A Pixel-Level Meta-Learner for Weakly Supervised Few-Shot Semantic Segmentation

Yuan-Hao Lee Fu-En Yang Yu-Chiang Frank Wang
Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, R.O.C.
ASUS Intelligent Cloud Services, Taiwan, R.O.C.
{r07942074, f07942077, ycwang}@ntu.edu.tw

Abstract

Few-shot semantic segmentation addresses the learning task in which only few images with ground truth pixel-level labels are available for the novel classes of interest. One is typically required to collect a large mount of data (i.e., base classes) with such ground truth information, followed by meta-learning strategies to address the above learning task. When only image-level semantic labels can be observed during both training and testing, it is considered as an even more challenging task of weakly supervised few-shot semantic segmentation. To address this problem, we propose a novel meta-learning framework, which predicts pseudo pixel-level segmentation masks from a limited amount of data and their semantic labels. More importantly, our learning scheme further exploits the produced pixel-level information for query image inputs with segmentation guarantees. Thus, our proposed learning model can be viewed as a pixel-level meta-learner. Through extensive experiments on benchmark datasets, we show that our model achieves satisfactory performances under fully supervised settings, yet performs favorably against state-of-the-art methods under weakly supervised settings.

1. Introduction

Recent advances in deep convolutional neural networks (CNNs) [19] have significantly improved the performances of several computer vision tasks. Among them, *semantic segmentation* aims at predicting class labels for each pixel in an image, with applications ranging from autonomous driving to medical imaging. With the help of CNNs, state-of-the-art semantic segmentation models including FCN [25], SegNet [1] and DeepLabs [5, 6, 7, 8] have all achieved very promising results and been successfully applied to the above applications. However, these models are generally trained in a fully supervised manner, requiring a huge amount of training data with pixel-level annotation for each category of interest. This substantially limits the scalability and practicality of these models, as collecting densely labeled data would be very time-consuming.

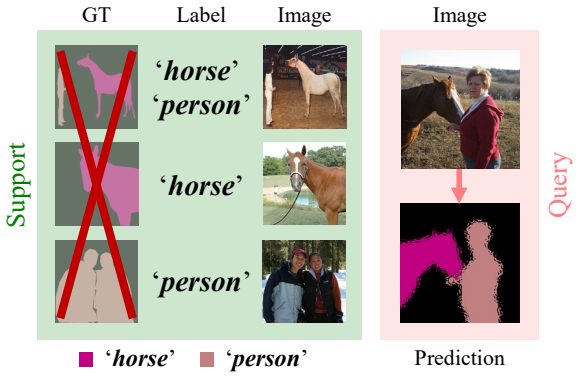

Figure 1: Illustration of weakly supervised few-shot segmentation. With only semantic labels but not pixel-level masks observed during few-shot training and testing, semantic segmentation of particular image categories can be achieved. Note that a 2-way 3-shot scheme is depicted.

Extended from the learning task of semantic segmentation and few-shot learning, *few-shot semantic segmentation* considers a more challenging setting in which only a few images are with ground-truth pixel-level labels for the (novel) classes of interest. In order to realize the learning of few-shot segmentation models, existing methods typically adopt meta-learning schemes [13, 30, 34], utilizing support and query images sampled from base categories (i.e., those with a sufficient amount of training data) for performing pixel-wise classification. Recent methods like AMP [38], PANet [46], FWB [27], PFENet [42] and ASGNet [20] choose to extract prototypes from support-set images using their ground truth masks, and expect such prototypes to sufficiently describe the associated semantic category. Nevertheless, these existing methods require pixel-level ground truth labels for each image of the base categories, and can-

not be easily extended to weakly supervised settings.

In order to alleviate the requirement of annotating pixel-level ground truth label information, we consider an even more challenging yet practical setting of weakly supervised few-shot segmentation, which requires only image-level labels collected for images in both base and novel categories, as depicted in Figure 1. As shown in the figure, we aim at inferring pixel-level labels from image-level labels in few-shot settings, followed by a meta-learning scheme enforced at pixel-level for segmentation purposes.

With this goal in mind, we propose a novel learning scheme in this paper, focusing on deriving a pixel-level meta learner for weakly supervised few-shot semantic segmentation. During the meta-training stage, our proposed learning framework observes only image-level labels and utilizes Classification Activation Maps (CAM) [59, 35] for prediction of pseudo pixel-level labels for each input image. While the class label embedding [2, 26] is exploited to bridge the gap between image and pixel-level information, our proposed method does not encounter any information leak since the class labels of both base and novel categories are *not* present in those of CAM.

