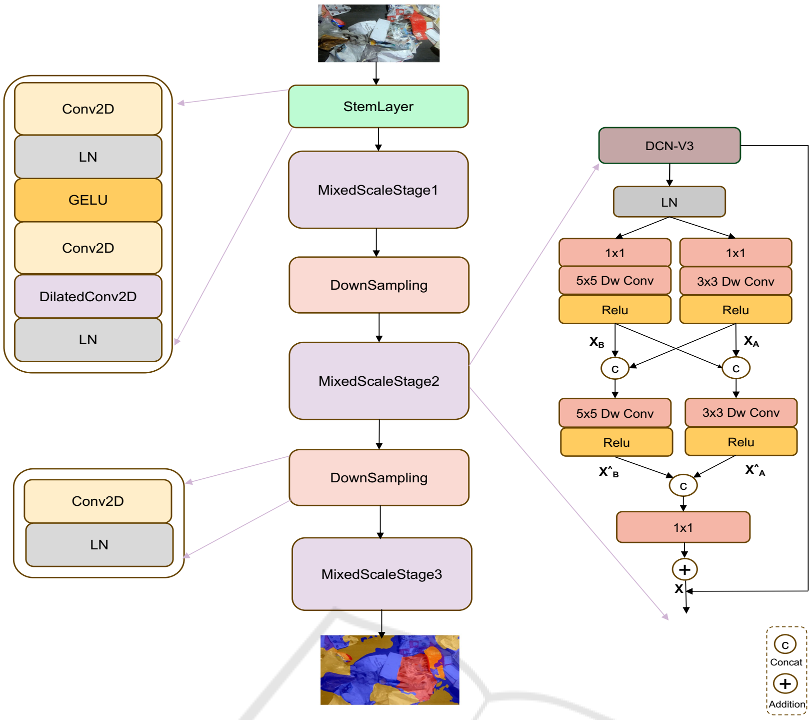
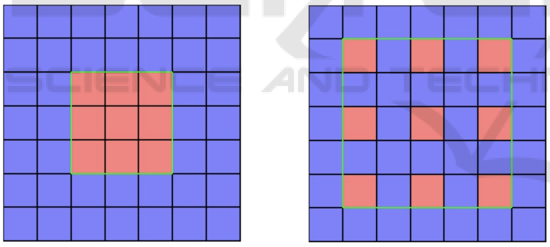


Figure 1: The overall architecture of our proposed model, primarily comprising the Sampling Layer, Convolution Layer, and Multi-Scale Feed-Forward Network (MFFN). 'LN' denotes Layer Normalization, while 'DW-Conv' stands for Depth-Wise Convolution. Each architectural modification is further elaborated in the respective sections detailing the individual blocks.



(a) Normal Conv2D (b) Dilated Conv, rate 2

Figure 2: An illustration of dilated convolution, showcasing the influence of the dilation rate on the effective receptive field. We utilize a dilation rate of 2 to facilitate the model's capability to capture more comprehensive spatial information from an extended receptive field.

cial for accurately identifying and segmenting waste in complex, cluttered environments where every neural connection can be vital for recognizing subtle differences. Instead, this method operates during the backward pass of the network, subtly modifying the main gradient signal flowing through each neuron. Specifically, branches of the gradient signal are randomly zeroed out based on a probability parameter p, and the remaining non-zeroed gradient signal branches are amplified in proportion to this probability. On one hand, it retains the integrity of the model's output signal to the response variable, ensuring that neural activity remains uninterrupted. On the other hand, it simultaneously acts as a regularizer, redistributing the gradient signal across the entire network. The core motivation to use Freezeconnect resides in its capacity to mitigate model overfitting in a gentle manner without altering the overall architecture of the network. By redistributing the gradient signal across the network without altering its overall architecture, Freezeconnect effectively reduces the risk of overfitting, ensuring that the model generalizes well to new, unseen cluttered environments, which is essential for the dynamic and varied nature of waste segmentation tasks. Experimental results, as detailed in subsequent sections, underscore the effectiveness of Freezeconnect in enhancing model robustness and performance, thereby positioning it as a valuable contribution

