WISNet: Pseudo Label Generation on Unbalanced and Patch Annotated Waste Images

Shifan Zhang[1] Hongzi Zhu[1*] Yinan He[2] Minyi Guo[1] Ziyang Lou[1] Shan Chang[3]
[1] Shanghai Jiao Tong University [2] Tencent AI Lab [3] Donghua University
{zhangshifan, hongzi, louworld12}@sjtu.edu.cn
heyinanda@alumni.sjtu.edu.cn guo-my@cs.sjtu.edu.cn
changshan@dhu.edu.cn

Abstract

Computer-vision-based assessment on waste sorting is desired to replace manpower supervision in Shanghai city. Due to the hardness of labeling a multitude of waste images, it is infeasible to train a semantic segmentation model for this purpose directly. In this work, we construct a new dataset consisting of $12,208$ *waste images, upon which seed regions* (i.e., *patches) are annotated and classified into* 21 *categories in a crowdsourcing fashion. To obtain pixel-level labels to train an effective segmentation model, we propose a weakly-supervised waste image pseudo label generation scheme, called WISNet. Specifically, we train a cohesive feature extractor with contrastive prototype learning, incorporating an unsupervised classification pretext task to help the extractor focus on more discriminative regions even with the same category. Furthermore, we propose an effective iterative patch expansion method to generate accurate pixel-level pseudo labels. Given these generated pseudo labels, a few-shot segmentation model can be trained to segment waste images. We implement and deploy WISNet in real-world scenarios and conduct intensive experiments. Results show that WISNet can achieve a state-of-the-art* 40.2% *final segmentation mIoU on our waste benchmark, outperforming all other baselines and demonstrating its efficacy. The dataset and code will be publicly available at:* https://github.com/shifan-Z/WISNet

1. Introduction

To reduce waste treatment overload and make the city more environmentally friendly, new waste sorting regulations have come into effect in Shanghai, one of the biggest metropolises in China. As many residents have little enthusiasm and knowledge for waste sorting, a team of over $30,000$ volunteers has been recruited to supervise trash sorting on around $21,000$ disposal sites across the city. To reduce manpower costs, an automated waste sorting quality assessment system is needed to analyze top-view images of trash bins, identify all waste categories, and estimate the proportion of incorrectly sorted waste.

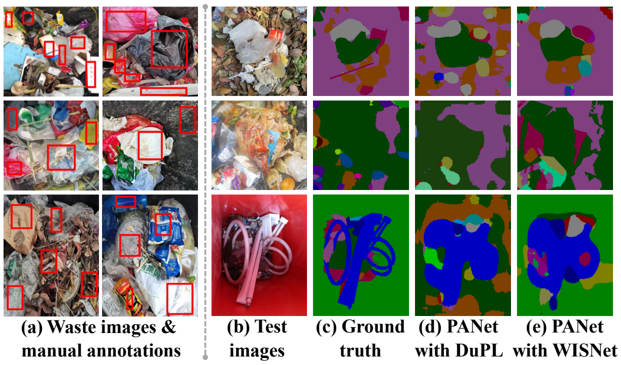

Figure 1. (a) Examples of waste images and patch-level annotations to minimize the manpower costs; (b) examples of test waste images; (c) ground truth; (d) and (e) are the segmentation results using PANet with labels generated by a SOTA model [34] and WISNet, respectively.

To recognize each category in a waste image, however, is quite distinct from classic image segmentation tasks due to three reasons. First, waste images are more complex with distorted and stained waste objects of various categories densely distributed and highly overlapped with each other. For instance, as illustrated in Figure 1 (a), it is often challenging for human eyes to recognize all categories in a waste image. Second, the distribution of waste categories is severely imbalanced, making the number of samples of each category quite imbalanced. Last but not least, there is no existing large-scale and high-quality community waste dataset available for waste image segmentation study.

In the literature, many segmentation models [3, 4] have been proven effective on various datasets, ranging from

[*] Corresponding author

biomedical images to real-world scenes, but they heavily rely on training datasets with pixel-level annotations. Recently, weakly-supervised learning, using simple labels such as image-level labels [5, 7, 31, 34, 35], scribbles [21] or bounding boxes [14, 19, 25, 30], greatly alleviates the difficulty of constructing expensive labeled data. Nevertheless, heavily stained waste objects severely impair the performance of these segmentation models. As illustrated in Figure 1, results of applying the PANet [33] trained with pseudo labels generated with DuPL [34], a SOTA solution to weakly-supervised segmentation, are unsatisfactory.

In this paper, we propose a weakly-supervised waste image pseudo label generation scheme, called *WISNet*, to facilitate the analysis of semantic composition of waste images. To boost the study, we first construct a waste dataset, named *ShanghaiWaste*, comprising $12,208$ images across 21 categories. To minimize annotation costs, as shown in Figure 1 (a), rectangle patches are used to label objects of the same category by volunteering sanitation workers, leading to a large set of $40,392$ patches in total. The core idea of WISNet is to first obtain pixel-level semantic information from patch-level annotations and then train a segmentation model using the generated masks as supervision. There are two main challenges in designing WISNet as follows.

First, the imbalanced distribution and patch-level annotations of waste images make generating accurate pixel-level labels for all categories challenging. To tackle this, we propose a framework of contrastive prototype learning, where both positive and negative feature prototypes of each waste category are generated and used to assign pseudo labels to unpatched pixels by matching the corresponding features of these pixels to the prototypes. Furthermore, to obtain stable results, we introduce an iterative patch expansion method, where newly generated pseudo labels are in turn used to update those prototypes. This process repeats until convergence. Consequently, all categories, regardless of sample abundance, can be effectively recognized.

