# Self-supervised Image-specific Prototype Exploration
# for Weakly Supervised Semantic Segmentation

Qi Chen[1], Lingxiao Yang[1], Jianhuang Lai[1,2,3] and Xiaohua Xie[1,2,3*]

[1]School of Computer Science and Engineering, Sun Yat-Sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

chenq377@mail2.sysu.edu.cn, {yanglx9, stsljh, xiexiaoh6}@mail.sysu.edu.cn

## Abstract

*Weakly Supervised Semantic Segmentation (WSSS) based on image-level labels has attracted much attention due to low annotation costs. Existing methods often rely on Class Activation Mapping (CAM) that measures the correlation between image pixels and classifier weight. However, the classifier focuses only on the discriminative regions while ignoring other useful information in each image, resulting in incomplete localization maps. To address this issue, we propose a Self-supervised Image-specific Prototype Exploration (SIPE) that consists of an Image-specific Prototype Exploration (IPE) and a General-Specific Consistency (GSC) loss. Specifically, IPE tailors prototypes for every image to capture complete regions, formed our Image-Specific CAM (IS-CAM), which is realized by two sequential steps. In addition, GSC is proposed to construct the consistency of general CAM and our specific IS-CAM, which further optimizes the feature representation and empowers a self-correction ability of prototype exploration. Extensive experiments are conducted on PASCAL VOC 2012 and MS COCO 2014 segmentation benchmark and results show our SIPE achieves new state-of-the-art performance using only image-level labels. The code is available at* https://github.com/chenqi1126/SIPE.

## 1. Introduction

Semantic segmentation aims to assign a semantic category label to each pixel in an image, which has been widely applied in autonomous driving [14], medical imaging [38] and remote sensing image interpretation [17]. Benefiting from Convolutional Neural Networks (CNNs), semantic segmentation has achieved remarkable progress in fully supervised manner. However, training a fully supervised segmentation model requires a large number of
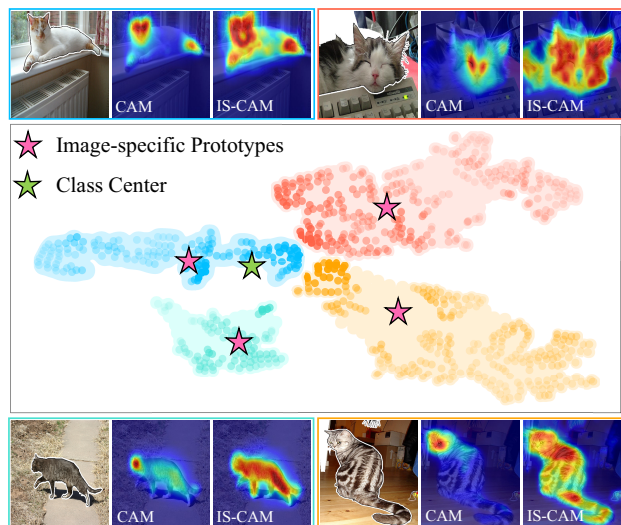
*Corresponding Author

Figure 1. Main motivation. We visualize the pixel-level feature distribution of four cat images by t-SNE [39]. The original CAM activates each pixel using class center (green star). Our method extracts image-specific prototypes (pink star) to generate Image-Specific CAM (IS-CAM) that captures more complete regions.

pixel-level annotations, which is notoriously expensive and time-consuming to collect. An alternative approach is to learn from weak labels, *e.g.*, image-level labels [2], bounding boxes [26, 49], scribbles [29, 34, 46] and points [3, 5]. Among these works, image-level labels based Weakly Supervised Semantic Segmentation (WSSS) has enjoyed great popularity within the community.

Most of existing methods leverage Class Activation Mapping (CAM) [53] technology to provide localization cues of target object. Specifically, these methods train a classifier and regard its learned weights as a general representation of each class, *i.e.*, **class center**. Then, this class center is used to correlate with image pixels to obtain localization maps as shown in Fig. 1. However, CAM tends to

focus on a few primary regions (cat's head) while ignoring other useful cues (cat's body). To explain the problem, we visualize pixel-level features of foreground extracted from a trained classification network. Those features are shown with four different colors and their transparency degree indicate the activation of CAM. We find that the class center always gives high activations to the close pixels (correspond to some primary regions) and ignores the distant pixels. The imbalanced activations lead to the incomplete localization map as demonstrated in Fig. 1. In addition, activating features on each image by the centroid of that features (pink star) can be beneficial to explore more complete regions (see Image-Specific CAM (IS-CAM) shown in Fig. 1). Therefore, this paper aims to tailor image-specific prototypes to adaptively describe the image itself.

To this end, we propose a novel weakly supervised semantic segmentation framework, called *Self-supervised Image-specific Prototype Exploration (SIPE)*. The proposed SIPE consists of an Image-specific Prototype Exploration (IPE) and a General-Specific Consistency (GSC) loss, which is illustrated in Fig. 2. Specifically, IPE is realized as two sequential steps to characterize prototypes, allowing to capture more complete localization maps. In the first step, we utilize inter-pixel semantics to explore spatial structure cues, locating robust seed regions of each class. Given the seed regions, we extract image-specific prototypes and then produce our IS-CAM by prototypical correlation. In addition, GSC is proposed to construct the consistency of general CAM and our specific IS-CAM. This self-supervised signal further optimizes the feature representation and empowers a self-correction ability of prototype exploration. Extensive experiments are conducted on Pascal VOC 2012 [11] and MS COCO 2014 [30] and results show that our SIPE achieves new state-of-the-art performance when only image-level labels are available.