4 EXPERIMENTATION

To evaluate our model, we conduct extensive experiments on the ZeroWaste-f dataset and report metrics such as mean Intersection over Union (mIoU) and pixel accuracy for comparison with previous works on the same dataset. Additionally, we compare our results with those achieved by SOTA models on the

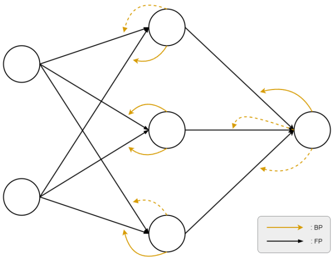


Figure 3: Freezeconnect randomly zeroes out branches of the gradient signal and amplifies the remaining branches. Each circle represents a neuron and solid lines indicate forward signal paths. Dotted lines denote selectively nullified gradients during backpropagation, emphasizing the regularization effect of the Freezeconnect approach in contrast to Dropout.

same dataset, including the DeepLabv3. Our proposed model outperforms all of these models on the ZeroWaste-f dataset, achieving a new state-of-the-art performance with a mIoU of 53.95 as well as an accuracy of 91.30. These results demonstrate the effectiveness of our approach for accurate and efficient segmentation as well as detection of waste materials in the ZeroWaste-f dataset. We further validate our results on the TrashCAN dataset. We first define setups requirement and detailed results are given in section 5.

4.1 Segmentation Task - Setup

Our proposed model is specifically designed for the task of semantic segmentation, which involves labeling each pixel in an image with its corresponding class or category. We use the open-source toolbox MMSegmentation (Contributors, 2020). We train the model using the UperNet framework (Xiao et al., 2018), initializing the backbone with pre-trained classification weights from the ImageNet (Deng et al., 2009) dataset. The model is trained for 120k iterations on the ZeroWaste-f dataset, employing the AdamW optimizer (Kingma and Ba, 2017) with an initial learning rate set at 6×10^{-5}. In accordance with established research protocols, we implement a learning rate decay schedule characterized by a polynomial decay strategy featuring a power exponent of 1.0 and set the crop size to 512 and the batch size to 16 for training. To evaluate the model's performance, we utilize the ZeroWaste-f as well as TrashCAN datasets. Further details in Section 5

4.2 Detection Task - Setup

To gear our model towards objection detection task and following established practices (Liu et al., 2021b; Wang et al., 2022c), we validate the performance of our model on top of the MaskRCNN framework. We initialize the backbone with pre-trained classification weights from ImageNet and trained our models for 120k iterations with default settings. This initialization provides an effective starting point for our detection models. Section 5.3 provides further detail. We benchmarked our method against both the ZeroWaste-f and TrashCAN datasets

5 RESULTS

Our primary focus is on segmentation, which we also extend to the task of object detection. In the following sections, we present a comprehensive analysis of the results obtained from our proposed approach in contrast to the baseline performance of the InternImage mode. Table 1 presents an ablation study, showcasing the improved contribution impact of different individual and grouped modules on the end result. Meanwhile, Table 2 provides a per-class comparison of the IoU metric, where our model demonstrates superior segmentation results across all classes, including those representing more complex waste shapes.

5.1 Ablation Study

The adoption of a combined loss function in our model enabled us to concurrently optimize the balance between two key objectives: maximizing the overlap between the predicted and ground truth segmentation masks, represented by the Dice loss, and minimizing the divergence in probability distributions between the predicted and true labels, as indicated by the Cross-Entropy (CE) loss. This dual-focus approach led to a substantial enhancement in model performance. Specifically, when compared to using either loss independently, the combination of both losses resulted in a 1.03% improvement in the mIoU (mean Intersection over Union) value, as shown in Table 1.