Second, assigning labels to unpatched pixels in a waste image is a non-trivial task since only pixels within patches have semantic labels. To deal with this challenge, it is key to obtain a supreme feature extractor, which can effectively discriminate complex waste objects of various categories in waste images. Observing the fact that the image characteristics of waste objects in the same category vary significantly, we incorporate a challenging classification task into the feature extractor training process. Specifically, all patches of each category are first classified into different subcategories in the feature space using a mature pre-trained image feature extractor. Then, in addition to the given patch-level annotations, the subcategory information is also used to supervise the classification task designed in the training of the feature extractor. As a result, the ability of the extractor to capture diverse attributes of waste images is greatly enhanced.

We implement a prototype system of WISNet and deploy the system in two residential complexes in Shanghai. We evaluate the performance of WISNet with real-world waste images and the results demonstrate that WISNet can achieve a SOTA average mIoU of 40.2% over all categories on our waste benchmark at the current stage.

The main contributions of our work are as follows:

- We construct a new waste dataset consisting of $40,392$ patches of 21 waste categories from $12,208$ images, which will be available for public study.
- We propose a new pseudo label generation scheme, called WISNet, integrating neat feature extraction and iterative patch expansion.
- We deploy a weakly-supervised waste image segmentation system at two pioneer residential complexes in Shanghai, gaining SOTA performance in real-world experiments.

2. Related Work

2.1. Waste Image Datasets

Existing waste image datasets [16, 24, 26] contain a limited number of images captured in simple scenes, typically featuring fewer than $1,000$ images with a few waste objects against clear backgrounds. Other datasets [11, 12] include a larger number of images but primarily focus on underwater environments. The TrashBox [17] collects images from the web, yet many of these may not accurately represent actual trash. The WIXray [27] provides valuable X-ray waste images, but expanding this dataset is challenging due to the uncommon practice of deploying X-ray machines at waste disposal sites. Consequently, there is a notable lack of datasets featuring mixed waste objects in trash bins with severe distortion and contamination. To address the gap, we construct a new waste dataset, namely ShanghaiWaste, which comprises $12,208$ images across 21 waste categories.

2.2. Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation, leveraging coarse annotations like image-level annotations [1, 9, 31, 34, 35], scribbles [21] or bounding boxes [19], aims to learn semantic segmentation without exhaustive labeling. Most SOTA methods [7, 31, 34] follow a similar strategy in which information is excavated from coarse annotations to obtain pseudo labels, aiming at revealing the accurate shapes and boundaries of object areas. Some methods [14, 25, 30] utilize traditional techniques like graph-based optimization and dense conditional random fields to refine initial regions. Others [13, 15] adopt a seed-and-expand principle, employing algorithms and loss functions to expand crude annotations. A few approaches [1, 2] extract pixel-wise affinity from weak annotations to propagate initial seeds, while some methods [10, 18, 36] leverage additional data like saliency maps or video frames to enhance performance. Most existing methods focus on two natural datasets COCO [22] and VOC [8], where all instances of object categories are labeled in different terms including image-level annotations, scribbles, and bounding boxes. In this paper, we introduce new weak annotations in the form of patch-level annotations for the ShanghaiWaste dataset.

2.3. Few-shot Learning

The distribution of objects in the real world is naturally imbalanced. Most existing datasets are well-designed and manually balanced to avoid performance degradation caused by class imbalance. However, in many cases, data of some categories are difficult to collect and imbalanced data distributions inevitably become a challenge. In recent years, research studies [28, 32] introduce the challenging problem of few-shot segmentation, in which new models are developed to learn to predict given only a few annotated examples. It can greatly help solve the data imbalance problem. In few-shot segmentation literature, metric-learning that measures the similarity between support images and query images for fine-grained mask prediction is a popular paradigm. Many few-shot learning segmentation models [6, 33, 37, 38] adopt the concept of prototypical networks that extract prototypes from support samples. The distribution of waste is inherently imbalanced. To address this issue, we propose a contrastive prototype learning framework based on few-shot learning.

3. The ShanghaiWaste Dataset

In Shanghai, sanitation workers upload photos taken from the top of waste bins to a data center via a smartphone app if they find trash is not properly sorted out. Such images are manually filtered according to the image quality, eliminating blur photos taken under extremely dark environments or taken by shaky hands. In total, we collect $15,090$ waste images from local authorities which were taken at different indoor and outdoor locations, ranging from residential areas to business districts, under different light conditions.

3.1. Waste Categories

In Shanghai, the current waste sorting method divides household waste into four general categories: wet waste, recyclables, hazardous waste, and dry waste. Wet waste refers to organic waste such as food waste and green waste, which can be composted with biotechnology to produce fertilizer and renewable energy. Recyclables include paper, plastic, glass, metal, and fabric, which can be recycled after comprehensive treatment to reduce pollution and save resources. Hazardous waste such as batteries, expired medicines and light bulbs contains heavy metal and other toxic substances that cause potential harm to human health and the environment, and thus need special treatment before disposal. As each super category contains a large variety of waste items with distinct appearances, we further define 21 categories of interest. Moreover, we explicitly define an additional background category, *i.e.*, trash bin, to distinguish trash bins from other waste in images. Table 1 presents the detailed statistics of images and patches across different waste categories. The data clearly highlights the imbalanced distribution of waste categories within the ShanghaiWaste dataset.

Super Category	**Category**	**# of Images**	**# of Patches**
Wet Waste	Wet (We.)	3522	4032
Recyclables	Fabric (Fa.)	444	664
	Cardboard (Ca.)	864	1328
	Paper (Pa.)	691	1040
	Glass (Gl.)	706	1538
	Metal (Met.)	224	484
	Leather (Le.)	27	43
	Shoes (Sh.)	71	105
	Foam Plastic (FP.)	651	1047
	Other Plastic (OP.)	1093	1554
	Plastic Bottle (PBo.)	1125	2179
	Tetrapak (Te.)	240	322
Dry Waste	Plastic Film (PF.)	5945	16096
	Tissue (Ti.)	277	470
Hazardous Waste	Battery (Ba.)	161	363
	Medicine (Me.)	617	1131
	Electronic (El.)	111	147
	Lamp (La.)	788	1836
	Paint Bucket (PB.)	21	28
	Pesticide (Pe.)	220	379
Background	Bin (Bi.)	5553	5606

Table 1. The ShanghaiWaste dataset contains $40,392$ patches of 21 categories of waste from the $12,208$ waste images.