Our main contributions are summarized as:

- We propose Self-supervised Image-specific Prototype Exploration (SIPE) to learn image-specific knowledge for weakly supervised semantic segmentation.

- We propose Image Prototype Exploration (IPE) that tailors image-specific prototypes for each image, which is achieved by structure-aware seed locating and background-aware prototype modeling. It enables the model to capture more complete localization maps.

- We propose a General-Specific Consistency (GSC) loss to effectively regularize the original CAM and IS-CAM, empowering the feature representation.

## 2. Related Work

Due to the low annotation costs, image-level labels based weakly supervised semantic segmentation has attracted in-

creasing attention. Most existing methods adopt Class Activation Mapping (CAM) to generate localization maps and then refine them as pseudo labels to train a fully supervised segmentation model. To achieve a high performance segmentation model, many strategies have been investigated to improve the quality of localization maps.

**Erasure and accumulation.** Erasure methods explore more object regions by intentionally removing the discriminative regions from the images [23, 37, 43] or feature maps [8, 18]. However, erasing most of the discriminative regions may confuse the classifier and result in false positives. To avoid this problem, some works accumulate multiple activations by applying well-designed sampling on dilated convolution rate [44], image scales [50], spatial location [24] and training process [20].

**Cross-image mining.** Considering the sharing semantics between images, some works design cross-image relation mining modules, such as cross-image affinity [13], max bipartite matching [31] and co-attention classifier [36] to excavate semantic context of weak labels. Furthermore, the collaborative information of multi-images is explored to capture the potential knowledge by graph convolution network [28] and self-attention mechanism [45].

**Background Modeling.** Many methods [21, 27, 47, 48] obtain precise background by using auxiliary saliency maps, which is laborious. Without auxiliary maps, Fan et al. [12] propose intra-class discriminator to separate the foreground and the background for each class. However, due to the object and scene diversity of images, it is quite tricky to learn a general intra-class discriminator for each class.

**Self-supervised Learning.** More recent, self-supervised methods mine potential information and build supervisory signals, which has been demonstrated a promising solution to narrow the supervision gap between fully and weakly supervised semantic segmentation. Wang et al. [42] apply consistency regularization on CAM from various transformed images to accomplish self-supervision learning. Chang et al. [4] introduce a self-supervised task that discovers sub-categories, which provides additional supervision to enhance feature representations.

In contrast to existing methods, we fully consider the distinctiveness of images, and introduce image-specific prototypes to discover complete regions and construct a self-supervised manner to empower feature representation.

## 3. Approach

This section elaborates the proposed SIPE framework for weakly supervised semantic segmentation as shown in Fig. 2. Firstly, we briefly review the preliminary of CAM. Then we describe the pipeline of exploring image-specific prototypes and Image-Specific CAM (IS-CAM). Finally, a self-supervised learning with General-Specific Consistency (GSC) is introduced to empower feature representation.
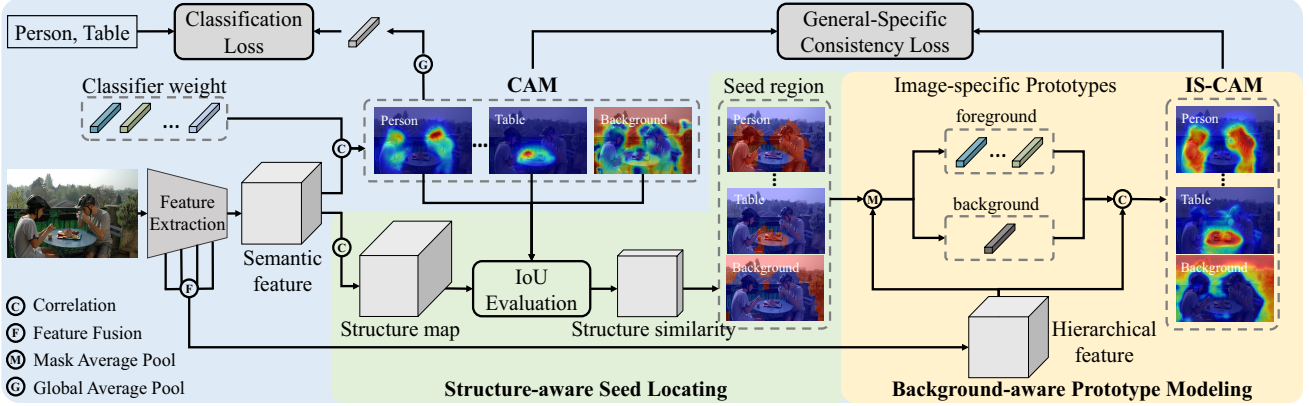
Figure 2. Overview of the proposed SIPE for weakly supervised semantic segmentation. It mainly consists of two proposed methods: an Image-specific Prototype (IPE) Exploration and a General-Specific Consistency (GSC) loss. Specifically, in our IPE, a structure-aware seed locating method is proposed to achieve more robust seed regions and a background-aware prototype modeling is developed to extract hierarchical features. In addition, we add consistent regularization between two types of CAM (*i.e.*, general CAM and our IS-CAM). This self-supervised signal effectively does correction in both CAM and IS-CAM.

