

COSNet: A Novel Semantic Segmentation Network using Enhanced Boundaries in Cluttered Scenes

Muhammad Ali¹ Mamoon Javaid² Mubashir Noman¹ Mustansar Fiaz³ Salman Khan¹

¹Mohamed bin Zayed University of AI, U.A.E ²Institute of Space Technology, Pakistan

³IBM Research

Abstract

Automated waste recycling aims to efficiently separate the recyclable objects from the waste by employing vision-based systems. However, the presence of varying shaped objects having different material types makes it a challenging problem, especially in cluttered environments. Existing segmentation methods perform reasonably on many semantic segmentation datasets by employing multi-contextual representations, however, their performance is degraded when utilized for waste object segmentation in cluttered scenarios. In addition, plastic objects further increase the complexity of the problem due to their translucent nature. To address these limitations, we introduce an efficacious segmentation network, named COSNet, that uses boundary cues along with multi-contextual information to accurately segment the objects in cluttered scenes. COSNet introduces novel components including feature sharpening block (FSB) and boundary enhancement module (BEM) for enhancing the features and highlighting the boundary information of irregular waste objects in cluttered environment. Extensive experiments on three challenging datasets including ZeroWaste-f [4], SpectralWaste [5], and ADE20K [50] demonstrate the effectiveness of the proposed method. Our COSNet achieves a significant gain of 1.8% on ZeroWaste-f and 2.1% on SpectralWaste datasets respectively in terms of mIoU metric. Source code is available at <https://github.com/techmn/cosnet>.

1. Introduction

With the upsurge of world population and urbanization, the production of waste is rising which may encounter detrimental effects on the environment [20]. The effective recycling of the waste objects such as plastic, metal, paper, etc., is therefore a critical task and requires careful handling procedures [4]. Although material recycling facilities (MRFs) or plants that separate and prepare recyclable materials utilize heavy machinery for recycling of waste ma-

terial. However, the manual labor may risk their health by getting exposed to the bare metal and unhygienic medical waste [4, 14]. It is therefore necessary to develop an automated system that can effectually separate the recyclable objects from the waste material. Furthermore, the translucent nature of the plastic objects and the presence of a cluttered background make the automated separation of the recyclable waste challenging.

In literature, various waste classification methods have been proposed [13, 28, 29, 40], they pose limitations in the presence of complex scenarios where there exists an unclear boundary information. Recently, Bashkirova et. al. [4] and Casao et al. [5] introduced challenging datasets, called ZeroWaste and SpectralWaste, for the task of waste object segmentation in a cluttered scenes, respectively. Current deep learning approaches [3, 7, 8, 25, 47, 49] have shown promising results on semantic segmentation task by utilizing the multi-scale contextual features in an encoder-decoder architecture. For instance, SegNet [3] introduces the encoder-decoder framework to benefit from both the fine grained boundary information of shallow layers as well as the rich semantic features of last layers, DeepLabv3 [7] exploits the atrous convolutions to capture multi-scale information for better segmentation performance, PSPNet [49] uses the pyramid pooling module to capture the multi-contextual representations, and DeepLabv3+ [8] further optimizes the method by using encoder-decoder framework and depth-wise separable convolutions. Furthermore, Xie et. al. [42] discuss the importance of boundary cues for accurate segmentation of transparent objects. Despite of utilizing the encoder-decoder architecture [3, 7] for capturing rich contextual information by pooling features and exploiting the fine detail information from shallow layers, the existing well-proven methods such as [8] still strive to accurately segment the waste objects in cluttered scenes. Conversely, transformer-based methods [25, 47] have shown remarkable performance for segmentation tasks due to their ability to capture global contextual relationships. However, these methods rely on a large volume of annotated data which

can be expensive and may require strenuous efforts especially for specialized applications [4]. Additionally, the dominance of global representations may limit the ability of transformers to focus on fine details that are necessary to segment the translucent and deformable-shaped objects in cluttered scenes. Therefore, an explicit mechanism is required to highlight the boundaries of such objects for better segmentation results.

To address the limitations of the aforementioned methods, we propose an effectual segmentation network, called COSNet (Cluttered Objects' semantic Segmentation Network), that highlights the boundary cues and utilizes the multi-scale information in an encoder-decoder architecture for better segmentation of waste objects in cluttered scenes. Inspired from the unsharp masking [30], we propose a feature sharpening block (FSB) that benefits from both enlarged receptive field and implicit feature boosting techniques. Specifically, we capture the multi-scale contextual representations using atrous convolutions to enlarge the receptive field and enhance the boundary details by exploiting an implicit sharpening module within our FSB module. Furthermore, we complement our COSNet by exploring explicit high-boost filtering. To achieve this, we propose a boundary enhancement module (BEM), as an intermediate network, to supplement the COSNet in obtaining finer feature representations. To summarize, our contributions are as below:

- We propose an effective backbone network that effectively utilizes the multi-scale feature representations and highlights the boundary information to accurately segment the waste material such as plastic objects in a cluttered background.
- We propose BEM that is utilized in an encoder-decoder segmentation framework for further boosting the feature representations obtained from the backbone network.
- Extensive experimentation on three challenging semantic segmentation datasets (ZeroWaste [4], SpectralWaste [5] and ADE20K [50]) reveal the merits of the proposed contributions.