Under the guidance of the produced pseudo pixel-level labels, we uniquely reinterpret the original problem as a *pixel-wise few-shot classification* task. That is, we view each pixel in support/query set images as individual samples, and turn the proposed model into a pixel-level meta learner for few-shot semantic segmentation. As confirmed later by our experiments, our model not only achieves satisfactory performances on standard fully supervised few-shot semantic segmentation tasks, it would perform favorably against several state-of-the-art approaches on benchmark datasets under weakly supervised settings.

The contributions of this work are summarized below:

- We address weakly supervised few-shot semantic segmentation, which requires only image-level labels during both training and testing.
- We bridge the gap between image and pixel-level labels using classification activation maps with no information leak, which allows prediction of pseudo pixel-level labels in weakly supervised settings.
- We uniquely design a pixel-level meta learner which enforces segmentation consistency across support and query set images during few-shot semantic segmentation. The proposed model can be realized in both fully supervised and weakly supervised settings.

2. Related Works

Semantic Segmentation. The task of semantic segmentation aims at performing pixel-level classification for each image. Most recent methods involve the use of deep convolutional neural networks (e.g., FCN [25]). Following works include SegNet [1], PSPNet [58], U-Net [32], DeepLabs [5, 6, 7, 8] and FastFCN [47]. A notable improvement is achieved by embedding features that contain multi-scale context information by applying dilated convolution [53, 5] or spatial pyramid pooling [58, 6]. These methods, however, are trained in a fully supervised manner and require a huge amount of pixel-level labels, which are time-consuming and laborious to obtain.

Weakly Supervised Semantic Segmentation. To alleviate the need for densely annotated ground truth data, some recent works address semantic segmentation in weak supervision, which utilize multiple-instance learning [43, 44], graph [55, 49, 28] and self-training based [50, 51, 56, 48] techniques. However, such weakly supervised methods cannot be easily extended to few-data or open-set scenarios.

Few-Shot Semantic Segmentation. Few-shot learning aims at learning models which would generalize to categories with only a limited amount of labeled data [12, 11]. *Meta-learning* [13, 30, 34] has been widely applied for this task, with the core idea of adapting the learning scheme from base to novel categories. For example, metric-based meta-learning algorithms such as ProtoNet [40] and RN [41] learn feature embeddings that exhibit proper distance metrics for classification and generalization. *Few-shot semantic segmentation*, on the other hand, aims at generalizing the ability of pixel-level classification across categories, while only a limited number of images are with ground truth pixel-level labels. OSLSM [36] is the first proposed method to tackle this problem, leveraging information learned from support-set images and outputs parameters for query image segmentation. PL [9] adopts metric learning methods [40] to extract prototypes of each semantic class, and measures their distances between feature maps of query images; CANet [54] adds an iterative optimization module to refine the predicted results; PFENet [42] generates additional prior masks to enrich the extracted features.

Moreover, PANet [46], FWB [27] and CRNet [23] propose to further leverage information from the support set by performing segmentation in the reversed direction (i.e., segmentation of the support set) for improved model learning. To exploit knowledge from the foreground objects, DAN [45] and SimPropNet [14] introduce attention mechanisms, PMMs [52] and ASGNet [20] employ multiple prototypes for a single category, and PPNet [24] proposes part-aware prototypes to capture fine-grained features. While some of the existing few-shot semantic segmentation methods present results in weak supervision settings (e.g., use of bounding boxes or scribbles [29, 46, 54] as guidance), they cannot produce satisfactory performance with only image-level annotation observed. More recently, [31, 4, 37] follow the few-shot setting using image-level supervision, but they