3.2. Patch-level Annotations

Identifying and annotating each pixel in waste images is challenging, even for humans. To balance practical usability and development costs, a crowd-sourcing method is adopted where sanitation workers as volunteers are asked to focus on large continuous regions of distinct waste objects. Rectangle boxes are used to annotate obvious waste materials of the same category, referred to as patches, as cues for the location of different waste objects. To ensure the annotation quality, each image is annotated by three workers independently, and their results are merged automatically by combining overlapping boxes with the same labels. We then review and refine these annotations to ensure quality. This process increases annotation efficiency and results in $40,392$ patches from $12,208$ waste images.

Figure 1 (a) shows examples of patch annotations, where pixels within a red box belong to the same category, and pixels outside remain unlabeled. Unlike bounding boxes, where pixel overlap can occur, all pixels within a patch exclusively pertain to a single object, ensuring there is no overlap between two patches. This feature proves to be highly valuable, serving as the foundation for the creation of a novel pseudo-label generation algorithm.

3.3. Comparison with Existing Waste Datasets

Compared to existing waste datasets, the ShanghaiWaste dataset stands out with several compelling advantages. First, the collected images reflects the real situation within trash bins, which is crucial for community waste sorting. Second, the waste images collected from widely distributed waste bins are much more complex for segmentation.

4. Design of WISNet

4.1. Overview

WISNet is designed to analyze the semantic composition of waste images by training an effective semantic segmentation model using an imbalanced dataset with weak annotations. As depicted in Figure 2, WISNet adopts a pipeline of weakly-supervised pseudo label generation scheme, which consists of three components as follows:

Pre-trained Unsupervised Classification (PUC) Considering the large disparity of image characteristics even in the same waste category, the PUC module leverages a pre-trained image feature extractor to divide each waste category into K more consistent subcategories in the feature space. The finer classification labels serve as one of the supervisory signals for the CFE module.

Cohesive Feature Extraction (CFE) The feature extractor aims to find a feature space where data points cluster around a prototype representation for each class. To train this extractor, we utilize the support-query setting from few-shot learning. The manual annotated masks of the query set serve as one of the supervisions for prototype extraction. Moreover, we further introduce a more demanding classification task in the CFE module, enhancing the network to capture diverse attributes within the same category. The classification supervision is provided by the PUC module.

Iterative Patch Expansion (IPE) For an image with patch annotation, the IPE module uses the feature extractor obtained in the CFE module to generate a pixel-level pseudo label. It begins by initializing the prototypes using the manually annotated masks as guidance. Subsequently, it iteratively updates the prototypes and expands the masks with the assistance of the original patch-level masks. Once the process converges, the final masks are upsampled to form pixel-level pseudo labels.

Finally, the pixel-level pseudo labels generated through the IPE can be used to train an underlying segmentation network. To deal with imbalanced waste categories, a classic few-shot segmentation model is trained. In each training episode, the segmentation model is trained to segment query images with prototypes extracted from a small support set of images with pseudo labels. Finally, given the obtained segmentation model, real-world waste images can be well segmented for automatic waste sorting quality assessment.

4.2. Pre-trained Unsupervised Classification

We observe that objects of the same waste category may have distinct image characteristics. For example, different types of papers have unique reflective properties. It would be beneficial if such subcategory information can be obtained and leveraged to train the backbone feature extractor. To this end, we employ unsupervised K-Means clustering to further divide each category $\mathcal{C}_i$, for $i \in [1, 21]$, into K subcategories, denoted as $\mathcal{C}_{ij}$, for $j \in [1, K]$.

Specifically, a mature feature extractor pre-trained on ImageNet is used to extract features from all images. For all patches labelled with category $\mathcal{C}_i$, denoted as $P^{\mathcal{C}_i}$, the clustering objective can be written as:

$$\min_{\mathbf{C}\in\mathbb{R}^{d\times K}} \sum_{k=1}^{|P^{\mathcal{C}_i}|} \min_{\mathcal{A}^{c_i}} \left\| MAP(E(P_k^{\mathcal{C}_i})) - \mathbf{C}\mathcal{A}_k^{c_i} \right\|^2 \quad (1)$$

$$\text{s.t., } \mathcal{A}_k^{c_i\top} 1_K = 1,$$

where $E(P_k^{\mathcal{C}_i})$ is the feature of the k-th patch in $P^{\mathcal{C}_i}$ obtained from the extractor and MAP is the masked average pooling, leading to a feature vector of d dimensions; $\mathbf{C}$ is a $d \times K$ cluster centroid matrix that needs to be learned; $|P^{\mathcal{C}_i}|$ denotes the number of patches in set $P^{\mathcal{C}_i}$; and $\mathcal{A}_k^{c_i} \in \{0, 1\}^K$ is the clustering assignment for patch $P_k^{\mathcal{C}_i}$. As a result, the derived $\mathcal{A}_k^{c_i}$ serves as the subcategory label of patch $P_k^{\mathcal{C}_i}$.

4.3. Cohesive Feature Extraction

To deal with the imbalance distribution of waste categories, we propose *contrastive* prototype learning to train the backbone feature extractor. The key idea of prototype learning [33] is to use a support set to generate feature prototypes which can be used to segment a query image by matching the extracted feature of the query image to the learned prototypes. In this way, the backbone can well recognize all categories either with abundant samples or with few samples. In addition to prototype learning, we use both positive and negative prototypes to enhance pixels corresponding to the same category having similar features. Furthermore, to empower the feature extractor to capture diverse attributes within the same category, we incorporate a more challenging classification task during training the backbone.