Furthermore, the incorporation of dilated convolution layers led to higher accuracies in our results, as demonstrated in Tables 1 and 2. Specifically, the segmentation accuracy increased by 0.35% compared to the baseline, indicating the model's enhanced ability to differentiate and recognize subtle discrepancies among various classes of waste. Further, the implementation of Freezeconnect's regularization allowed the model to redistribute its gradient signal across different pathways, leading to more uniform learning throughout the network. This potential innovation in optimization landscapes contributed to improved gen-

Table 1: Ablation Table demonstrating the incremental impact of module additions on model performance with the ZeroWaste-f Dataset.

Dil. Conv	CE+DICE	FConnect	MSFN	mIoU	Acc
-	-	-	-	45.61	83.36
✓	-	-	-	45.96	86.02
✓	✓	-	-	46.64	86.05
✓	✓	✓	-	47.60	86.84
-	✓	-	✓	51.05	|87.45
✓	✓	✓	✓	**52.80**	**91.63**

eralization, as evidenced by an almost 1% improvement when added, as depicted in Table 1. The Multiscale Feed-forward Network (MSFN) architecture was designed with a focus on efficiency and effectiveness, making it a versatile solution for various computer vision applications. It yielded significantly improved results, with a performance enhancement of 3.45% in the mIoU value over the previous values (Table 1). We also present the model's confusion matrix in Figure 4, emphasizing the significantly higher prediction accuracy achieved for the soft plastic and metal classes compared to the other two classes.

5.2 Semantic Segmentation Task

As depicted in Table 3, for the task of semantic segmentation, our proposed model consistently outperforms the previously established state-of-the-art model on the ZeroWaste-f dataset (Liu et al., 2021b). Our model achieves the highest Intersection over Union (IoU) of 53.95 on the ZeroWaste-f Dataset, as shown in Table 3. The IoU is a key metric in evaluating the quality of segmentation, measuring the overlap between the predicted segmentation and the ground truth. This high IoU score demonstrates the precise segmentation capabilities of our model. Figure 5 presents a few sample results, while additional examples can be found in the Appendix.

While transformer-based models have emerged in recent works (Liu et al., 2022b), our results provide further evidence supporting the continued relevance and effectiveness of CNN-based models in vision tasks. This is particularly pronounced when dealing with smaller-sized datasets, where our CNN-based model has shown superior performance. The success of our model in handling such data limitations underscores the robustness of CNNs in tackling real-world vision tasks, even in scenarios with restricted training samples.

Per Class Comparison

In our comprehensive analysis, we evaluated the impact of incorporating various modules on the performance of our model using per-class metrics as given in Table 1. The results showed consistent improvements for all classes. Introducing the Multi-Scale Feature Network (MSFN) led to significant increases in Intersection over Union (IoU) scores for most classes, indicating improved segmentation accuracy. We observed promising outcomes for soft-plastic, which is typically challenging to predict or segment due to its transparent nature. Additionally, we noted improved results for rigid plastic and metal, owing to the distinct characteristics of their compositions. However, the cardboard class exhibited a different behavior. This can be attributed to the nature of cardboard shapes in the dataset, which are less deformed compared to other classes, making the MSFN's contribution relatively less pronounced. Nevertheless, it is important to note that the IoU for the cardboard class did increase, highlighting the positive impact of the MSFN on segmentation results, albeit to a lesser extent than for classes with more deformed shapes. Also metal and cardboard are less in numbers as compared to the other two classes.

Our findings underscore the varying influence of the MSFN on segmentation performance across different classes, with the level of deformation in objects playing a crucial role in determining the extent of improvement. These insights hold significant potential for optimizing segmentation models and customizing them to the specific characteristics of individual classes.

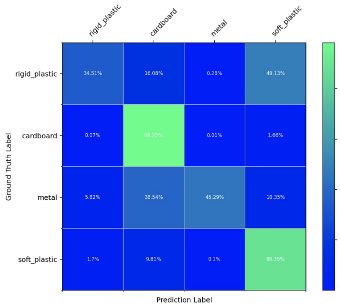


Figure 4: The confusion matrix of our model's semantic segmentation performance on the ZeroWaste-f dataset.