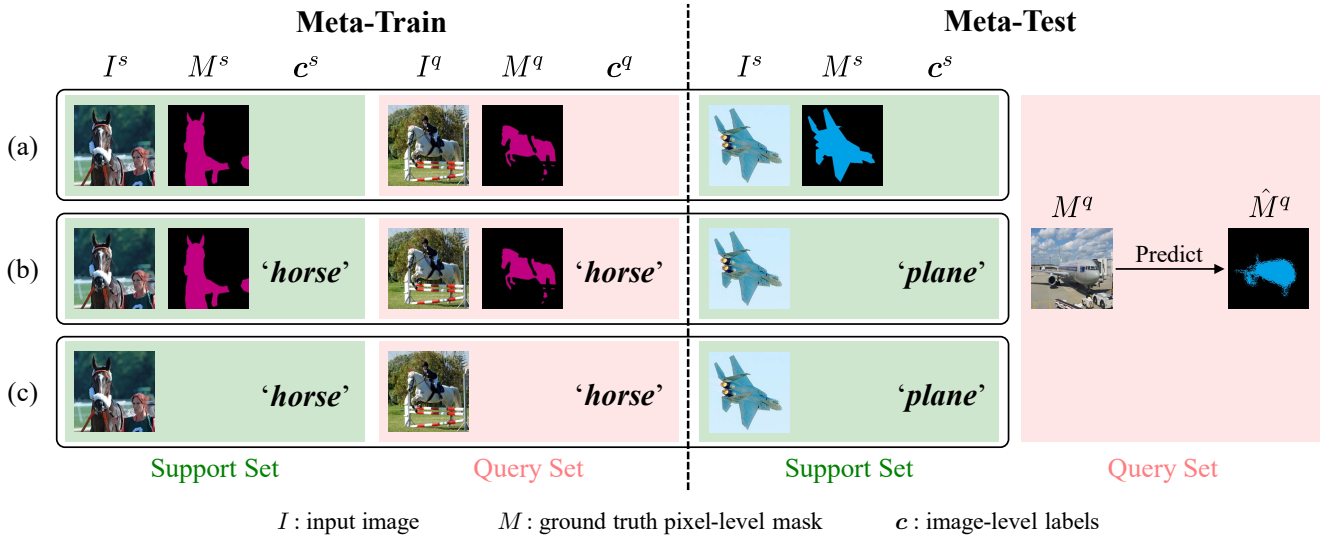

Figure 2: Comparisons of different *few-shot* semantic segmentation schemes. (a) *Fully Supervised* [38, 46, 54, 27, 42]: M required in meta-training and meta-testing; (b) *Loosely Weakly Supervised* [37]: both M and c available during meta-training, while only c observed in meta-testing; (c) *Weakly Supervised*: only image-level labels c available in both phases.

Method	Weak Supervision	Few-Shot Setting
DeepLab [5]	-	-
EDAM [48]	✓	-
PFENet [42]	-	✓
Co-att [37]	△	✓
Ours	✓	✓

Table 1: Comparisons of different semantic segmentation methods. Existing methods cope with either weak supervision or few-shot settings, while ours combine both. Note that [37] still requires full supervision during training, as detailed in Figure 2(b).

still require collection of ground truth pixel-level masks for base-class images during meta-training. As depicted in Figure 2 and Table 1, to the best of our knowledge, we are the first to tackle few-shot semantic segmentation using only image-level annotations during both (meta) training and testing stages.

3. Proposed Method

3.1. Notation and Problem Formulation

For the sake of completeness, we define the notations which will be used in this paper. The semantic classes are denoted as $\mathcal{C}$, which are split into two disjoint subsets: base categories $\mathcal{C}_{\text{base}}$ and novel categories $\mathcal{C}_{\text{novel}}$ (i.e., $\mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}} = \mathcal{C}$ and $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \varnothing$). Note that the novel categories are with only a few samples available during training (typically less than 5 per class). For each input RGB image $I \in \mathbb{R}^{H\times W\times 3}$, we denote its image-level labels as $\boldsymbol{c} \subseteq \mathcal{C}$, and the associated ground truth pixel-level semantic mask as $M \in \{\boldsymbol{c} \cup 0\}^{H\times W}$, where 0 indicates the background pixels. For each training episode, we sample a support/query pair from the image dataset $\mathcal{D}_{\text{train}} = \{(S_j, Q_j)\}_{j=1}^{n_{\text{train}}}$ that contain the same set of base categories, while the testing episodes consist of those from $\mathcal{C}_{\text{novel}}$ (i.e., $\mathcal{D}_{\text{test}} = \{(S_j, Q_j)\}_{j=1}^{n_{\text{test}}}$).