Specifically, for each category $\mathcal{C}_i$, for $i \in [1, 21]$, a subset of training images, denoted as $I^{\mathcal{C}_i} = \{(X_k^{\mathcal{C}_i}, Y_k^{\mathcal{C}_i}), k \in \mathbb{N}\}$, is first constructed, where $X_k^{\mathcal{C}_i}$ is the k-th image that contains at least one patch labelled with category $\mathcal{C}_i$ in the set and $Y_k^{\mathcal{C}_i}$ is the corresponding label of $X_k^{\mathcal{C}_i}$. Particularly, location (x, y) in $Y_k^{\mathcal{C}_i}$, *i.e.*, $Y_k^{\mathcal{C}_i}(x, y)$, is set to 1 or 0 if the corresponding $X_k^{\mathcal{C}_i}(x, y)$ is labelled with $\mathcal{C}_i$ or $\mathcal{C}_j$, $j \neq i$,

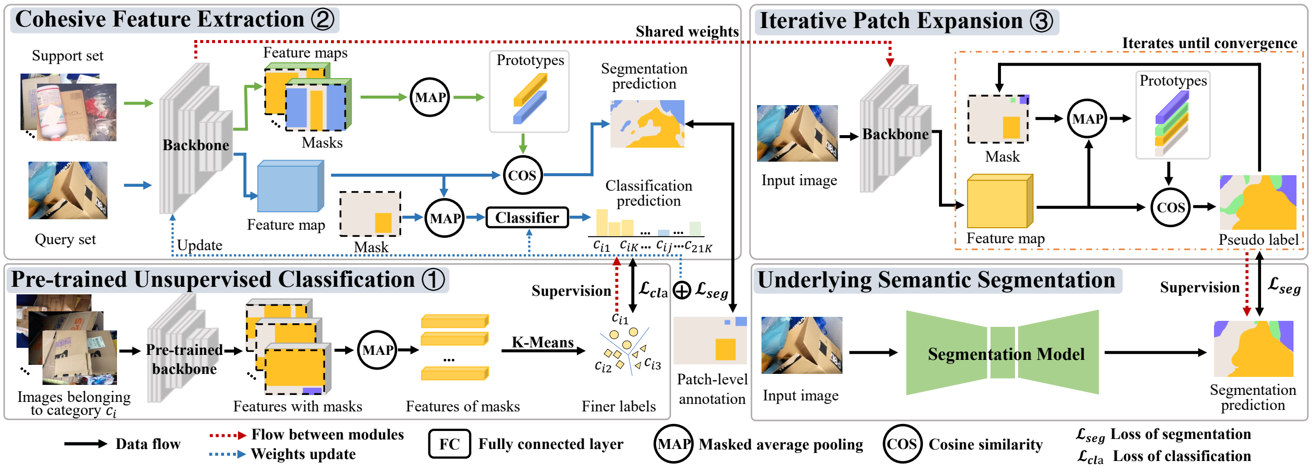

Figure 2. Framework of WISNet. WISNet consists of three components: 1) the Pre-trained Unsupervised Classification (PUC) module initially generates finer classification labels by unsupervised learning; 2) in the Cohesive Feature Extraction (CFE) module, a feature extractor is trained through contrastive learning and supervised by both the patch annotations and the finer labels generated by PUC; 3) the trained extractor in CFE is then used in the Iterative Patch Expansion (IPE) module to generate pseudo labels, which finally serve as supervision for training a downstream segmentation model.

respectively; otherwise, $Y_k^{\mathcal{C}_i}(x, y)$ is not annotated and will be ignored in the training procedure.

Then, in each training episode, a support set S of N images and a query set Q of one image are randomly sampled from $I^{\mathcal{C}_i}$. Both the support and query images are fed into the feature extractor, producing support feature set $\mathcal{S}^{\mathcal{C}_i}$ and query feature set $\mathcal{Q}^{\mathcal{C}_i}$. Two feature prototypes, *i.e.*, the positive prototype $P_1^{\mathcal{C}_i}$ representing the feature template of category $\mathcal{C}_i$ and the negative prototype $P_0^{\mathcal{C}_i}$ squeezing representative information of other categories, are calculated using masked average pooling over $\mathcal{S}^{\mathcal{C}_i}$ as follows:

$$P_t^{\mathcal{C}_i} = \frac{1}{|S|} \sum_{k=1}^{|S|} \frac{\sum_{x=1}^{H} \sum_{y=1}^{W} \mathcal{S}_k^{\mathcal{C}_i}(x, y) \mathbb{1}_t(Y_k^{\mathcal{C}_i}(x, y))}{\sum_{x=1}^{H} \sum_{y=1}^{W} \mathbb{1}_t(Y_k^{\mathcal{C}_i}(x, y))} \quad (2)$$

where $t \in \{0, 1\}$ and $\mathcal{S}_k^{\mathcal{C}_i}$ is the feature of the k-th image in S; H and W are the height and width of a feature, respectively; and $\mathbb{1}_t(Y_k^{\mathcal{C}_i}(x, y))$ is an indicator function that equals 1 when $Y_k^{\mathcal{C}_i}(x, y)) = t$ and equals 0 when $Y_k^{\mathcal{C}_i}(x, y)) \neq t$.