Table 2: A comparison of individual classes from the ZeroWaste-f dataset against the hierarchical stacking of different model modules.

Class	Baseline	IoU (CE+Dice)	IoU (Prev.+ Dilated Conv)	IoU (Prev.+ MSFN)	IoU (Prev.+ F.Connect)
Rigid Plastic	27.58	27.72	28.06	30.67	**34.88**
Soft Plastic	61.86	70.16	71.65	84.88	**88.47**
Cardboard	51.36	80.93	88.83	89.47	**92.12**
Metal	3.41	4.5	4.64	5.83	**10.88**

Table 3: Segmentation results on the validation and test sets of ZeroWaste-f dataset.

	Validation		**Test**	
Methods	mIoU	Acc	mIoU	Acc
EPS(Lee et al., 2021)	13.75	59.96	13.91	60.65
CCT(Ouali et al., 2020)	30.79	84.80	29.32	85.91
Deeplab(Chen et al., 2016)	46.0	87.0	52.30	84.79
ReCo(Liu et al., 2022a)	49.49	89.58	44.12	88.36
InternImage(Wang et al., 2022b)	50.20	90.00	50.40	89.80
ZegClip(Zhou et al., 2023)	50.50	90.10	51.05	90.20
Ours model	**52.80**	**91.63**	**53.95**	**91.30**

Table 4: A Comparison of state-of-the-art detectors on the ZeroWaste-f test data set.

Models	**AP**	**AP50**	**AP75**	**APs**	**APm**	**APl**
Deformable Transformer (Zhu et al., 2021)	20.20	32.50	21.2	0.4	9.0	22.2
MaskRCNN (Dina Bashkirova and Saenko, 2022)	22.8	34.9	24.4	4.6	10.6	25.8
TridentN (Li et al., 2019)	24.2	36.3	26.6	4.7	10.8	26.05
Internimage (Wang et al., 2022a)	24.7	37.3	26.95	5.8	12.7	27.1
Ours model	**26.40**	**38.05**	**27.70**	**8.1**	**16.89**	**30.90**

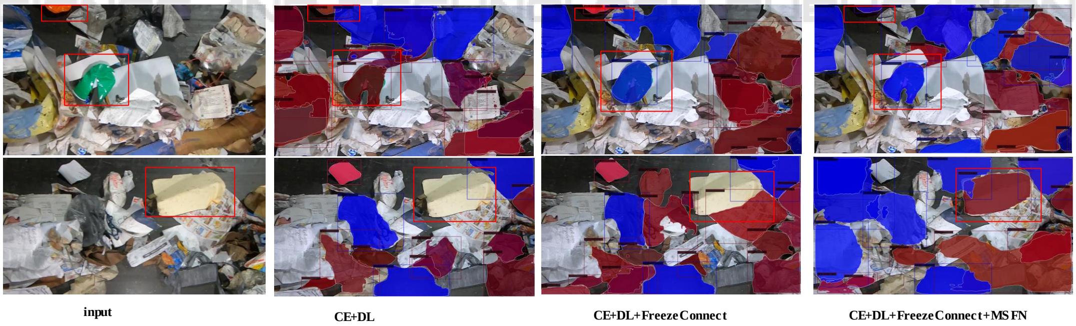


Figure 5: Random samples of the segmentation and detection results on the ZeroWaste-f Dataset. Figure shows input image with application of different modules for accurate and multiscale segmentation and detection.

Test Results Comparison

We evaluated the performance of our proposed model using the basic setup described in section 4.1. As shown in Table 3, our model achieved the best accuracy among all models tested on this dataset. Specifically, our model achieved a validation score of 52.80 %, which is 2.3% higher than the ZegClip score of validation score of 50.50. Similarly, the testing score is 2.90% higher with improved accuracy as compared to the deeplab. This improvement demonstrates the effectiveness of our proposed model in accurately segmenting waste objects in the given dataset. Segmentation results on the TrashCAN dataset as given in Table 6 in Appendix also validate the performance of our model. This shows the improved generalization of our method while further detail is given in the Appendix in section 4 where we give details of our assessment test.