In standard fully supervised N-way K-shot settings, each support set $S_j = \{(I_i^s, M_i^s)\}_{i=1}^{N\times K}$ contains $N \times K$ image/label example pairs (K pairs from each of N categories), and the query sets are denoted as $Q_j = \{(I_i^q, M_i^q)\}_{i=1}^{n_q}$ where n_q is the number of query images in a single episode. As for the *weakly supervised* setting considered in this work, only image-level labels are available during training and testing, so that support/query sets are denoted as $S_j = \{(I_i^s, \boldsymbol{c}_i^s)\}_{i=1}^{N\times K}$ and $Q_j = \{(I_i^q, \boldsymbol{c}_i^q)\}_{i=1}^{n_q}$, respectively. Our proposed weakly supervised learning framework is able to produce pseudo masks of each image given only its class label (as depicted in Figure 3), followed by our pixel-level meta-learner for few-shot semantic segmentation (see Figure 4 for the complete framework).

3.2. Pixel-Level Pseudo Label Generation

In the weakly supervised scenario, only the class labels are available for the image data during both training and inference. Thus, we first present a module that is able to generate pseudo pixel-level semantic masks, guiding the following segmentation process. As shown in Figure 3, this pseudo pixel-level label generation can be viewed as the process of semantics-oriented heatmap extraction. As

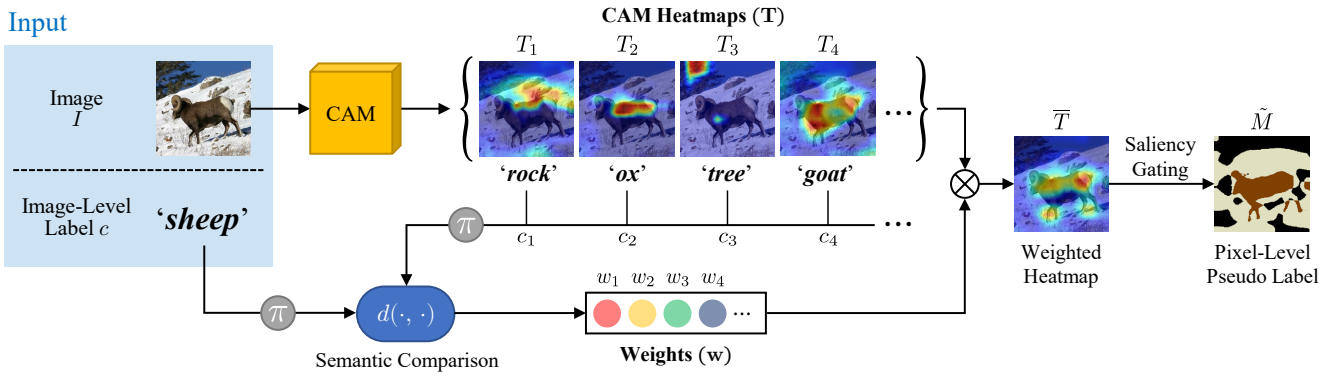

Figure 3: Illustration of pixel-level pseudo label generation. Given an input I and its image-level label c, the CAM module extracts heatmaps $\mathbf{T}$ for each training class c_k of CAM (not in $\mathcal{C}_{\text{base}}$ nor $\mathcal{C}_{\text{novel}}$). $\mathbf{w}$ denotes the visual similarity for the heatmap of each CAM category, which is measured by the distance between the word embeddings of the associated class labels. The output heatmap $\overline{T}$ is converted into the pseudo mask $\tilde{M}$ for I via saliency gating. Note that π represents the word embedding function.

adopted in previous works like [18], Classification Activation Maps (CAM) [59] have been utilized to localize discriminative regions in images that are informative in classifying image-level labels. In order to produce the heatmap for the input image I as its pseudo pixel-level labels, we follow [46, 54] and apply a VGG-16 [39] network as our CAM backbone. It is worth noting that, the CAM backbone is pre-trained on a reduced subset of ImageNet [33], in which the images do not belong to either the base or novel categories in our segmentation task. This would minimize possible leakage of semantic information.

With CAM obtained, we extract per-class heatmaps $\mathbf{T}$ of the input image I:

$$\text{CAM}(I) = \mathbf{T} = [T_1, T_2, \ldots, T_{N_{\text{CAM}}}], \tag{1}$$

where $T_k \in [0,1]^{H \times W}$ denotes the heatmap of the k-th image class in CAM, whereas N_{CAM} is the total number of pre-training image categories.