## 3.1. Class Activation Mapping

Given an input image and a pretrained classification network, the class activation map $M_f = \{M_k\}_{k=1}^K$ over $K$ foreground classes can be represented as follows:

$$M_k = ReLU(\boldsymbol{\theta}_k^T \boldsymbol{F}_s), \quad \forall k \in K, \tag{1}$$

where $\boldsymbol{F}_s$ is the semantic feature from the last layer of the network, $\boldsymbol{\theta}_k$ denotes the $k$-th classifier weight, and thus $M_k$ is the $k$-th class-specific activation map. Following previous works, CAM is further normalized to $[0, 1]$ by the maximum value along the spatial axes so that it could be regarded as the probability for each class.

Considering the importance of background in segmentation task, we follow [42] to estimate the background activation map $M_b$ based on $M_f$. Since CAM tends to cover object regions partially, the estimated background often contains high responses in foreground regions, which will bring considerable noise. To reduce such confusion, we weaken the background probability by introducing an attenuation coefficient $\alpha = 0.5$:

$$M_b = \alpha(1 - \max_{1 \le k \le K} M_k). \tag{2}$$

We combine the processed background activation map with the foreground activation map as a whole, *i.e.* $M = M_f \cup M_b$, to help model background knowledge.

## 3.2. Image-specific Prototype Exploration

Image-specific prototype is proposed to represent the feature distribution of each class, allowing to capture more complete regions. Different from the prototype representation in few-shot segmentation [33, 40, 52], there is no

ground truth pixel-level mask in WSSS. To explore image-specific prototypes for characterizing a feature distribution, we design an efficient two-step pipeline. The first step provides robust class-wise seed regions and the second step aggregates these seeds on a comprehensive feature space to achieve accurate image-specific representation.

**Structure-aware Seed Locating.** A straightforward approach to obtain seeds is empirically selecting thresholds for CAM [19], but it is difficult to use a fixed threshold on different images due to the diversity of objects and scenarios. Although CAM pays more attention to discriminative regions, it also produces weak activations on the remaining regions. It means that CAM has the potential to provide the spatial structure of semantic objects. Besides, a pixel's spatial structure can be constituted by clustering high correlation pixels. For an image, we can determine each pixel's category by comparing its spatial structure with CAMs. Based on above analysis, we propose a structure-aware seed locating method by exploring inter-pixel semantics to capture the spatial structure and employing CAMs as the templates to match the optimal category.

Fig. 3 illustrates the proposed method with selected foreground and background examples. Firstly, for an arbitrary pixel $i$, we take its semantic feature vector $\boldsymbol{f}^i$ as the query to compute semantic correlation with all pixels in that feature map. Since pixels with high correlation scores are more likely to belong to the same class, these high correlation pixels can highlight the spatial structure. Therefore, we define the spatial structure of a pixel by inter-pixel semantic correlation:

$$\boldsymbol{S}^i(j) = ReLU(\frac{\boldsymbol{f}^i \cdot \boldsymbol{F}_s(j)}{||\boldsymbol{f}^i|| \cdot ||\boldsymbol{F}_s(j)||}), \tag{3}$$
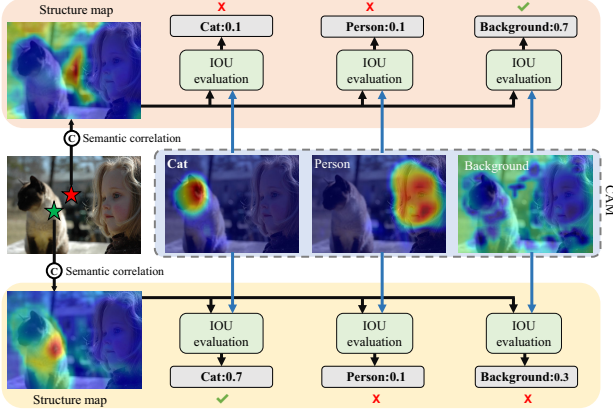
Figure 3. Illustration of structure-aware seed locating of selected foreground (green) and background pixels (red). The structure map are obtained by semantic correlation. Then the class-wise structure similarity can be evaluated by IoU and the maximal one is selected as the final class.
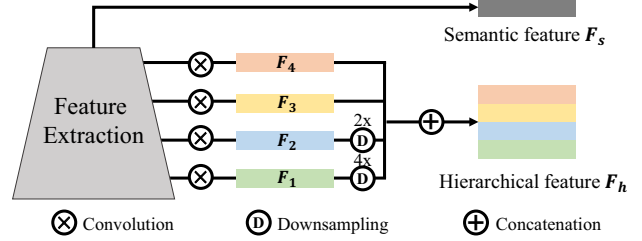


Figure 4. Illustration of the modified backbone for extracting hierarchical feature representations. The semantic feature is extracted from the last layer and the hierarchical feature is the fusion of four stages of the backbone.

where $\cdot$ is dot product and $j$ is spatial index over the feature map as well as the structure map. $\boldsymbol{S}^i(j)$ denotes the semantic correlation between pixel $i$ and $j$, and $\boldsymbol{S}^i$ is the structure map of pixel $i$. We suppress negative correlations by ReLU function to eliminate the influence of irrelevant pixels.

Secondly, we evaluate the class-wise IoU between the structure map of pixel $i$ and CAM as the structure similarity:

$$\boldsymbol{C}_k^i = \frac{\sum_j \boldsymbol{M}_k(j)\boldsymbol{S}^i(j)}{\sum_j [\boldsymbol{M}_k(j) + \boldsymbol{S}^i(j) - \boldsymbol{M}_k(j)\boldsymbol{S}^i(j)]}. \quad (4)$$

Here, $\boldsymbol{C}_k^i$ denotes the structure similarity for pixel $i$ with respect to $k$-th class. $j$ is spatial index over the activation map as well as the structure map. From Fig. 3, we can see that the foreground pixel (green star) correlates with the cat's body and achieves the highest IoU with CAM of the cat category. Additionally, the background pixel (red star) is not associated with foreground pixels, so it is more likely to belong to the background class.