2. Related Work

2.1. Waste Detection and Segmentation

Waste detection has recently gained immense popularity due to various challenges including cluttered backgrounds, translucent objects, illumination variations, and objects' varying shapes and sizes. Since the advent of deep learning, various convolutional neural network (CNNs) based methods have been introduced for waste sorting [13, 16, 28, 29, 33, 35, 40]. Qin et al. [33] proposed a garbage classification

framework by introducing a small amount of background and noise to alleviate the overfitting and improve the robustness. Xia et al. [40] exploit MobileViTv3 [36] features to capture the global and local features for garbage detection. Mao et al. [28] utilize DenseNet121 [19] to introduce a genetic algorithm that fine-tunes the parameters of the fully connected layers of the model. Feng et al. [13] used the EfficientNet [22] as a backbone feature extractor and introduced the efficient channel attention (ECA) and coordinate attention (CA) modules to capture the waste objects. Meng et al. [29] employ FPN along with a focal loss to reduce the imbalance between foreground and background samples to enhance the detector effect. Tian et al. [35] exploit the MobileNetV3 [18] and introduce the CBAM [39] to enhance the spatial features for garbage recognition. Hossein et al. [16] introduce a recyclable waste classification (RWC-Net) to classify the wastes.

Though these above-mentioned waste classification methods demonstrate great performance, these approaches present limitations in various scenarios including a wide range of waste types and unstructured backgrounds. Various efforts have been made to segment the waste from the background [1, 4–6, 10, 32, 34, 48]. Bashkirova et al. [4] utilizes DeepLabv3+ as backbone features and exploits atrous spatial pyramid pooling (ASPP) to encode the multi-scale semantic features for segmentation maps. Qi et al. [32] uses non-uniform data augmentation tailored to real waste classification scenarios. Yudin et al. [48] propose a hierarchical neural network, called H-YC, that can perform waste classification and segmentation. Corrigan et al. [10] propose a method to detect underwater litter using Mask RCNN [15]. Casao et al. [5] benefits from RGB and HSI modalities for waste object segmentation. Cascina et al. [6] utilize different architectures for constrained semantic segmentation models for segmenting recyclable waste. Rahman et al. [34] employ generalized semantic segmentation techniques over the kitchen food waste. Ali et al. [1] introduce a feature amplification network (FANet) that is capable of capturing the semantic information at multi-stages using a feature enhancement module. Existing methods still struggle to localize the objects due to cluttered backgrounds and unclear boundaries between the waste objects and backgrounds.

2.2. Semantic Segmentation

Recently, deep learning techniques have shown promising performance over semantic segmentation tasks [2, 8, 17, 23, 25, 26, 37, 43, 44, 47]. Chen et al. [8] extended the DeepLabv3 [7] by utilizing depthwise separable convolution to both Atrous Spatial Pyramid Pooling and decoder modules. MiniNet-v2 [2] is built on ConvNext for efficient semantic segmentation for real-time robotic applications. Wang et al. [37] introduce the deformable convolution as the core operator to enhance the receptive field for semantic

segmentation. Lin et al. [23] propose a multi-head mixed convolution (MHMC) operation along with a scale-aware aggregation (SAA) module to exploit multi-scale context features. Hou et al. [17] leverage the convolutional modulation by utilizing a large kernel to enlarge the receptive field. Xie et al. [43] propose Segformer which exploits transformers with lightweight multilayer perception (MLP) decoders combining both local attention and global attention to render powerful representations. Liu et al. [26] reexamine the design spaces to test the limits of what a pure ConvNet is and propose ConvNext. Yang et al. [44] propose a FocalNet that modulates the features at different granularities to exploit features from multiple scales. Fan et al. [12] redesigned the BiSeNet [46] and propose a detail aggregation module that utilizes the static Laplacian filter to highlight the boundaries of the objects for better segmentation performance. Similarly, Kim et al. [21] introduce B2Inst model for instance segmentation task that uses static Laplacian kernel to extract the boundaries of the predicted and ground truth masks. Afterwards, boundary loss is utilized to learn the boundaries of the object instances.

In contrast, we propose a multi-contextual features extraction and sharpening module that strives to capture implicit local information and sharpening information to handle complex shapes and sizes of waste objects. Furthermore, our explicit boundary enhancement module learns salient features to handle the background clutter and delineates the translucent waste objects from backgrounds.

3. Background

In image processing, unsharp masking [30] is a common technique to enhance the edges and details of the objects in an image. The unsharp masking is a filter that emphasizes the high-frequency content. In the unsharp masking technique, a blurred (unsharp) version of the image is generated and subtracted from the original input image. The output, called the mask, focuses on the edges and fine details of the image. Later the mask is added to the original image which results in a sharpened image emphasizing the texture and details of the image.

Let $f(x, y)$ and $\bar{f}(x, y)$ be the input and its blurred version in latent space, and the mask $g_{mask}(x, y)$ can be obtained using equations as:

$$g_{mask}(x, y) = f(x, y) - \bar{f}(x, y). \quad (1)$$

Finally, the sharpened feature map $g(x, y)$ can be obtained follows as:

$$g(x, y) = f(x, y) + k * g_{mask}(x, y), \quad (2)$$

where $k \in (k \geq 0)$ represents the weight. When $k == 1$ we have unsharp masking and when $k > 1$ the process is referred to as high boost filtering. Motivated by this, we

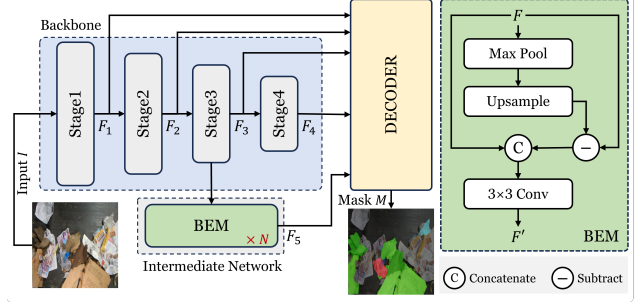


Figure 1. Illustration of the overall architecture of the proposed segmentation framework named COSNet. The proposed framework utilizes enhanced backbone network to obtain rich multi-scale representations through FSBs (sec. 4.2.1) of the backbone network. The intermediate network further enhances the boundary details of the third-stage features by means of a boundary enhancement module (BEM). Finally, the multi-scale feature maps, F_i where $i \in 1, 2, 3, 4, 5$, are passed to the decoder to obtain the segmentation mask M .

leverage our sharpening module (SM) and boundary enhancement module (BEM) as described in Sec. 4.2.1 and 4.3, respectively.