With the above per-class heatmaps observed, we next leverage the word embedding features of each class label as intermediate representations, with the associated similarity indicating the weight for each of the N_{CAM} categories. More specifically, as depicted in Figure 3, we extract word embedding features of each class label in CAM and that of the input image. We perform pairwise comparisons between these features to obtain weighting factors w_k for each CAM category c_k:

$$w_k = d(\pi(c), \pi(c_k))^{-1}, \tag{2}$$

where $\pi(\cdot)$ represents the word embedding function. We note that $d(\cdot,\cdot)$ denotes the distance metric, and we use consine similarity in our work. Thus, categories that are semantically similar with each other (e.g., *goat*/*sheep*) would result in a higher weight, and vice versa for those that are dissimilar (e.g., *goat*/*tree*). As a result, a weighted heatmap $\overline{T}$ can be obtained by averaging the above N_{CAM} heatmaps, which is calculated as

$$\overline{T} = \mathbf{w} \cdot \mathbf{T}^\top = \sum_{k=1}^{N_{\text{CAM}}} w_k T_k. \tag{3}$$

As the CAM heatmaps identify only regions that are helpful in terms of classification, more detailed structural information such as edges and boundaries would not be well described. To this end, we impose a class-agnostic saliency map on the weighted heatmap as a gating mechanism, alleviating the presence of false-positive pixels in the generated pseudo labels. Here we note that, to comply with our weakly supervised setting, the saliency maps are obtained via a network pre-trained with only foreground/background information without any categorical supervision. In other words, minimized leakage of semantic information is also enforced at this stage.

3.3. Pixel-Level Meta-Learner for Few-Shot Semantic Segmentation

We now detail our proposed meta-learning scheme for few-shot semantic segmentation. Under the guidance of CAM, the pseudo pixel-level labels obtained in Section 3.2 tend to contain only partial discriminative areas, and may not sufficiently cover the entire foreground object. In order to learn few-shot segmentation models with only such information from pseudo labels, we present a unique pixel-level meta-learning framework, as illustrated in Figure 4. In our pixel-level meta-learner, we utilize DeepLabv3+ [8] as the feature extractor, together with the introduced pseudo label generation and learning modules for pixel-wise classifica-