Finally, pixel $i$ is assigned to the category with maximal similarity:

$$\boldsymbol{R}_k^i = \begin{cases} 1, & \text{if } k = \arg\max_{k'} \boldsymbol{C}_{k'}^i, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

By repeating this process for all pixels of the image in parallel, the seed regions $\boldsymbol{R}$ of both foreground and background classes are located as shown in Fig. 2.

**Background-aware Prototype Modeling.** In this section, we simultaneously model foreground and background prototypes. Considering that background does not have specific semantics, it is difficult to explore representative

background prototypes on semantic feature space. Instead, features from shallow layers contain rich low-level visual information (*e.g.* color, texture), which is more suitable to model background-related information. Therefore, we modify the architecture of the backbone to capture hierarchical features for effective prototype representation.

Fig. 4 illustrates the architecture of modified backbone. Specifically, we add four convolution layers to extract multi-scale outputs. Then the multi-scale outputs are resized to the same size and concatenated to form the hierarchical feature $\boldsymbol{F_h}$. Thus image-specific prototypes $\boldsymbol{P}_k$ of foreground and background can be formulated as the centroid of seed regions in hierarchical feature space:

$$\boldsymbol{P}_k = \frac{\sum_i \boldsymbol{F_h}^i * \mathbb{1}(\boldsymbol{R}_k^i = 1)}{\sum_i \mathbb{1}(\boldsymbol{R}_k^i = 1)}, \quad (6)$$

where $i$ indexes the spatial locations and $\mathbb{1}(\cdot)$ outputs 1 if the argument is true or 0 otherwise. This process performs class-wise compression on the seed pixels, achieving $K$ foreground prototypes and one background prototype.

With these image-specific prototypes, the Image-Specific CAM (IS-CAM) is computed as follows:

$$\tilde{\boldsymbol{M}}_k(j) = ReLU\left(\frac{\boldsymbol{F_h}(j) \cdot \boldsymbol{P}_k}{||\boldsymbol{F_h}(j)|| \cdot ||\boldsymbol{P}_k||}\right), \quad (7)$$

where $\tilde{\boldsymbol{M}}_k(j)$ is the $k$-th IS-CAM at pixel $j$. The correlation is bounded in $[-1, 1]$, and followed by ReLU to remove negative correlations.

Compared to the original CAM that takes classifier weight as the class center to compute the correlation of each pixel, the proposed IS-CAM utilizes prototypes tailored for every image to achieve more complete object regions. Besides, background prototype modeling provides high-quality background localization cues, which in turn help determine accurate foreground regions.

### 3.3. Self-supervised Learning with GSC

To further take advantage of image-specific knowledge, we introduce a self-supervised learning paradigm. The

overall training loss consists of multi-label classification loss and the General-Specific Consistency (GSC) loss,

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{gsc}. \tag{8}$$

The classification loss is computed by a multi-label soft margin loss between the image-level category label $y$ and the prediction $\hat{y}$, which is obtained by averaging over foreground maps generated by CAM.

$$\mathcal{L}_{cls} = \frac{1}{K} \sum_{i=1}^{K} y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i)), \tag{9}$$

where $\sigma$ is the sigmoid activation function.

The GSC is employed to minimize the difference between original CAM activated by classifier weight and IS-CAM activated by image-specific prototypes. The mathematical definition of this consistency regularization is formulated as the L1 normalization of the two kinds of CAM:

$$\mathcal{L}_{gsc} = \frac{1}{K+1} ||\boldsymbol{M} - \tilde{\boldsymbol{M}}||_1, \tag{10}$$

where $\boldsymbol{M}, \tilde{\boldsymbol{M}}$ denotes the original CAM and IS-CAM respectively. The loss is averaged over $K$ foreground classes and one background class.

With this consistency, the image-specific knowledge is injected into the feature representation, and the collaborative optimization is accomplished in the training cycles. IS-CAM forces the original CAM to pay attention to the absent object regions, which implicitly narrows the feature distance between the discriminative and missing pixels. Besides, the enhanced semantic and hierarchical features are favorable to capture more comprehensive and accurate image-specific prototypes and improve the quality of localization maps.

## 4. Experiments

In this section, we first elaborate the experimental settings including dataset, evaluation metric and implementation details. Second, we compare our method with state-of-the-art approaches on the PASCAL VOC 2012 dataset [11] and MS COCO 2014 dataset [30]. Third, we conduct a series of ablation studies to verify the effectiveness of the proposed method.

### 4.1. Experimental Settings

**Dataset and Evaluated Metric.** We evaluate our proposed method on PASCAL VOC 2012 segmentation benchmark [11] with 20 foreground classes and one background class. The official dataset split contains 1,464 images for training, 1,449 for validation and 1,456 for testing. Following the common experimental protocol in semantic segmentation, we take additional annotations from SBD [15]

to build an augmented training set with 10,582 images. Another MS COCO 2014 dataset has totally 81 classes and contains 80k train and 40k validation images, which is challenging for weakly supervised semantic segmentation. Note that only image-level classification labels are available during network training for both datasets. Mean intersection over union (mIoU) is used as a metric to evaluate segmentation results. The results for the PASCAL VOC test set are obtained from the official evaluation server.