After obtaining $P_1^{\mathcal{C}_i}$ and $P_0^{\mathcal{C}_i}$, the dense prediction on query image is made based on the cosine similarity between the feature $\mathcal{Q}^{\mathcal{C}_i}$ and both prototypes. More specifically, a similarity map $\mathcal{M}_t^{\mathcal{C}_i}$ for each prototype $P_t^{\mathcal{C}_i}$, $t \in \{0, 1\}$ is calculated as:

$$\mathcal{M}_t^{\mathcal{C}_i}(x, y) = \frac{\mathcal{Q}^{\mathcal{C}_i}(x, y) P_t^{\mathcal{C}_i}}{||\mathcal{Q}^{\mathcal{C}_i}(x, y)|| \cdot ||P_t^{\mathcal{C}_i}||}. \quad (3)$$

Given all $\mathcal{M}_0^{\mathcal{C}_i}$ and $\mathcal{M}_1^{\mathcal{C}_i}$ for $i \in [1, 21]$, the category of each pixel in the query image is predicted using the $argmax$ scheme. We define a segmentation loss $\mathcal{L}_{seg}$ as the cross-entropy loss between the segmentation predictions and the patch-level annotations.

Finally, each query feature point $\mathcal{Q}^{\mathcal{C}_i}(x, y)$ is not only used to predict which category but used to predict which subcategory $\mathcal{C}_{ij}$ for $j \in [1, K]$. We define a classification loss $\mathcal{L}_{cla}$ as the cross-entropy loss between the subcategory predictions and the the subcategory labels as stated in the above subsection. In the end, both $\mathcal{L}_{seg}$ and $\mathcal{L}_{cla}$ are linearly combined to supervise the training.

4.4. Iterative Patch Expansion

As patches are annotated with rectangular boxes, they provide cues about the spatial location of the target regions but miss the important pixel-wise information of object shapes and boundaries. To expand the original rectangular patches to more accurate and complete pseudo labels, we propose an iterative patch expansion scheme, where the patch-level annotations serve as the initial mask and are progressively refined through repeated iterations of two steps, *i.e.*, label assignment and prototype update, until the process converges.

More specifically, in the initialization stage, given an image X, its annotation Y containing patches of m categories, and the feature F extracted with the backbone feature extractor, MAP is used on the feature F with label Y as the initial mask. Without the loss of generality, assume this process yields $m + 1$ prototypes, denoted as $P^{\mathcal{C}_0}, P^{\mathcal{C}_1}, \ldots, P^{\mathcal{C}_m}$, where $P^{\mathcal{C}_0}$ represents the prototype for all unannotated regions, and $P^{\mathcal{C}_i}$ for $i \in [1, m]$ is the prototype for each annotated category in Y.

During the iteration, in the label assignment step, a refined mask is generated based on the cosine similarity between the feature F and the derived prototypes. To enhance the accuracy of the mask, the regions of the initial mask containing m categories are overlaid onto the refined mask for pixels within a patch that entirely belongs to a specific category. In the prototype update step, MAP is used on the feature F with the refined mask to yield refined $m+1$ prototypes. When the difference between two consecutive masks is less than ϵ, the iteration process is considered converged. The final mask serve as the pseudo label of this input image.

5. Evaluation

5.1. Methodology

5.1.1. Implementation

We implement WISNet using PyTorch and train all models on a Linux server equipped with a 3.80GHz Intel i7-9800X CPU and a GeForce RTX 3090Ti. In Pre-trained Unsupervised Classification (PUC) and Cohesive Feature Extraction (CFE), VGG16 [29] is used as the backbone. The number of subcategories K is set to 3 in PUC. In CFE, we use SGD for training over $30,000$ iterations. In Iterative Patch Expansion, the convergence criterion ϵ is set to 0.1%. In Underlying Semantic Segmentation, a 1-way 5-shot setting is adopted in the training stage. In the testing stage, support sets contain all 21 waste categories.

5.1.2. Datasets

We train and evaluate WISNet on two datasets: ShanghaiWaste and ShanghaiWaste-Seg, respectively.

- *ShanghaiWaste:* We train WISNet on the ShanghaiWaste dataset, which contains $12,208$ waste images and $40,392$ annotated patches, split into $11,108$ training samples and $1,100$ validation samples.
- *ShanghaiWaste-Seg:* We evaluate WISNet on an extra constructed dataset, ShanghaiWaste-Seg, featuring pixel-level labels. The annotations were created using the open-source tool Labelme. This dataset comprises 928 real-world waste images, carefully selected to minimize the presence of unrecognizable objects. These images cover all 21 waste categories, with a detailed distribution across super categories shown in Table 2. The table reveals that the waste dataset is inherently imbalanced in terms of category distribution.

Super Category	**Wet**	**Dry**	**Rec.**	**Har.**	**Bin**
Ratio of Images (%)	62.7	85.0	91.0	28.0	82.4
Ratio of Area (%)	15.5	30.7	21.1	5.3	27.4

Table 2. Category distribution of ShanghaiWaste-Seg. “Rec.” and “Har.” denote Recyclables and Hazardous Waste, respectively.

5.1.3. Metrics

We use mean intersection over union (mIoU) as the metric to evaluate segmentation performance across 21 categories.

5.2. Choosing a Proper Segmentation Model

Different segmentation models can be applied in WISNet. In this experiment, we examine three existing segmentation models, *i.e.*, DeepLab V3 [4], PANet [33] and ASGnet [20] as candidates, which have been proved effective on fully supervised segmentation tasks on common datasets such as Pascal VOC [8] and COCO [22]. We follow the authors' settings to train the three candidates on ShanghaiWaste and test on ShanghaiWaste-Seg.

The results are shown in Table 3, where PANet consistently outperform the other two models for two reasons. First, among the three models, PANet and ASGnet are specially designed for few-shot segmentation while DeepLab V3 encounters severe performance degradation dealing with the imbalanced ShanghaiWaste dataset. Second, despite the fact that ASGnet can achieve the SOTA performance on Pascal VOC and COCO, its structure is designed specially for fully supervised learning with accurate pixel-wise labels and fails in learning from coarse annotations. As a result, WISNet integrates PANet as its segmentation model.