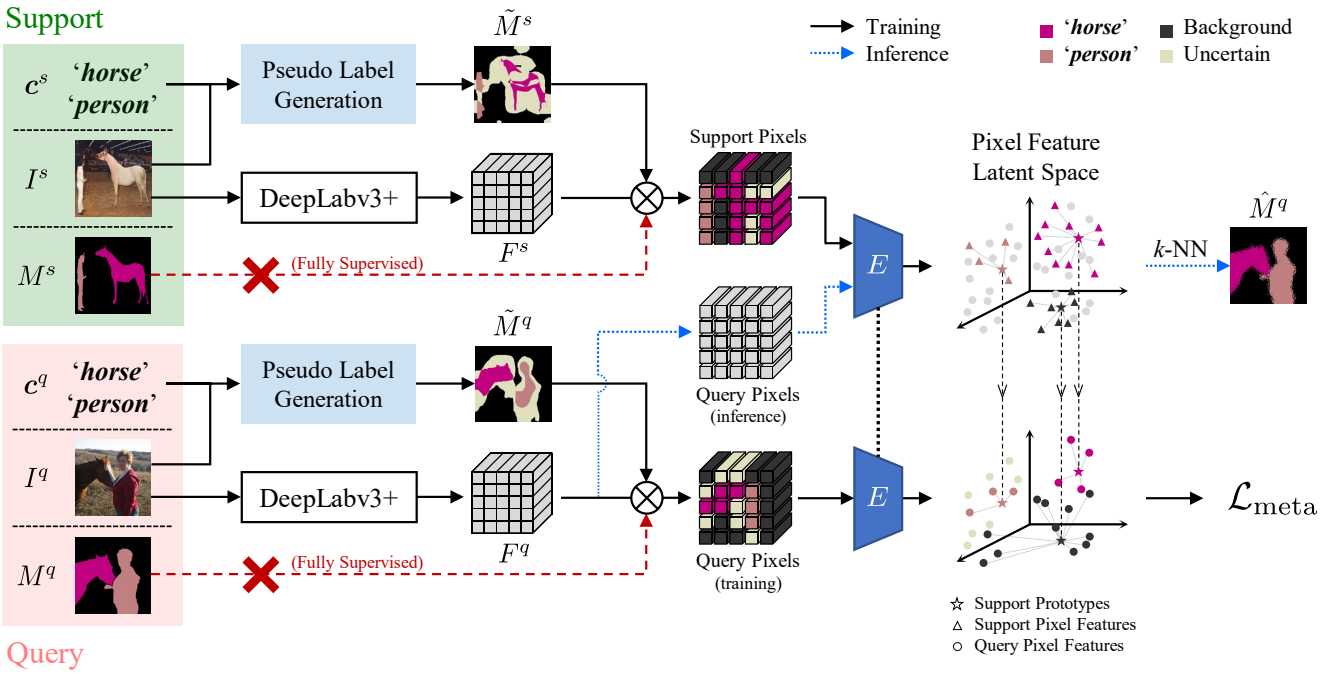

Figure 4: Architecture of our pixel-level meta-learner for weakly supervised segmentation. During (meta) training, DeepLabv3+ extracts pixel-wise feature maps F^s and F^q for support and query inputs, with the produced pseudo pixel-level labels $\tilde{M}^s$ and $\tilde{M}^q$, respectively. Our meta-learner encoder E transforms the above F into a latent space, in which the pixel-level prototypical loss $\mathcal{L}_{\text{meta}}$ can be calculated based on $\tilde{M}$ for segmentation purposes. During inference (i.e., meta-testing), segmentation $\hat{M}^q$ for the query can be performed by pixel-wise k-NN classification using F^s and $\tilde{M}^s$.

tion. It is worth noting that, the DeepLabv3+ is pre-trained on categories *not* appeared in $\mathcal{C}$, and this module remains fixed throughout the meta-learning process. In other words, no semantic information is leaked from its training stage. By removing the final classification layers of DeepLabv3+, we obtain a pixel-wise feature map $F^s \in \mathbb{R}^{H\times W\times d}$ from the support image I^s, where (H, W) is the original image dimension and d is the feature channel size.

Next, we label each spatial location in F^s using the generated support pseudo mask $\tilde{M}^s$. By randomly sampling a fixed number of pixels from each background/foreground category, the resulting pixel-wise features are then collected into a set of *support pixel features*. It is worth noting that, we choose not to use all the labeled pixel features. This is not only because that such pseudo labels might not match the ground truth ones (although not available), this sampling mechanism also makes the meta-learning process more robust against weak labels.

Likewise, a pixel-wise feature map $F^q \in \mathbb{R}^{H\times W\times d}$ of the query image I^q is also obtained via the same DeepLabv3+ feature extractor, resulting in a total of $H \times W$ d-dimensional *query pixel features*. All support and query pixel features are then jointly embedded into a latent space via the learnable encoder E. With the support pixel features and pseudo labels obtained, we define the prototypes p_c for each associated category $c \in \{\boldsymbol{c}^s \cup 0\}$ as

$$p_c = \frac{\sum_l E(F_l^s)\mathbb{1}[\tilde{M}_l^s = c]}{\sum_l \mathbb{1}[\tilde{M}_l^s = c]}, \tag{4}$$

where l iterates over each pixel, and $\mathbb{1}[\cdot]$ is an indicator function which only outputs 1 when the condition holds.

Inspired by [40, 9, 42], we advance the prototypical loss $\mathcal{L}_{\text{meta}}$ as the objective during the meta-training stage, which is calculated by accumulating the distance between each query pixel (with pseudo labels) and its corresponding pixel-level prototype from the support sets:

$$\mathcal{L}_{\text{meta}} = -\frac{\sum_c \sum_l \exp(-d(E(F_l^q), p_c))\mathbb{1}[\tilde{M}_l^q = c]}{\sum_c \sum_l \mathbb{1}[\tilde{M}_l^q = c]}. \tag{5}$$

During inference (i.e., meta-testing), the labels of the embedded query pixel features are determined by performing pixel-wise k-nearest neighbors classification, as depicted by the blue dotted arrows in Figure 4. The final predicted query semantic mask $\hat{M}^q$ is then compared with the ground truth M^q for performance evaluation.