**Implementation Details.** In our experiments, the ImageNet [9] pretrained ResNet50 [16] is adopted as backbone with output stride of 16, where fully connected layer is replaced by a classifier with output channels of 20. The augmentation strategy is the same to [25], including random flipping, random scaling and crop. The model is trained with a batch size of 16 on 2 Nvidia A100 GPUs. SGD optimizer is adopted to train our model for 5 epochs, with a momentum of 0.9 and a weight decay of 1e-4. The learning rates for the backbone and the newly added layers are set as 0.1 and 1, respectively. We use poly learning scheduler decayed with a power of 0.9 for the learning rate.

**Inference** At the inference stage, the network generates foreground and background seeds by hierarchical features and activates to localization maps. Instead of checking various mIoU scores over the training set to obtain pseudo labels as other works [25, 42], we directly compute pseudo labels with the background localization maps.

### 4.2. Comparison with State-of-the-arts

**Improvements on localization maps.** We first evaluate mIoU on localization maps, where those maps are generated by the proposed IS-CAM. Tab. 1 presents the comparison with other advanced methods on PASCAL VOC 2012 train set. Among these compared methods, ECS [37] provides the best results with a mIoU of 56.6%. Our proposed SIPE achieves the state-of-the-art performance of 58.6%. Furthermore, we report the performance with denseCRF post-processing. The results show that our SIPE with denseCRF improves the mIoU to 64.7% and outperforms all other methods, which may be benefited from our high-quality localization maps. We interpret this performance gain comes from the capability of SIPE to generate complete localization maps for foreground and background. As the localization maps can capture satisfying boundaries of semantic objects, the denseCRF is less confusing while refining the CAMs. Fig. 5 shows the visual comparison of foreground localization maps on PASCAL VOC 2012 train set. It can be observed that our SIPE is effective in capturing the whole semantic regions under various scenes, such as different object scales, crowded objects, and multiple categories. The
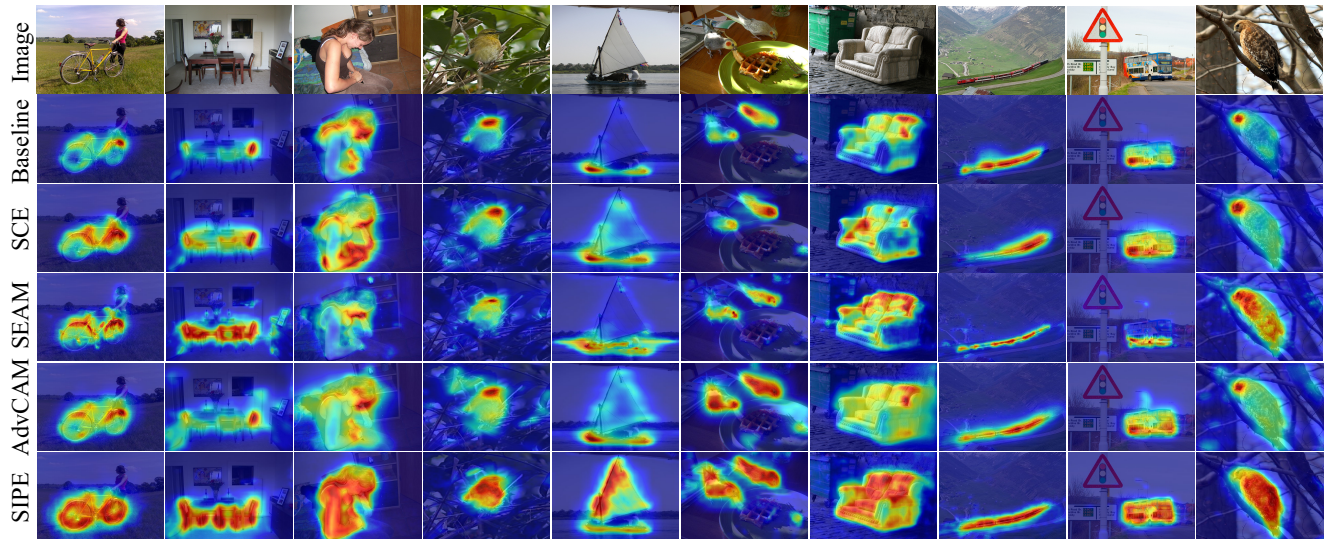
Figure 5. Visual comparison of localization maps generated by different methods on PASCAL VOC 2012 train set. From top to down: original image, Baseline, SCE [4], SEAM [42], AdvCAM [25] and our SIPE.

Table 1. mIoU (%) of localization maps on PASCAL VOC 2012 *train* set. The best results are shown in bold.

| Method | Pub. | Local. Maps | +denseCRF |
|--------|------|-------------|-----------|
| Baseline | - | 50.1 | 54.3 |
| SCE [4] | CVPR20 | 50.9 | 55.3 |
| SEAM [42] | CVPR20 | 55.4 | 56.8 |
| EDAM [45] | CVPR21 | 52.8 | 58.2 |
| AdvCAM [25] | CVPR21 | 55.6 | 62.1 |
| ECS [37] | ICCV21 | 56.6 | 58.6 |
| CSE [23] | ICCV21 | 56.0 | 62.8 |
| SIPE (Ours) | | **58.6** | **64.7** |

obtained high-quality localization maps will further benefit our segmentation results.