5.3. Performance Comparison

To our knowledge, weakly-supervised semantic segmentation (WSSS) for patch-level annotations has not been exploited before. We compare WISNet with three WSSS models based on bounding box annotations, *i.e.*, SDI [14], BANA [23] and BBAM [19], and four models based on image-level annotations, *i.e.*, AEFT [35], PPC [7] , ACR [31] and DuPL [34]. These models are trained on the ShanghaiWaste dataset to generate pixel-level pseudo labels. Subsequently, two segmentation models, PANet and DeepLab, are trained on the pseudo labeles and tested on the ShanghaiWaste-Seg dataset.

Table 4 lists the segmentation results, showing that WISNet outwits all other methods. With DeepLab and PANet, WISNet achieves mIoUs of 31.8% and 40.2%, exceeding the second-highest methods (ACR and PPC) by 2.7% and 4.1%, respectively. Figure 3 shows the qualitative results of the top models: PPC, ACR, DuPL, and WISNet, respectively. It is clear to see that WISNet can recognize both common categories and rare categories better and make more accurate predictions on object locations.

5.4. Ablation Study

5.4.1. Effectiveness of Different Components

Experiments are conducted with consistent training settings and various component combinations to verify the effectiveness of each component. As shown in Table 5, both Pre-trained Unsupervised Classification and Iterative Patch Expansion contribute to performance improvement.

Model	**Ba.**	**Bi.**	**Ca.**	**Me.**	**El.**	**Fa.**	**FP.**	**Gl.**	**La.**	**Le.**	**Met.**	**OP.**	**PB.**	**Pa.**	**Pe.**	**PF.**	**PBo.**	**Sh.**	**Te.**	**Ti.**	**We.**	**mIoU**
DeepLab	55.4	66.1	19.5	8.4	21.0	34.6	34.4	29.6	26.4	0.0	3.0	3.8	0.0	2.6	48.9	57.9	12.6	2.4	0.4	1.7	53.3	23.0
PANet	43.9	69.3	29.8	23.2	14.8	44.0	59.5	32.8	37.0	47.8	12.3	15.0	62.0	14.3	56.3	48.7	17.3	30.0	22.1	27.0	57.3	**36.4**
ASGnet	24.1	15.6	0.0	22.4	12.0	28.1	28.8	7.8	22.6	10.1	13.2	34.5	12.3	20.5	42.0	4.4	18.2	4.0	5.0	40.0	61.7	20.4

Table 3. Performance comparison of existing semantic segmentation models trained on the ShanghaiWaste dataset with patch-level annotations. Methods from top to bottom are DeepLab V3 [4], PANet [33] and ASGnet [20].

Ref.	**Seg.**	**Ba.**	**Bi.**	**Ca.**	**Me.**	**El.**	**Fa.**	**FP.**	**Gl.**	**La.**	**Le.**	**Met.**	**OP.**	**PB.**	**Pa.**	**Pe.**	**PF.**	**PBo.**	**Sh.**	**Te.**	**Ti.**	**We.**	**mIoU**
SDI	DeepLab	26.9	70.0	22.9	7.3	28.8	31.6	34.6	33.8	29.6	0.0	**2.4**	4.9	0.0	4.3	48.6	56.8	13.0	0.0	0.0	1.6	53.3	22.4
BANA	DeepLab	0.2	48.7	18.2	13.2	0.0	17.9	20.0	20.3	23.0	0.0	0.9	2.0	0.0	1.3	9.3	43.0	6.6	0.0	0.0	0.0	47.2	12.9
BBAM	DeepLab	0.0	4.2	5.1	7.8	0.0	26.4	11.1	21.9	14.9	0.0	0.0	0.0	0.0	0.0	0.0	40.0	0.7	0.0	0.0	0.0	46.8	8.5
AEFT	DeepLab	59.7	69.1	11.8	18.8	30.4	41.3	22.5	42.5	32.6	0.0	0.1	3.5	0.0	12.3	62.7	62.9	14.8	**38.4**	0.0	8.1	60.4	28.2
PPC	DeepLab	62.4	77.1	20.2	**23.4**	9.1	39.6	44	**39.1**	35.9	0.0	0.2	5.4	0.0	11.9	68.1	**71.2**	12.5	0.0	2.0	6.2	**66.6**	28.3
ACR	DeepLab	**69.1**	**79.2**	16.0	16.1	25.0	36.1	39.1	36.9	**42.4**	0.0	0.0	3.7	0.0	10.3	59.2	63.6	19.6	17.5	5.8	10.5	61.9	29.1
DuPL	DeepLab	51.0	71.3	11.6	12.8	28.1	37.4	**48.5**	36.5	35.5	0.0	0.5	5.7	0.0	9.3	46.7	66.8	16.4	0.0	0.0	2.6	64.1	26.0
WISNet	DeepLab	61.3	78.3	**22.7**	18.6	**32.5**	**42.9**	47.7	33.7	42.3	0.0	0.2	**7.6**	**13.3**	**15.0**	**68.9**	67.4	**23.2**	5.0	**6.9**	**16.6**	64	**31.8**
SDI	PANet	34.9	67.8	18.5	**26.8**	8.1	36.3	40.8	16.7	33.4	63.4	5.2	4.7	62.0	7.4	27.5	46.9	20.9	26.4	13.8	19.6	52.0	30.1
BANA	PANet	24.2	61.5	21.5	11.6	15.2	40.3	43.9	14.7	35.6	32.1	4.7	5.9	56.3	5.0	34.5	42.7	**24.5**	28.2	5.9	17.7	53.0	27.6
BBAM	PANet	41.9	45.8	11.9	8.8	16.0	22.2	23.5	19.8	12.5	37.0	5.5	8.9	39.6	13.7	32.7	22.2	5.7	18.6	10.5	17.0	33.1	21.3
AEFT	PANet	33.4	69.8	35.3	22.2	17.5	44.2	60.1	34.3	36.5	55.8	8.4	15.5	66.1	14.5	51.4	59.3	12.0	22.9	14.9	15.3	58.7	35.6
PPC	PANet	22.9	71.7	32.3	24.2	18.2	42.6	58.1	32.5	**43.6**	62.7	8.7	16	63.9	13.7	48.9	**60.1**	12.9	29.0	18.2	16.6	**60.4**	36.1
ACR	PANet	44.8	**74.1**	33.6	21.2	18.8	41.6	61.6	30.2	42.5	43.2	8.4	13.4	70.2	14.6	46.3	54.4	13.8	31.8	15.9	15.4	57.5	35.9
DuPL	PANet	30.4	65.4	32.8	24.1	**22.8**	43.0	58.6	29.5	39.0	69.1	**13.0**	15.3	53.2	11.3	44.8	57.4	14.2	20.9	11.9	10.0	46.7	34.0
WISNet	PANet	**45.2**	70.8	**37.3**	25.5	19.1	**45.3**	**62.2**	**36.8**	41.2	**77.1**	12.0	**16.8**	**72.2**	**15.7**	**56.0**	56.9	15.5	**42.3**	**21.2**	**20.8**	54.3	**40.2**