5.3 Detection Task

As depicted in Table 4 for the detection task, our proposed model consistently outperforms the previously established state-of-the-art model on the ZeroWaste-f dataset. For TrashCAN results are given in the Appendix in section 3, where we have compared our results with EfficientDet (Majchrowska et al., 2022) and MaskRCNN(Dina Bashkirova and Saenko, 2022)

Test Results Comparison

Our experiments highlight the superior performance of our model when compared to other models in the field of object detection. Specifically, when employing Mask RCNN as our framework, our model showcased exceptional capabilities and outperformed the alternative approaches. Table 4 presents a comprehensive comparison of various performance metrics when benchmarked on ZeroWaste-f, illustrating the efficacy of our model in terms of accuracy, and precision. These findings validate the effectiveness and robustness of our proposed approach for object detection too. The TrashCAN dataset's comparative results are in the Appendix as Table 1. By leveraging the Mask RCNN framework, our model demonstrates a significant improvement in performance, providing accurate and reliable detection results. A few sample results are in Figure 5 while other results are in Appendix. In summary, our experiments clearly demonstrate that our model, in conjunction with the Mask RCNN framework, excels in object detection and underscores its potential for various practical applications where precise and reliable object detection is essential.

6 CONCLUSION

In this paper, we presented a unique computer vision-based network configuration for efficient segmentation of waste objects in cluttered and disorganized environments, tailoring the InternImage model for segmentation tasks. Our proposed approach incorporates a dilated convolution layer in the stem layer to learn spatially more complex representations and integrate information from wider receptive fields. Furthermore, we utilized a mutli-scale feedforward network and leveraged a combination of cross-entropy and Dice loss functions to enhance segmentation accuracy. We also enhanced our results by employing the benefits of Freezeconnect. We evaluated our model using the ZeroWaste-f dataset as well as the TrashCAN dataset, which contains overlapping objects in heavily cluttered scenes, achieving state-of-the-art results compared to the baseline model.

Our block configuration method enhances waste sorting and recycling by accurately separating recyclables in challenging conditions. This model can also be adapted for future waste management tasks, such as tracking waste flow. Its precise segmentation allows for the effective tracing of waste materials, facilitating better understanding of waste distribution and the optimization of processing routes.

REFERENCES

Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli, B., Bargal, S. A., and Saenko, K. (2021). Zerowaste dataset: Towards deformable object segmentation in extreme clutter.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2020). Attention augmented convolutional networks.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Contributors, M. (2020). MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dina Bashkirova, Mohamed Abdelfattah, Z. Z. J. A. F. A. P. H. V. A. B. C. S. A. B. and Saenko, K. (2022). Zerowaste dataset: Towards deformable object segmentation in cluttered scenes.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Fulton, M. S., Hong, J., and Sattar, J. (2020). Trash-ICRA19: A bounding box labeled dataset of underwater trash. Retrieved from the Data Repository for the University of Minnesota, https://doi.org/10.13020/x0qn-y082.

Gundupalli, S. P., Hait, S., and Thakur, A. (2017). A review on automated sorting of source-separated municipal solid waste for recycling. *Waste Management*, 60:56–74. Special Thematic Issue: Urban Mining and Circular Economy.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hong, J., Fulton, M., and Sattar, J. (2020). Trashcan: A semantically-segmented dataset towards visual detection of marine debris.

Huang, Y., Kang, D., Jia, W., Liu, L., and He, X. (2022). Channelized axial attention–considering channel relation within spatial attention for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1016–1025.

Jo, S. and Yu, I. (2021). Puzzle-cam: Improved localization via matching partial and full features. *CoRR*, abs/2101.11253.

Jung, H. and Oh, Y. (2021). Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1336–1344.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Lee, S., Lee, M., Lee, J., and Shim, H. (2021). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505.