4. Method

In this section, we describe the proposed approach in detail.

4.1. Overall Architecture

In Fig. 1, we illustrate the overall architecture of the proposed segmentation framework named COSNet. The proposed framework consists of a backbone network, an intermediate network, and a decoder. The backbone network consists of four stages and utilizes FSBs to provide the multi-scale feature representations. The intermediate network uses the BEM to further enrich the feature maps of the third stage. The utilization of intermediate network is task-relevant such as further enhancement of object boundaries in the case of translucent objects and may be replaced by the identity module. Finally, the feature representations from the backbone and intermediate networks are fed to a decoder that effectively fuses the multi-scale feature representations to obtain segmentation masks.

4.2. COSNet Backbone

The backbone network comprises of four stages to obtain the multi-scale feature representations F_i where $i \in 1, 2, 3, 4$ as shown in Fig. 2. Each stage utilizes a series of feature sharpening blocks (FSBs) to obtain rich feature representations. The COSNet backbone takes an input image I and feeds it to a convolution stem layer that projects the image to feature space and downsamples the spatial resolution

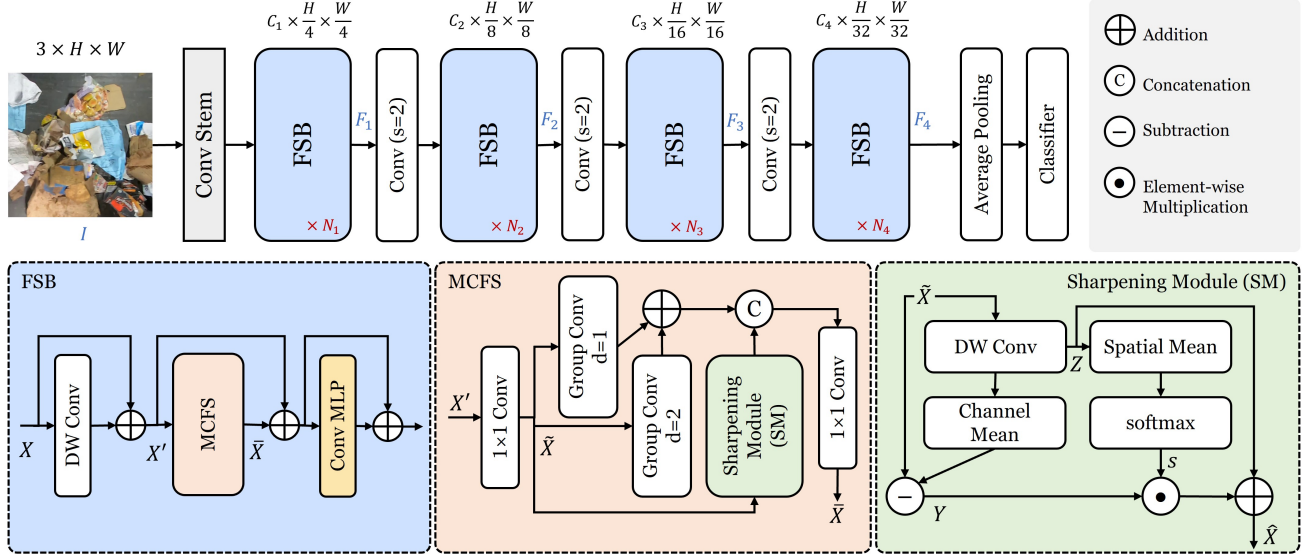


Figure 2. Illustration of the backbone network of the proposed COSNet. The backbone network extracts multi-scale features at four scale levels by utilizing feature sharpening blocks (FSBs). The core component of FSB is multi-contextual features extraction and sharpening (MCFS) module that uses dilated convolutions to extract multi-contextual representations and a learnable sharpening module to enhance the boundary information.

by four times. The last three stages use a convolution layer with a stride value of two to reduce the spatial resolution of the feature maps before passing them to feature sharpening blocks.

4.2.1 Feature Sharpening Block (FSB)

The design of the FSB is illustrated in Fig. 2. The FSB is composed of a depth-wise convolution layer, a multi-contextual feature sharpening module (MCFS) and a convolutional multi-layer perceptron (MLP). Being a key module of the FSB, MCFS is responsible for capturing the multi-contextual information while focusing on the object boundaries for obtaining the better segmentation maps. Inspired by [7], we utilize atrous convolutions to capture multi-contextual features. Although multi-contextual information plays a significant role in better segmentation performance [7, 8], however, it struggles to highlight boundary cues for translucent and cluttered objects. For this reason, we introduce an implicit feature sharpening module (SM) that highlights the boundary information of objects in cluttered scenarios.