**Improvements on segmentation results.** To further evaluate the performance of our methods, we train fully supervised models using generated pseudo labels and compare the segmentation results with the state-of-the-arts. Following the common practice [25], the pseudo labels are refined by IRN [1] and used to train DeepLabV2 [7]. Tab. 2 presents the comparison with state-of-the-art methods on PASCAL VOC 2012 val and test sets. Using only image-level labels, our SIPE outperforms previous methods with 68.8% mIoU on val set and 69.7% mIoU on test set. In addition, our method performs favorably against NSROM [48] and EPS [27], which introduce saliency maps as auxiliary

Table 2. Comparison with other state-of-the-arts on PASCAL VOC 2012 *val* and *test* sets. The best results among methods only using image-level labels are shown in bold.

| Model | Pub. | Sup. | Backbone | Val | Test |
|-------|------|------|----------|-----|------|
| DSRG [19] | CVPR18 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 61.4 | 63.2 |
| SeeNet [18] | NIPS18 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 63.1 | 62.8 |
| FickleNet [24] | CVPR19 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 64.9 | 65.3 |
| OAA+ [20] | ICCV19 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 65.2 | 66.4 |
| G-WSSS [28] | AAAI21 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 68.2 | 68.5 |
| NSROM [48] | CVPR21 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 70.4 | 70.2 |
| EPS [27] | CVPR21 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 71.0 | 71.8 |
| AuxSegNet [47] | ICCV21 | $\mathcal{I} + \mathcal{S}$ | ResNet38 | 69.0 | 68.6 |
| IRN [1] | CVPR19 | $\mathcal{I}$ | ResNet50 | 63.5 | 64.8 |
| ICD [12] | CVPR20 | $\mathcal{I}$ | ResNet101 | 64.1 | 64.3 |
| SCE [4] | CVPR20 | $\mathcal{I}$ | ResNet101 | 66.1 | 65.9 |
| SEAM [42] | CVPR20 | $\mathcal{I}$ | ResNet38 | 64.5 | 65.7 |
| BES [6] | ECCV20 | $\mathcal{I}$ | ResNet101 | 65.7 | 66.6 |
| MCIS [36] | ECCV20 | $\mathcal{I}$ | ResNet101 | 66.2 | 66.9 |
| CONTA [10] | NIPS20 | $\mathcal{I}$ | ResNet101 | 66.1 | 66.7 |
| LIID [32] | TPAMI20 | $\mathcal{I}$ | ResNet101 | 66.5 | 67.5 |
| A2GNN [49] | TPAMI21 | $\mathcal{I}$ | ResNet101 | 66.8 | 67.4 |
| AdvCAM [25] | CVPR21 | $\mathcal{I}$ | ResNet101 | 68.1 | 68.0 |
| CDA [35] | ICCV21 | $\mathcal{I}$ | ResNet38 | 66.1 | 66.8 |
| ECS [37] | ICCV21 | $\mathcal{I}$ | ResNet38 | 66.6 | 67.6 |
| CSE [23] | ICCV21 | $\mathcal{I}$ | ResNet38 | 68.4 | 68.2 |
| CPN [51] | ICCV21 | $\mathcal{I}$ | ResNet38 | 67.8 | 68.5 |
| SIPE (Ours) | | $\mathcal{I}$ | ResNet38 | 68.2 | 69.5 [1] |
| SIPE (Ours) | | $\mathcal{I}$ | ResNet101 | **68.8** | **69.7** [2] |

---

[1] http://host.robots.ox.ac.uk:8080/anonymous/NGICBM.html
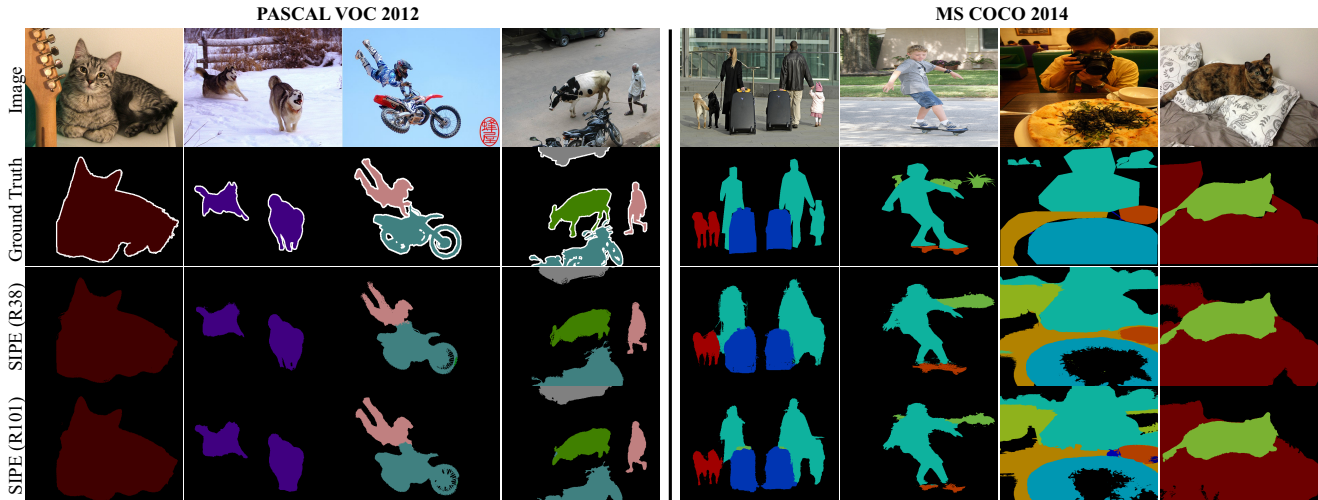[2] http://host.robots.ox.ac.uk:8080/anonymous/UU6VNX.html

Figure 6. Qualitative segmentation results on the validation set of PASCAL VOC 2012 and MS COCO 2014.