Table 4. Performance comparison between WISNet and seven SOTA weakly-supervised segmentation methods, *i.e.*, SDI [14], BANA [23] and BBAM [19] designed based on bounding box annotations, AEFT [35], PPC [7] , ACR [31] and DuPL [34] designed based on image-level annotations. PANet [33] trained with WISNet pseudo labels outwits other methods.

Baseline	**PUC**	**IPG**	**IPE**	**mIoU**
✓				36.4
		✓		38.4(+2.0)
	✓	✓		39.5(+3.1)
	✓	✓	✓	40.2(+3.8)

Table 5. Experiments of the proposed components. Baseline refers to PANet. PUC, IPG, and IPE stand for Pre-trained Unsupervised Classification, Iterative Patch Generation, and Iterative Patch Expansion, respectively. IPE makes more extensive use of patch-level annotations compared to IPG.

Seed Cues	**Refine Method**	**Segmentation Model**	**mIoU(%)**
Manual Annotation	/	DeepLab	23.0
	IPE		31.8
	/	PANet	36.4
	IPE		40.2
CAM	/	DeepLab	28.9
	PSA		28.9
	PSA + IPE		30.1
	IRN		30.1
	IRN + IPE		31.3
	IPE		29.7
	/	PANet	34.4
	PSA		34.4
	PSA + IPE		36.1
	IRN		36.6
	IRN + IPE		37.6
	IPE		35.1

Table 6. Evaluation of the quality of generated labels. The top block compares the segmentation performance when patch-level manual annotations are provided. The bottom block compares the performance when only image-level annotations are provided and CAM is used as seed cue.

5.4.2. Effectiveness of Iterative Patch Expansion (IPE)

We investigate the effectiveness of IPE by evaluating the quality of the generated labels. To do this, we train semantic segmentation models using various labels and test their performance on the ShanghaiWaste-Seg dataset. Specifically, we conduct experiments using both manual patch-level annotations and image-level annotations. For manual annotations, we compare the results of the model trained solely with manual annotations against those trained with labels generated by IPE. For image-level annotations, following the widely accepted pipeline, we utilize class activation map (CAM) as the initial cues for objects.

We assess several pseudo-label generation methods, including PSA [1], IRN [2], and our IPE, both indepen-

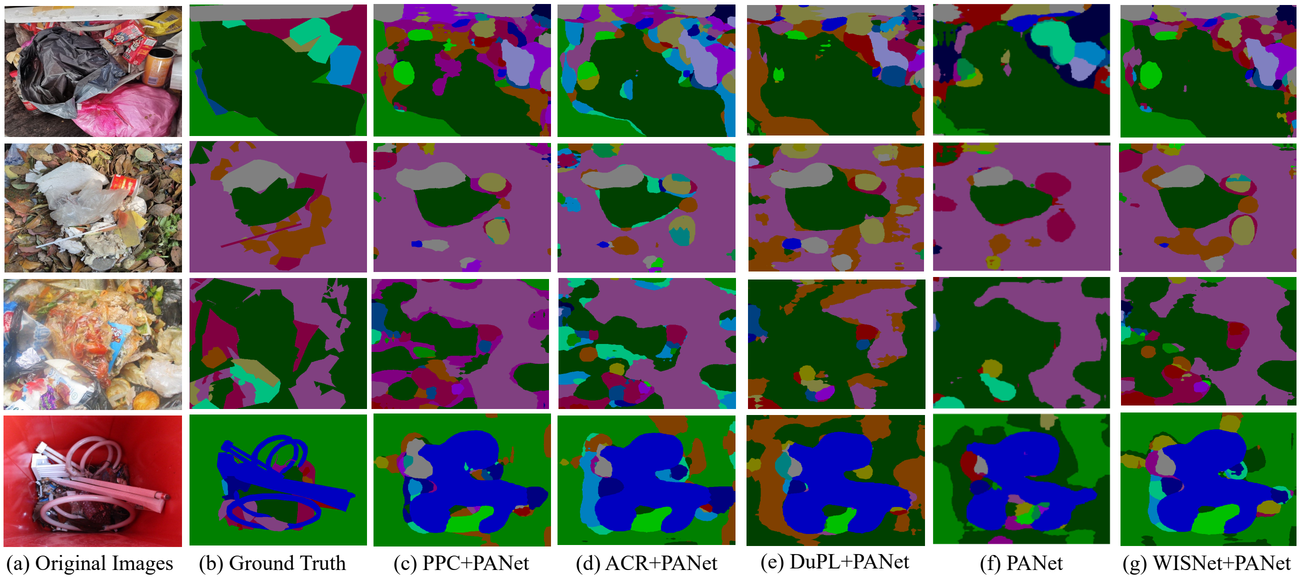

Figure 3. Visualized examples of segmentation, where (a) are waste images, (b) are ground truth, (c)-(e) are results of PANet [33] trained with pseudo labels generated by PPC [7], ACR [31] and DuPL [34] . (f)-(g) are the results of PANet trained with patch-level annotations and pseudo labels generated by WISNet. PANet trained with labels generated by WISNet achieves the best performance.

dently and in combination with segmentation models such as DeepLab and PANet. It can be seen in Table 6 that pseudo labels generated by IPE alone achieve competitive quality compared to those refined by PSA and IRN. Furthermore, IPE can be used to further refine the noisy labels generated by PSA and IRN, thereby enhancing the performance of the segmentation models.