Given feature maps X' , MCFS first projects the features by using a 1×1 convolution layer to obtain features $\tilde{X} \in \mathbb{R}^{\tilde{C} \times \tilde{H} \times \tilde{W}}$. Then, we utilize two group convolution layers having different dilation values to extract multi-contextual features. In parallel, feature maps $\tilde{X}$ are passed to sharpening module (SM) to get the boundary enhanced features. Lastly, we concatenate the multi-contextual fea-

tures and the boundary enhanced features along channel dimension and utilize a 1×1 convolution layer to get enhanced feature representations.

The sharpening module (SM) first utilizes a depth-wise convolution layer to encode local representations and separately enhance the structural information in each channel and obtain features Z . Afterwards, we compute mean of the feature maps (Z) along channel dimension to obtain the averaged response of feature maps. To highlight the fine details, we subtract the averaged response from input features $\tilde{X}$ to get features Y . Additionally, we compute the spatial mean of each channel of features Z and apply softmax operation to obtain a sharpening factor ($s \in \mathbb{R}^{\tilde{C} \times 1 \times 1}$) for each channel of feature maps Y . We multiply the sharpening factors s with the features Y and add it to the Z to obtain the sharpened features $\hat{X}$.

4.3. Boundary Enhancement Module (BEM)

Our explicit boundary enhancement module is utilized to further enhance the feature maps of the third stage of the backbone network as shown in Fig. 1. The third stage features corresponds to high level semantics as well as preserves the fine details of the regions. The idea is to focus on the high frequency details of the feature maps similar to the process of highlighting the boundaries by subtracting the blurred image from the original one [30]. BEM takes input feature maps F and utilize max pooling operation to extract the significant features. The pooled features have smaller spatial resolution compared to the input features F . There-

Table 1. Comparison of our framework with state-of-the-art methods for semantic segmentation on the test set of the ZeroWaste-f dataset. We report the results in terms of mIoU and pixel accuracy. The best results are highlighted in bold text.

Method	mIoU (%) $\uparrow$	Pix. Acc. (%) $\uparrow$
CCT [31]	29.32	85.91
ReCo [24]	52.28	89.33
DeepLabv3+ [8]	52.13	91.38
FANet [1]	54.89	91.41
InternImage-T [38]	54.33	91.84
COSNet (Ours)	56.67	91.91

Table 2. Class-wise IoU (%) score of our COSNet framework with DeepLabv3+ on the test set of the ZeroWaste-f dataset. The proposed COSNet achieves significant gains in IoU score for all classes. The best results are highlighted in bold text.

	DeepLabv3+	Ours
Background	91.02	91.44
Cardboard	54.47	59.13
Soft Plastic	63.18	65.92
Rigid Plastic	24.82	37.24
Metal	27.14	29.61

fore, pooled features are upsampled back to the input spatial resolution. To highlight the fine details, we perform the subtraction operation between the input F and upsampled features. The highlighted features are concatenated with input features F and a 3×3 convolution layer is utilized to obtain the enhanced feature maps F' .

5. Experiments

Here we discuss the datasets, implementation details, quantitative, and qualitative comparison of our method with state-of-the-art methods.

5.1. Datasets and Evaluation Metrics

To evaluate the performance of proposed framework on cluttered scenes, we utilize ZeroWaste-f [4] and SpectralWaste [5] datasets. Additionally, we report the performance of our framework on a well known public semantic segmentation dataset ADE20K [50].

ZeroWaste-f [4] is a challenging waste segmentation dataset that contains object classes of four material types including cardboard, soft plastic, rigid plastic, and metal. The dataset comprises of 3002, 572, and 929 images in train, validation, and test sets respectively.

SpectralWaste [5] is another recently introduced dataset which has a cluttered background and translucent waste objects. It is collected from a waste processing facility and contains synchronized RGB and hyperspectral images. The dataset comprises of six object classes containing cardboard, film, basket, filaments, video tape, and trash bags.

The dataset is challenging due to the presence of clutter and brightness variations. SpectralWaste dataset contains 514, 167, and 171 images in train, validation, and test sets respectively. In our experiments, we only utilize the RGB images of the dataset.

ADE20K [50] is a well known public semantic segmentation dataset comprising of 150 object classes in diverse scenarios. The dataset contains 20210 training and 2000 validation images.

For performance evaluation, we report the *mean intersection over union* (mIoU) and *pixel accuracy* score on the respective datasets.

5.2. Implementation Details

The proposed method is implemented in Pytorch by utilizing MMSegmentation library [9]. We utilize NVIDIA RTX A6000 GPU for experimentation. In training phase, we initialize the backbone with ImageNet [11] pretrained weights and optimize the model by using AdamW [27] optimizer with initial learning rate of $9e-5$. The learning rate is decayed by utilizing polynomial scheduler. We train the model for 40k iterations for the ZeroWaste-f [4] and SpectralWaste [5] datasets, and 160k iterations for ADE20K [50] dataset. The shorter side of images are randomly resized to 512 pixels and we utilize random crop of 512×512 pixels during training. We train the COSNet backbone on ImageNet1k [11] for 300 epochs. For each segmentation experiment, we load the ImageNet1k pre-trained weights in the backbone and utilize the UperNet [41] decoder to obtain a segmentation mask.

5.3. Quantitative Results

5.3.1 ZeroWaste-f [4]

We report the semantic segmentation performance of COSNet on ZeroWaste-f dataset in Tab. 1. Among existing methods, DeepLabv3+ [8] provides promising results by achieving the mIoU score of 52.13% while FANet [1] achieves the mIoU score of 54.89%. However, our proposed COSNet performs favorably and sets new state-of-the-art score of 56.67% in terms of mIoU metric.

We further compare the class-wise IoU score of the existing best DeepLabv3+ (as reported in [4]) with our COSNet in Tab. 2. We notice that there is a significant improvement in IoU for all class objects especially rigid plastic. The reason is that COSNet enhances the boundaries of the objects that supplement the model’s performance to better discriminate the objects in cluttered scenes.