Table 3. Comparison with other state-of-the-arts on MS COCO 2014 *val* set. The best results are shown in bold.

| Model | Pub. | Sup. | Backbone | Val |
|---|---|---|---|---|
| DSRG [19] | CVPR18 | $\mathcal{I} + \mathcal{S}$ | VGG16 | 26.0 |
| G-WSSS [28] | AAAI21 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 28.4 |
| EPS [27] | CVPR21 | $\mathcal{I} + \mathcal{S}$ | ResNet101 | 35.7 |
| AuxSegNet [47] | ICCV21 | $\mathcal{I} + \mathcal{S}$ | ResNet38 | 33.9 |
| SEC [22] | ECCV16 | $\mathcal{I}$ | VGG16 | 22.4 |
| IRN [1] | CVPR19 | $\mathcal{I}$ | ResNet50 | 32.6 |
| IAL [41] | IJCV20 | $\mathcal{I}$ | VGG16 | 27.7 |
| SEAM [42] | CVPR20 | $\mathcal{I}$ | ResNet38 | 31.9 |
| CONTA [10] | NIPS20 | $\mathcal{I}$ | ResNet101 | 33.4 |
| CSE [23] | ICCV21 | $\mathcal{I}$ | ResNet38 | 36.4 |
| SIPE (Ours) | | $\mathcal{I}$ | ResNet38 | **43.6** |
| SIPE (Ours) | | $\mathcal{I}$ | ResNet101 | 40.6 |

Table 4. Effect of the main contributions. CAM: orginal CAM, IPE: image-specific prototype exploration, GSC: general-specific consistency.

| CAM | IPE | GSC | mIoU (%) |
|---|---|---|---|
| ✓ | | | 50.1 |
| ✓ | ✓ | | 53.2 |
| ✓ | ✓ | ✓ | 58.6 |

forms EPS [27] by 4.9%. These outstanding performances over existing state-of-the-arts on both datasets confirm the effectiveness of our SIPE, which well explores image-specific prototypes via self-supervised learning paradigm.

### 4.3. Ablation Studies

**Effect of the main contributions.** We conduct an ablation study to verify the effect of the proposed two key contributions, i.e., Image-specific Prototype Exploration (IPE) and General-Specific Consistency (GSC). As shown in Tab. 4, with image-specific prototypes, our IS-CAM can improve the original CAM by 2.1% on mIoU score. To enhance the feature representation, GSC is introduced for self-supervised training. The proposed GSC improves the quality of IS-CAM (5.4%) by a clear margin. By combining these two methods, our full method performs significantly better than the original CAM. Fig. 7 visualizes CAM and IS-CAM with different settings. From the first two rows, we can observe that image-specific prototypes can activate more useful regions. Additionally, our IS-CAM shows that the proposed method produces more clear background activations than that from CAM. When training our model with GSC, the quality of localization maps is obviously improved, especially for background.

labels for this task. For a fair comparison, we also train the model with ResNet38 following the default setting as [42]. Our SIPE achieves 69.5% mIoU on test set, exceeding the existing methods that use ResNet38 backbone. The qualitative segmentation results on val set are shown in Fig. 6. Based on our SIPE, DeepLabV2 shows more robustness to various challenging scenarios, such as different object scales, multiple objects and multiple categories.

In Tab. 3, we also evaluate our method in MS COCO 2014 dataset [30]. The same training script with the experiment on PASCAL VOC 2012 is employed, but no refinement with IRN due to the large computation cost. Our method achieves 43.6% mIoU with ResNet38 backbone on the validation set, which is 7.2% higher than previous SOTA CSE [23]. Using ResNet101 backbone, we also outper-
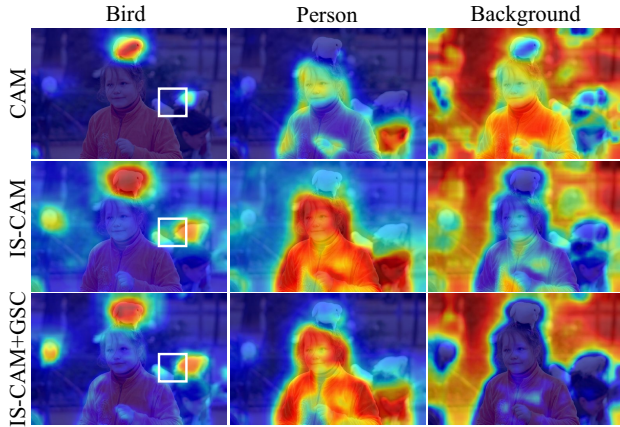
Figure 7. Visualization of CAM and IS-CAM with different settings. Our IS-CAM shows more complete regions of person. Moreover, the proposed GSC further refines the background activations. The boundary is well captured as the white box.
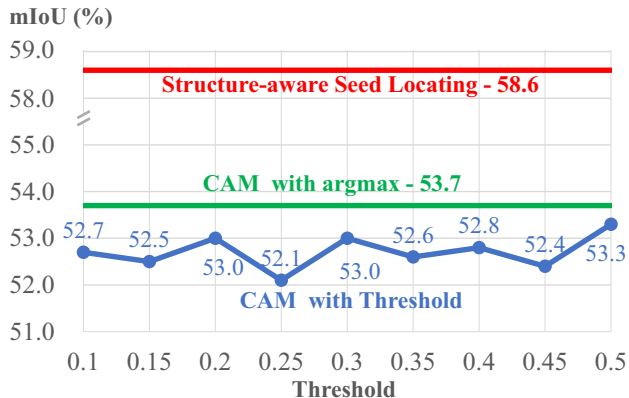


Figure 8. Ablation study of structure-aware seed locating. The proposed structural-aware seed locating method is consistently better than CAM with fixed threshold, as well as CAM with adaptive threshold.