Li, Y., Chen, Y., Wang, N., and Zhang, Z. (2019). Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6054–6063.

Liu, S., Zhi, S., Johns, E., and Davison, A. J. (2021a). Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*.

Liu, S., Zhi, S., Johns, E., and Davison, A. J. (2022a). Bootstrapping semantic segmentation with regional contrast.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. (2022b). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows.

Majchrowska, S., Mikołajczyk, A., Ferlin, M., Klawikowska, Z., Plantykow, M. A., Kwasigroch, A., and Majek, K. (2022). Deep learning-based waste detection in natural and urban environments. *Waste Management*, 138:274–284.

Mao, A., Mohri, M., and Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*.

"Minaee, S., "Boykov, Y. Y., "Porikli, F., "Plaza, A. J., "Kehtarnavaz, N., and "Terzopoulos, D. (2021). Image segmentation using deep learning: A survey.

Ouali, Y., Hudelot, C., and Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12671–12681.

Pathak, D., Krähenbühl, P., and Darrell, T. (2015). Constrained convolutional neural networks for weakly supervised segmentation.

Proença, P. F. and Simões, P. (2020). Taco: Trash annotations in context for litter detection.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., and Shlens, J. (2021). Scaling local self-attention for parameter efficient visual backbones.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: Self-attention with linear complexity.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. (2022a). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al. (2022b). Internimage: Exploring large-scale vision foundation models with deformable convolutions.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., and Qiao, Y. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2022c). PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424.

Wu, H., Zhang, J., Huang, K., Liang, K., and Yu, Y. (2019). Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. (2018). Unified perceptual parsing for scene understanding.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Zamir, S. W., Arora, A., Khan, S. H., Hayat, M., Khan, F. S., and Yang, M. (2021). Restormer: Efficient transformer for high-resolution image restoration. *CoRR*, abs/2111.09881.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113.

Zhou, Z., Lei, Y., Zhang, B., Liu, L., and Liu, Y. (2023). Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection.

APPENDIX

Hyper Parameters for Weight Loss

Different hyperparameters for loss weight give us different mIoU values. We present these results in Figure 6.

Overall Pipeline

The complete pipeline can be seen as a block diagram as in Figure 7. The images from benchmarked datasets are passed through segmentation and detection frameworks using our model. These backbones are pre-trained with classification weights.

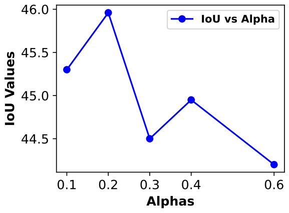


Figure 6: Variation in mIoU Values with Different Loss Weight Hyperparameters.

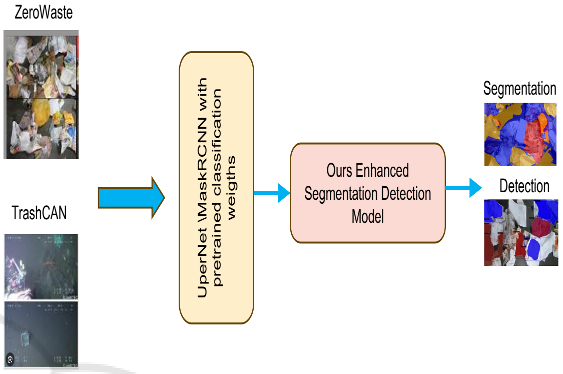

ZeroWaste
TrashCAN
UperNet \MaskRCNN with pretrained classification weigths
Ours Enhanced Segmentation Detection Model
Segmentation
Detection

Figure 7: Pipeline Overview: This figure displays the flow from input image through backbone stages to the segmentation and detection models, culminating in the output, showcasing the process at a block level.