5.3.2 SpectralWaste [5]

Tab. 3 compare the performance of COSNet with the existing state-of-the-art waste segmentation methods as well

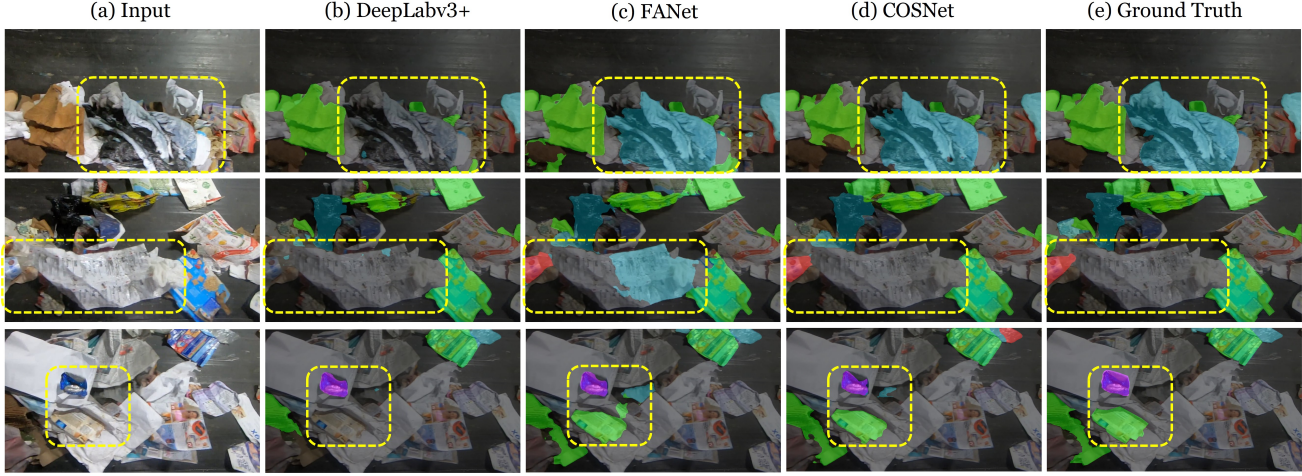


Figure 3. Here we present the segmentation results of our COSNet on ZeroWaste-f [4] dataset. Our method can accurately segment the different waste material types compared to the DeepLabv3+ [8] and recently introduced FANet [1] as highlighted in yellow boxes.

Table 3. Comparison of our framework with state-of-the-art methods for semantic segmentation on the test set of the SpectralWaste [5] dataset. We report the results in terms of mIoU and individual class IoUs. The best results are highlighted in bold text. The * represents that the reported results are from SpectralWaste [5] having smaller backbone networks.

Method	mIoU (%) $\uparrow$	IoU per Class (%) $\uparrow$					
		Film	Basket	Cardboard	Video Tape	Filament	Trash Bag
MiniNet-v2* [2]	44.5	63.1	58.9	55.4	30.6	10.0	49.2
SegFormer-B0* [43]	48.4	66.9	71.3	48.9	33.6	15.2	54.6
FANet [1]	67.83	72.47	82.98	75.26	41.28	67.65	67.36
InternImage-T [38]	47.99	42.38	82.8	69.1	41.39	16.5	35.77
COSNet (ours)	69.96	77.61	83.65	75.14	42.95	69.06	71.38

as the methods reported in [5]. First two rows of the table demonstrate the performance of the networks that utilize shallower backbones as reported in [5]. For fair comparison, we compare the results of COSNet with the recent waste segmentation method FANet [1]. As illustrated from the table, proposed COSNet achieves a significant gain of 2.1% in terms of mIoU metric. In case of class-wise comparison, FANet [1] has slightly better IoU score of 75.3% for *Cardboard* objects compared to 75.1%. However, COSNet surpasses the FANet [1] for the other five classes by obtaining significant gains of approximately 4.02% and 5.14% for *Trash Bag* and *Film* object classes, respectively.

5.3.3 ADE20K [50]

Tab. 4 presents the comparison of our COSNet with the existing well-established backbone networks on ADE20K [50] dataset. Despite having *less number of parameters*, our COSNet outperforms the existing methods and achieves a gain of 0.3% in terms of mIoU metric. We further compare the average latency on the test set of ADE20k dataset using crop size of 512×512 on Nvidia A100 GPU. These results

further validate the effectiveness of FSB for semantic segmentation task.

5.4. Qualitative Results

We present the qualitative performance of our framework in Fig. 3, 4 and 7 on ZeroWaste-f, SpectralWaste and ADE20K datasets respectively. Fig. 3 shows that the proposed COSNet can effectively segment different waste material types such as soft plastic, rigid plastic, metal, and cardboard compared to the existing methods including DeepLabv3+ [8] and FANet [1]. Although, segmentation results of FANet [1] seems to be promising, however, it introduces more false segmentation as highlighted in the second and third rows of the figure.

Similarly, for the challenging scenarios of SpectralWaste [5] dataset, we show that the proposed COSNet can better segment various types of waste objects in Fig. 4 as compared to recently introduced FANet [1]. In the first and second row, our COSNet detects the trash bag and video tape objects which are not detected by FANet [1]. Likewise, proposed COSNet accurately detects the cardboard objects in the third row of Fig. 4 as highlighted in yellow boxes.