(*fully sup.* / *weakly sup.*)		1-shot						5-shot	
Method	Backbone	Split-0	Split-1	Split-2	Split-3	Mean	Δ	Mean	Δ
1-way									
Co-att [37]	VGG-16	49.5	65.5	50.0	49.2	53.5	—	51.7	—
AMP [38]	VGG-16	41.9 / 10.6	50.2 / 14.1	46.7 / 7.6	34.7 / 10.9	43.4 / 10.8	32.6	46.9 / 14.7	32.2
PANet [46]	VGG-16	42.3 / 25.7	58.0 / 33.4	51.1 / 28.8	41.2 / 20.7	48.1 / 27.1	21.0	55.7 / 37.7	18.0
PFENet [42]	ResNet-50	61.7 / 33.4	69.5 / 42.5	55.4 / 43.6	56.3 / 39.9	60.8 / 39.9	20.9	61.9 / 44.8	17.1
Ours	VGG-16	38.3 / 36.5	57.6 / 51.7	54.0 / 45.9	40.1 / 35.6	47.5 / **42.4**	**5.1**	50.6 / **45.5**	**5.1**
2-way									
PANet [46]	VGG-16	45.1† / 24.5	45.1† / 33.6	45.1† / 26.3	45.1† / 20.3	45.1 / 26.2	18.9	53.1 / 36.6	16.5
Ours	VGG-16	36.5 / 31.5	51.8 / 46.7	48.5 / 41.4	38.9 / 31.2	43.9 / **37.7**	**6.2**	49.3 / **43.0**	**6.3**

(a) PASCAL-5^i

(*fully sup.* / *weakly sup.*)		1-shot						5-shot	
Method	Backbone	Split-0	Split-1	Split-2	Split-3	Mean	Δ	Mean	Δ
1-way									
PANet [46]	VGG-16	20.9† / 12.7	20.9† / 8.7	20.9† / 5.9	20.9† / 4.8	20.9 / 8.0	12.9	29.7 / 13.9	15.8
Ours	VGG-16	26.0 / 24.2	14.5 / 12.9	20.0 / 17.0	18.3 / 14.0	19.7 / **17.0**	**2.7**	27.0 / **17.5**	**9.5**
2-way									
Ours	VGG-16	18.2 / 17.4	12.2 / 9.5	9.1 / 10.4	6.5 / 7.1	11.5 / **11.1**	**0.4**	14.8 / **11.9**	**2.9**

(b) MS COCO

Table 2: Performance evaluation on (a) PASCAL-5^i and (b) MS COCO in terms of mean-IoU (Mean) and performance difference Δ due to change of settings. The numbers before and after '/' indicate results under fully and weakly supervised settings, respectively. Note that [37] considers a loosely weakly supervised setting and requires ground truth pixel-level masks during training, while [42] utilizes a stronger backbone (ResNet-50) compared to others (VGG-16).

4. Experiments

Datasets. We follow the evaluation protocol in [36] and conduct experiments on the PASCAL-5^i dataset. It contains a total of 20 object categories from the PASCAL VOC 2012 [10] and the extended SDS [15] datasets, which are evenly divided into 4 splits ($i = 0, 1, 2, 3$). Additionally, we consider the MS COCO 2014 [21] dataset, which contains 80 object categories and thus is more challenging. Following the settings of [46], we divide MS COCO into 4 splits with 20 categories each. For both datasets, three of the splits are used as base (training) classes, with the remaining one as novel (testing) classes in each experiment.

Evaluation Metrics. We follow [36, 57, 46, 54, 27] and apply the mean intersection over union (mean-IoU) as the evaluation metric, which computes separate IoUs for each foreground category, and then averages them along with the background class. Following the protocol in [46], the mean-IoU in each evaluation is calculated by the average of 5 runs, with each run randomly sampling 1,000 episodes from the testing set.

Implementation Details. In our experiments, all images are normalized and reshaped to 129×129 pixels. For pseudo pixel-level label generation, we choose the VGG-16 [39] network as the CAM backbone, with weights pre-trained over a reduced subset of ILSVRC 2012 [33] with categories in $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ removed. For semantic labels, we use the word embedding vectors pre-trained on Wikipedia using fastText [3]. We apply DSS [16] to extract saliency maps for the gating mechanism in Fig. 3, which is a model pre-trained over MSRA-B [22] without categorical supervision.

For the segmentation model, we use a DeepLabv3+ [8] pre-trained over irrelevant categories, which remains fixed throughout the meta-learning process. The encoder E is a multilayer perceptron with two hidden layers, and the output dimension (for the latent space) is set to 64. All experiments are implemented by PyTorch, and are run using a single NVIDIA Titan RTX graphics card with 24GB of video memory.