**Effect of structure-aware seed locating.** To verify the effectiveness of the proposed structure-aware seed locating, we compare the mIoU with other seed generation approaches including applying threshold or argmax operations on CAM. As shown in Fig. 8, applying different thresholds on CAM exhibits fluctuations and can only achieve maximal mIoU of 53.3%. In addition, simply applying argmax results in slightly performance gains (+0.4%). These methods are still hard to find compete regions because both of them only depend on the single pixel's probability. In contrast, the proposed structure-aware seed locating largely outperforms the above methods owing to the proposed structure information.

**Effect of prototype modeling.** We conduct ablation study of prototype modeling on localization maps. The IoUs of four options concerning feature and Background

Table 5. Ablation study of combinations of feature selection and Background Prototype Modeling (BPM) via GSC. Each item reports mIoUs for two pseudo labels where the former are generated by searching the best background threshold [42] and the latter are generated by estimated background map. The best results are shown in bold.

| Feature | w/o BPM | | w/ BPM | |
| --- | --- | --- | --- | --- |
| | Threshold | Estimated | Threshold | Estimated |
| Semantic | 56.0 | 54.6 | 55.9 | 54.3 |
| Hierarchical | 53.9 | 51.2 | **57.6** | **58.6** |

Prototype Modeling (BPM) are presented in Tab. 5: (1) Semantic feature without BPM, (2) Semantic feature with BPM, (3) Hierarchical feature without BPM, and (4) Hierarchical feature with BPM. We report threshold-based and estimated-based pseudo labels for each type of option. Our SIPE adopts option (4) where hierarchical feature is learned for image-specific background modeling. It shows that our method achieves the highest performance in both values among all options. From (1) to (2), the mIoU is basically unchanged because the background often does not have specific semantics. From (1) to (3), the mIoU drops over 2% since the low-level information brings confusion to foreground localization. It is worth noting that only when hierarchical feature and background modeling are employed at the same time, the mIoU of the estimated-based pseudo labels exceed that bases on threshold, which strongly proves the effectiveness of estimating background cues.

## 5. Conclusion

In this paper, we propose Self-supervised Image-specific Prototype Exploration (SIPE) for weakly supervised semantic segmentation. In our framework, an Image-specific Prototype Exploration (IPE) is proposed to achieve more favorable localization maps. It is achieved by structure-aware seed locating and background-aware prototype modeling. In addition, a General-Specific Consistency (GSC) loss is developed to efficiently regularize the general CAM and Image-Specific CAM (IS-CAM), empowering the feature representation. Extensive experiments show that our SIPE sets new state-of-the-art performance on two well-known benchmarks using image-level labels.

# References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 6, 7

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, pages 549–565, 2016. 1

[4] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 2, 6

[5] Hongjun Chen, Jinbao Wang, Hong Cai Chen, Xiantong Zhen, Feng Zheng, Rongrong Ji, and Ling Shao. Seminar learning for click-level weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6920–6929, October 2021. 1

[6] 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, 2020. 6

[7] 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 Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 6

[8] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[10] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2020. 6, 7

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 2, 5

[12] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceed-

[13] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. 2

[14] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1

[15] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 991–998, 2011. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[17] Mohammad D Hossain and Dongmei Chen. Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:115–134, 2019. 1

[18] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 547–557, 2018. 2, 6

[19] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 3, 6, 7

[20] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 2, 6

[21] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1754–1761, 2021. 2

[22] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, 2016. 7

[23] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6994–7003, October 2021. 2, 6, 7

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In

*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 2, 6

[25] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, June 2021. 5, 6

[26] 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 Conference on Computer Vision and Pattern Recognition*, pages 2643–2652, 2021. 1

[27] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5505, June 2021. 2, 6, 7

[28] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1984–1992, May 2021. 2, 6, 7

[29] 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. 1

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014. 2, 5, 7

[31] Weide Liu, Chi Zhang, Guosheng Lin, Tzu-Yi Hung, and Chunyan Miao. Weakly supervised segmentation with maximum bipartite graph matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2085–2094, 2020. 2

[32] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[33] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020. 3

[34] Zhiyi Pan, Peng Jiang, Yunhai Wang, Changhe Tu, and Anthony G. Cohn. Scribble-supervised semantic segmentation by uncertainty reduction on neural representation and self-supervision on neural eigenspace. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7416–7425, October 2021. 1

[35] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. 2021. 6

[36] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised seman-

tic segmentation. In *European Conference on Computer Vision*, pages 347–365, 2020. 2, 6

[37] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7283–7292, October 2021. 2, 5, 6

[38] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 1

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[40] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019. 3

[41] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *International Journal of Computer Vision*, 128(6):1736–1749, 2020. 7

[42] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2, 3, 5, 6, 7, 8

[43] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1576, 2017. 2

[44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2

[45] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, June 2021. 2, 6

[46] Jingshan Xu, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang. Scribble-supervised semantic segmentation inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15354–15363, October 2021. 1

[47] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised se-

mantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 6, 7

[48] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2623–2632, June 2021. 2, 6

[49] Bingfeng Zhang, Jimin Xiao, Jianbo Jiao, Yunchao Wei, and Yao Zhao. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 6

[50] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772, 2020. 2

[51] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 6

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

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, June 2016. 1