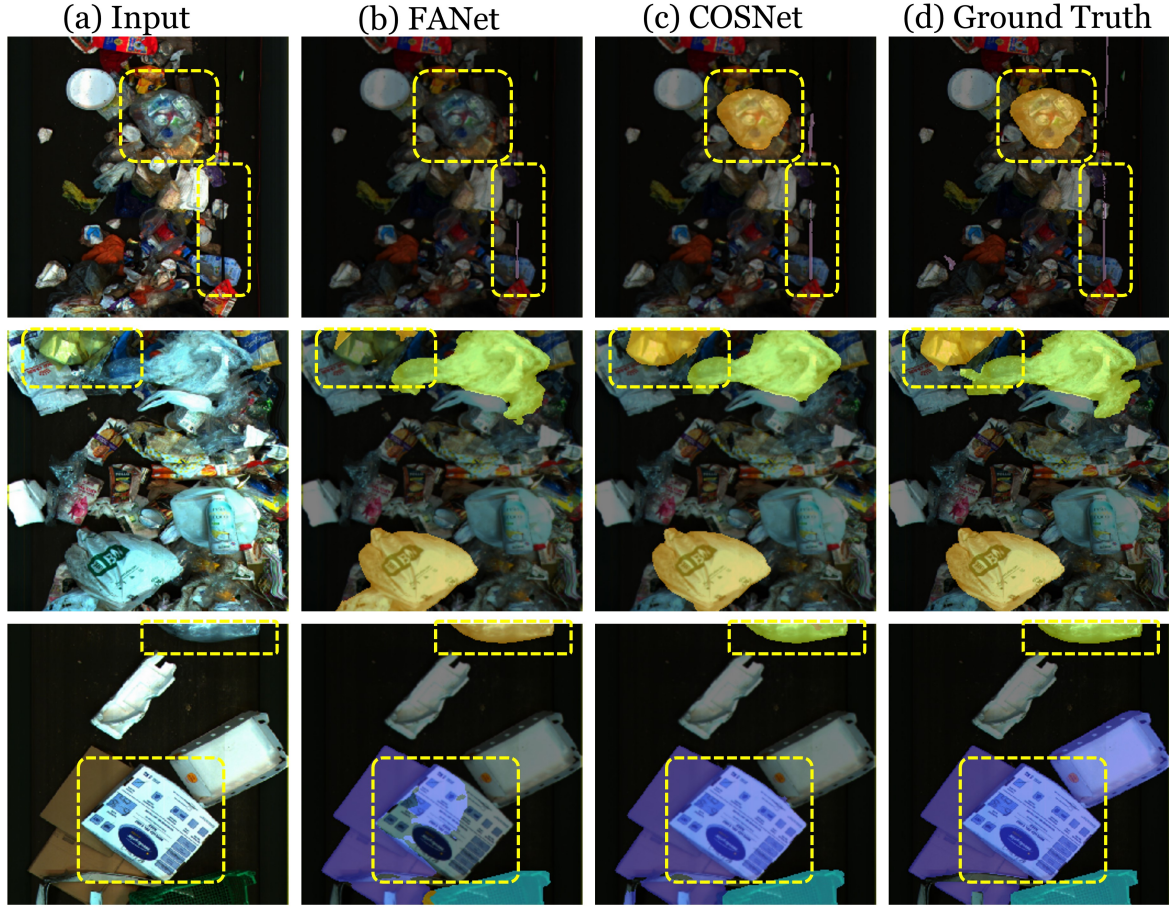


Figure 4. Comparison of the segmentation results of proposed COSNet on Spectral Waste (RGB images) dataset. The proposed COSNet can accurately segment the different waste material types compared to the FANet as highlighted in yellow boxes.

Table 4. Comparison of COSNet with state-of-the-art methods on ADE20K [50] dataset. We report the results in terms of mIoU metric. Here, SS refers to single scale and MS refers to multi-scale evaluation. The best results are highlighted in bold.

Method	Params (M)↓	Avg. Latency (ms)↓	mIoU(SS/MS) (%)↑
Swin-T [25]	60	90	44.5/45.8
FocalNet-T [44]	61	88	46.5/47.2
ConvNext-T [26]	60	76	46.7/-
FocalAtt-T [45]	62	-	45.8/47.0
InternImage-T [38]	59	105	47.9/48.1
COSNet (ours)	37	85	47.3/ 48.4

While Fig. 3 and 4 present the challenging scenarios in cluttered background, Fig. 7 illustrate the diverse scenes of ADE20k dataset having greater number of class object. Nevertheless, our COSNet performs favorably and is able to segment the objects with clear boundaries.

5.5. Ablation Experiments

We present the effectiveness of proposed contributions on ZeroWaste-f [4] dataset in Tab. 5. In the first row of Tab. 5, MCFS, SM, and BEM modules are absent from the

Table 5. Ablation study of enhancement modules in COSNet on ZeroWaste-f dataset. We notice that utilization of the sharpening module (SM) in MCFS and BEM in COSNet provides the best performance highlighted in bold text.

	Exp. No.	MCFS	SM	BEM	mIoU (%)↑	Pix. Acc. (%)↑
COSNet	1	×	×	×	37.48	88.29
	2	✓	×	×	46.31	90.44
	3	✓	✓	×	52.38	91.29
	4	✓	✓	✓	56.67	91.91

network. As a result, the performance of the network is significantly reduced to 37.48%.