Segmentation and Detection Results on TrashCAN

Detection results for TrashCAN dataset are given in Table 5 Where we compare our model's performance against other models on TrashCAN. We further display results of our model in Figures

respectively. These indicate the performance of our model for both of these tasks. We evaluated the performance of our model on the TrashCAN dataset, which is a challenging dataset for waste paper segmentation and detection. The results are shown in Table 5, where we compare our model's performance against other state-of-the-art models. Our model achieved the best performance on all metrics, including mean intersection over union (mIoU) and precision.

We also visualized the detection and segmentation results of our model in Figures 8 . Figure 1 shows the segmentation results of our model on a sample image from the TrashCAN dataset. The different colors in the image represent different waste paper categories. Figure 2 shows the detection results of our model on the same image. The bounding boxes in the image show the location of the detected waste paper objects.

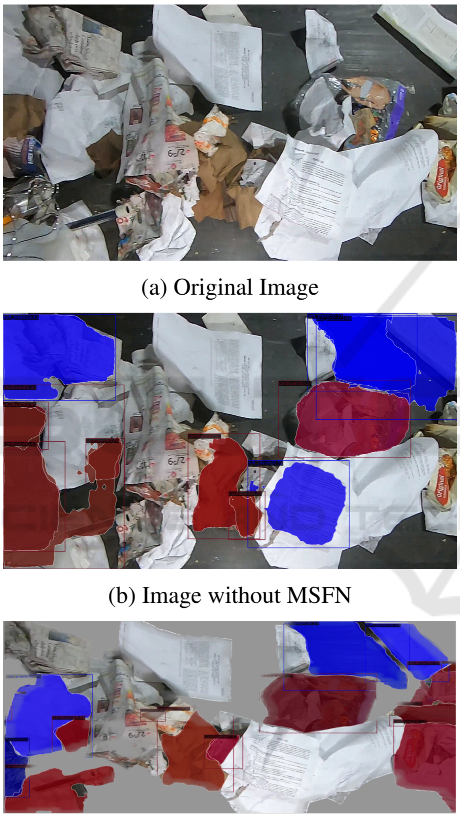
These results indicate that our model is able to

Table 5: Comparison of state-of-the-art detectors on the TrashCAN dataset.

Models	AP	AP50	AP75	APs	APm	APl
EfficientDet	29.1	51.21	87.83	28.22	30.20	40.0
MaskRCNN	30.0	55.3	29.4	23.2	31.7	48.6
Internimage	30.80	54.3	31.4	26.2	32.7	47.6
Our model	**32.91**	**55.06**	**36.70**	**33.09**	**34.00**	**49.51**

Table 6: Segmentation results on the TrashCAN.

Methods	mIoU	Acc
EfficientDet	45.39	80.50
InternImage (Wang et al., 2022b)	47.92	81.85
Ours model	**50.98**	**85.28**



(a) Original Image

(b) Image without MSFN

(c) Image WITH MSFN + Freezeconnect

Figure 8: Sample Images with Ablation Effects.

accurately segment and detect waste paper objects in cluttered environments. This makes our model a promising candidate for applications such as waste recycling and waste sorting.

Robustness Evaluation

One of the critical aspects of assessing the robustness of any machine learning model, especially in the fields of segmentation and object detection, is conducting evaluations across diverse datasets. These datasets often contain varying scenarios, environmental conditions, and object classes. By subjecting your model to different datasets, you aim to ensure that it can generalize effectively and provide reliable results in a wide range of real-world situations.

To gain a comprehensive understanding of our model's robustness and effectiveness, we conducted a thorough comparison with various other models commonly used in segmentation and object detection tasks. We also perform statistical testing on the basis of mIoU values with the null -hypothesis of the mean of mIoU values of our model and the interchange is the same. We computed 5 values for both models and then computed the result based on p-value and t-statistic and we found out that the importance of p is 0.067 with a t-statistic of 2.18. As the value of p is more than 0.067 it almost validates our hypothesis, further, our findings indicate that, within the context of Waste datasets, there is significant potential for our model's utility. This implies that continued refinement and development of our model could yield substantial benefits for waste management applications, thereby contributing to the improved results and practical significance of our research.