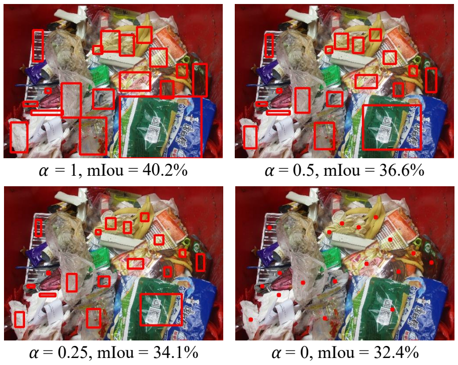
α = 1, mIou = 40.2%
α = 0.5, mIou = 36.6%
α = 0.25, mIou = 34.1%
α = 0, mIou = 32.4%

Figure 4. Cropped annotations of various size and the corresponding segmentation performance.

5.4.3. Impact of Patch Annotation Method

We further study the impact of patch annotation method in terms of patch size to the segmentation performance. Clearly, there is a trade-off between annotation cost and quality. For fair comparison, we synthesize patches of smaller size based on original annotations to represent low-quality annotations. Specifically, given a patch of size S, it is cropped in the center to get a smaller patch of size αS. We choose $\alpha \in \{0, 0.25, 0.5, 1\}$, where 0 means the patch is reduced to a point. Examples of cropped annotations and performance are shown in Figure 4. We can see that WISNet is robust to moderate fluctuation of quality of annotations and is still applicable with only point supervision.

6. Conclusion

In this paper, we have constructed ShanghaiWaste, a waste image dataset consisting of $12,208$ image samples with $40,392$ seed patch annotations of 21 waste categories, which will be available for public study. Moreover, we have proposed a weakly-supervised few-shot pseudo label generation scheme, called WISNet, for imbalanced and weakly annotated waste images dataset. We have implemented and deployed a WISNet-enabled waste image segmentation system in Shanghai city and conducted extensive experiments. Our system can achieve a SOTA segmentation mIoU of 40.2% on real-world waste images. In the future, we will continue to contribute more samples to ShanghaiWaste and study advanced end-to-end semantic segmentation methods to improve the overall performance of WISNet.

Acknowledgement

This research was supported by National Natural Science Foundation of China (Grants No. 62432008).

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4981–4990. IEEE Computer Society, 2018. 2, 7

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2209–2218. Computer Vision Foundation / IEEE, 2019. 2, 7

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 833–851. Springer, 2018. 1, 6, 7

[5] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 2

[6] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 79. BMVA Press, 2018. 3

[7] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 2, 6, 7, 8

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 3, 6

[9] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 2

[10] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, pages 371–388. Springer, 2018. 2

[11] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic detection of marine litter using deep visual detection models. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 5752–5758. IEEE, 2019. 2

[12] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 2

[13] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7014–7023. IEEE Computer Society, 2018. 2

[14] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1665–1674. IEEE Computer Society, 2017. 2, 6, 7

[15] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 695–711. Springer, 2016. 2

[16] Marek Kraft, Mateusz Piechocki, Bartosz Ptak, and Krzysztof Walas. Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. *Remote. Sens.*, 13(5):965, 2021. 2

[17] Nikhil Venkat Kumsetty, Amith Bhat Nekkare, Sowmya Kamath, et al. Trashbox: Trash detection and classification using quantum transfer learning. In *2022 31st Conference of Open Innovations Association (FRUCT)*, pages 125–130. IEEE, 2022. 2

[18] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6807–6817. IEEE, 2019. 2

[19] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 2, 6, 7

[20] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8334–8343. Computer Vision Foundation / IEEE, 2021. 6, 7

[21] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 2

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and

C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 3, 6

[23] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6913–6922. Computer Vision Foundation / IEEE, 2021. 6, 7

[24] Harsh Panwar, PK Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, Sudhansh Sharma, and Iqbal H Sarker. Aquavision: Automating the detection of waste in water bodies using deep transfer learning. *Case Studies in Chemical and Environmental Engineering*, 2:100026, 2020. 2

[25] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1742–1750. IEEE Computer Society, 2015. 2

[26] Pedro F Proença and Pedro Simoes. Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*, 2020. 2

[27] Lingteng Qiu, Zhangyang Xiong, Xuhao Wang, Kenkun Liu, Yihan Li, Guanying Chen, Xiaoguang Han, and Shuguang Cui. Ethseg: An amodel instance segmentation network and a real-world dataset for x-ray waste inspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2283–2292, 2022. 2

[28] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017. 3

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[30] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3136–3145. Computer Vision Foundation / IEEE, 2019. 2

[31] Weixuan Sun, Yanhao Zhang, Zhen Qin, Zheyuan Liu, Lin Cheng, Fanyi Wang, Yiran Zhong, and Nick Barnes. All-pairs consistency learning forweakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 826–837, 2023. 2, 6, 7, 8

[32] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638, 2016. 3

[33] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9196–9205. IEEE, 2019. 2, 3, 4, 6, 7, 8

[34] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3534–3543, 2024. 1, 2, 6, 7, 8

[35] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 326–344. Springer, 2022. 2, 6, 7

[36] Zeng Yu, Yun-Zhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7222–7232. IEEE, 2019. 2

[37] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5217–5226. Computer Vision Foundation / IEEE, 2019. 3

[38] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S. Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Trans. Cybern.*, 50(9):3855–3865, 2020. 3