Adding the MCFS component in the network while removing the SM provides reasonably better mIoU score of 46.31%. Correspondingly, introducing the SM in the MCFS further improves the performance of the network by 6.1%. Finally, the addition of the BEM module as an intermediate network in COSNet further enhances the performance and achieves the state-of-the-art result of 56.67% in terms of mIoU score. It is evident from the Tab. 5 that boundary information improves the performance of the COSNet

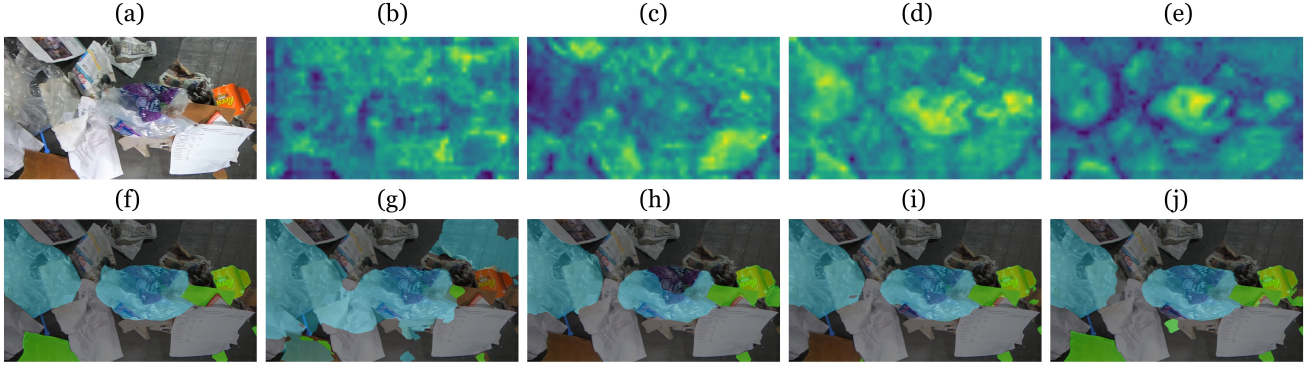


Figure 5. Demonstration of the proposed modules of COSNet on ZeroWaste-f [4] dataset. (a) is the input image, (b)-(e) refer to features of the third stage for the rows (1-4) in Tab. 5, respectively. (f) is the ground truth mask, (g)-(j) denote the corresponding prediction masks for rows (1-4) in Tab. 5, respectively. The above results clearly indicate the efficacy of the proposed modules when integrated in the network.

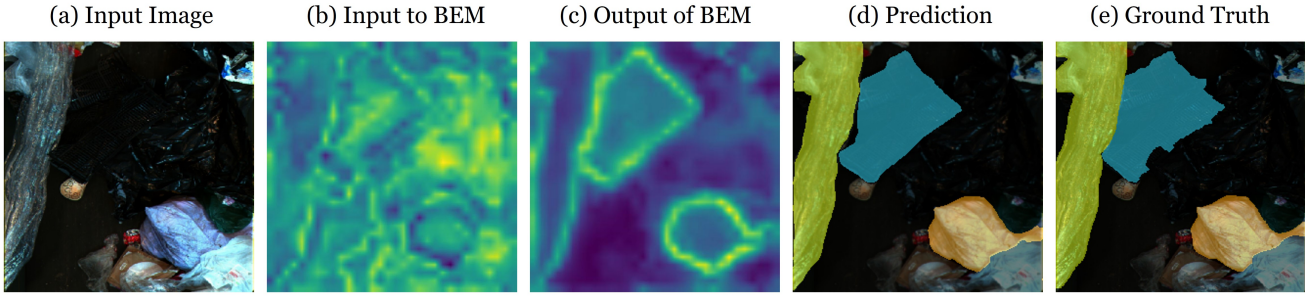


Figure 6. Demonstration of the enhancement capabilities of proposed BEM on SpectralWaste [5] dataset. (a) is the input image to the model, (b) shows the feature map visualization of the third stage of backbone network that is input to BEM, (c) shows enhanced feature maps using BEM highlighting the boundaries of the segmented objects, (d) presents the predicted segmentation mask, and (e) is the ground truth mask.



Figure 7. Segmentation results of COSNet on ADE20K validation set. We notice that our framework can effectively segment the objects by using boundary information.

for the segmentation task which validates the efficacy of the proposed contributions.

In addition to the quantitative results reported in Tab. 5, we demonstrate the effectiveness of the proposed modules in Fig. 5 and 6 respectively. It is evident from the Fig. 5 that the introduction of MCFS, SM, and BEM modules in the network improves the segmentation capabilities of the

COSNet. Similarly, as depicted in Fig. 6, the BEM highlights the boundary information and enhances the features of the third stage by suppressing the unnecessary information. To summarize, COSNet effectively utilize the boundary cues for better segmentation of waste objects in cluttered scenarios.

6. Conclusion

In this work, we introduce a semantic segmentation framework named COSNet that utilizes the multi-contextual representations and boundary cues to improve the segmentation performance in cluttered scenes. COSNet utilizes an enhanced backbone to highlight the boundary information and capture rich feature representations. Additionally, intermediate network comprising of BEM further enhances the object boundaries and suppress the irrelevant information for better segmentation results. Extensive experiments on three challenging datasets reveal the merits of the proposed framework. The potential future direction is to extend the work for segmentation of transparent and camouflaged objects in real-time scenarios.

References

- [1] Muhammad Ali, Mamoon Javaid, Mubashir Noman, Mustansar Fiaz, and Salman Khan. Fanet: Feature amplification network for semantic segmentation in cluttered background. In *IEEE ICIP 2024*, 2024. 2, 5, 6
- [2] Iñigo Alonso, Luis Riazuelo, and Ana C. Murillo. Mininet: An efficient semantic segmentation convnet for real-time robotic applications. *IEEE Transactions on Robotics*, 36(4):1340–1347, 2020. 2, 6
- [3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2016. 1
- [4] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes, 2022. 1, 2, 5, 6, 7, 8
- [5] Sara Casao, Fernando Peña, Alberto Sabater, Rosa Castellón, Darío Suárez, Eduardo Montijano, and Ana C. Murillo. Spectralwaste dataset: Multimodal data for waste sorting automation, 2024. 1, 2, 5, 6, 8
- [6] Elisa Cascina, Andrea Pellegrino, and Lorenzo Tozzi. Resource constrained semantic segmentation for waste sorting. *arXiv preprint arXiv:2310.19407*, 2023. 2
- [7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 1, 2, 4
- [8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 4, 5, 6
- [9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020. 5
- [10] Brendan Chongzhi Corrigan, Zhi Yung Tay, and Dimitrios Konovessis. Real-time instance segmentation for detection of underwater litter as a plastic source. *Journal of Marine Science and Engineering*, 11(8):1532, 2023. 2
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5
- [12] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *CVPR*, pages 9716–9725, June 2021. 3
- [13] Zhicheng Feng, Jie Yang, Lifang Chen, Zhichao Chen, and Linhong Li. An intelligent waste-sorting and recycling device based on improved efficientnet. *International Journal of Environmental Research and Public Health*, 19(23):15987, 2022. 1, 2
- [14] Sathish Paulraj Gundupalli, Subrata Hait, and Atul Thakur. A review on automated sorting of source-separated municipal solid waste for recycling. *Waste Management*, 60:56–74, 2017. Special Thematic Issue: Urban Mining and Circular Economy. 1
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [16] Md Mosarraf Hossen, Molla E Majid, Saad Bin Abul Kashem, Amith Khandakar, Mohammad Nashbat, Azad Ashraf, Mazhar Hasan, Ali K Ansaruddin Kunju, Saidul Kabir, and Muhammad EH Chowdhury. A reliable and robust deep learning model for effective recyclable waste classification. *IEEE Access*, 2024. 2
- [17] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3
- [18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2
- [19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2
- [20] Silpa Kaza, Lisa C. Yao, Perinaz Bhada-Tata, and Frank Van Woerden. What a waste 2.0: A global snapshot of solid waste management to 2050. *Urban Development*, 2018. 1
- [21] Myungchul Kim, Sanghyun Woo, Dahun Kim, and In So Kweon. The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation, 2020. 3
- [22] Brett Koonce and Brett Koonce. Efficientnet. *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123, 2021. 2
- [23] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6015–6026, 2023. 2, 3
- [24] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J. Davison. Bootstrapping semantic segmentation with regional contrast, 2022. 5
- [25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1, 2, 7
- [26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 7
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5
- [28] Wei-Lung Mao, Wei-Chun Chen, Chien-Tsung Wang, and Yu-Hao Lin. Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling*, 164:105132, 2021. 1, 2
- [29] Jing Meng, Ping Jiang, Jianmin Wang, and Kai Wang. A mobilenet-ssd model with fpn for waste detection. *Journal*

- of *Electrical Engineering & Technology*, 17(2):1425–1431, 2022. 1, 2
- [30] Phillip A. Mlsna and Jeffrey J. Rodríguez. Chapter 19 - gradient and laplacian edge detection. In Al Bovik, editor, *The Essential Guide to Image Processing*, pages 495–524. Academic Press, Boston, 2009. 2, 3, 4
- [31] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12671–12681, 2020. 5
- [32] Jianchun Qi, Minh Nguyen, and Wei Qi Yan. Nuni-waste: novel semi-supervised semantic segmentation waste classification with non-uniform data augmentation. *Multimedia Tools and Applications*, pages 1–19, 2024. 2
- [33] Jiongming Qin, Cong Wang, Xu Ran, Shaohua Yang, and Bin Chen. A robust framework combined saliency detection and image recognition for garbage classification. *Waste Management*, 140:193–203, 2022. 2
- [34] Raiyan Rahman, Mohsena Chowdhury, Yueyang Tang, Huayi Gao, George Yin, and Guanghui Wang. Kitchen food waste image segmentation and classification for compost nutrients estimation. *arXiv preprint arXiv:2401.15175*, 2024. 2
- [35] Xueyong Tian, Liwei Shi, Yuanqing Luo, and Xinlong Zhang. Garbage classification algorithm based on improved mobilenetv3. *IEEE Access*, 2024. 2
- [36] Shakti N Wadekar and Abhishek Chaurasia. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*, 2022. 2
- [37] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 2
- [38] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 5, 6, 7
- [39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2
- [40] Zhongyi Xia, Houkui Zhou, Huimin Yu, Haoji Hu, Guangqun Zhang, Junguo Hu, and Tao He. Yolo-mtg: a lightweight yolo model for multi-target garbage detection. *Signal, Image and Video Processing*, pages 1–16, 2024. 1, 2
- [41] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018. 5
- [42] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild, 2020. 1
- [43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc., 2021. 2, 3, 6
- [44] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 2, 3, 7
- [45] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *ArXiv*, abs/2107.00641, 2021. 7
- [46] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 334–349. Springer, 2018. 3
- [47] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation, 2021. 1, 2
- [48] Dmitry Yudin, Nikita Zakharenko, Artem Smetanin, Roman Filonov, Margarita Kichik, Vladislav Kuznetsov, Dmitry Larichev, Evgeny Gudov, Semen Budenny, and Aleksandr Panov. Hierarchical waste detection with weakly supervised segmentation in images from recycling plants. *Engineering Applications of Artificial Intelligence*, 128:107542, 2024. 2
- [49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1
- [50] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1, 2, 5, 6, 7