4.1. Comparison with State-of-the-art Methods

Since no previous work was designed to address this weakly supervised few-shot segmentation task, we choose to implement modified versions of state-of-the-art supervised methods by replacing the ground truth masks for training with our generated pseudo labels (i.e., as introduced in Section 3.2). As depicted in Figure 4, our proposed framework can be trained in a fully supervised fashion (i.e., using ground truth pixel-level masks during training). Thus, we include this fully supervised version of our model as the performance upper bounds. We compare our results mainly with methods using the same backbone [37, 38, 46], with

† Split-wise results not reported in the original paper.

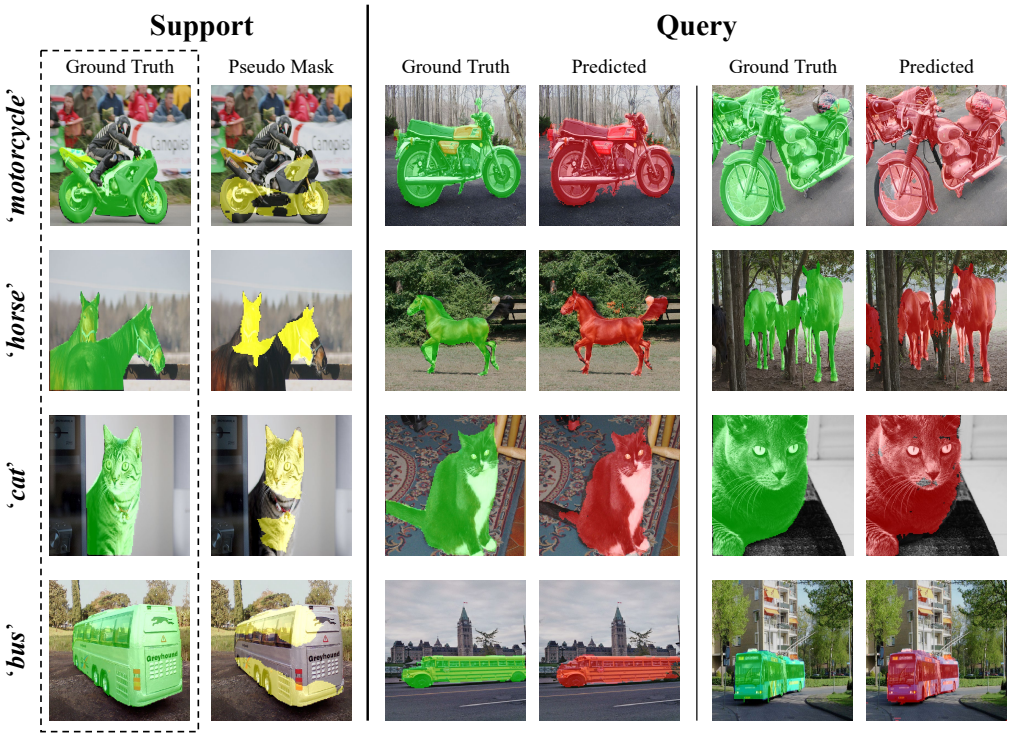

Figure 5: Example 1-way 1-shot segmentation results on PASCAL-5^i. We have green, yellow and red pixels represent ground truth pixel-level foreground labels, generated pseudo masks, and the predicted outputs, respectively. Note that our weakly supervised setting does not observe ground truth pixel-level labels for the support set images (in dotted frames).

the exception of [42] for cross-backbone analysis.

PASCAL-5^i. We first compare the performances of different methods on PASCAL-5^i, with results listed in Table 2a. In the first row of this table, we consider a loosely weakly supervised model of [37] (as illustrated in Figure 2b). For other methods (including ours) in this table, we present results under both fully and weakly supervised settings. While our fully supervised model achieved comparable performance with state-of-the-art methods in the standard 1-way 1-shot setting, our model reported a significant improvement over PANet [46] by 15.3% (42.4% v.s. 27.1%) in the weakly supervised setting. It is worth noting that PFENet [42] is trained using a stronger backbone (ResNet-50), while all other methods (including ours) utilize VGG-16. Nevertheless, our model still outperforms [42] by a considerable margin of 2.5% in the weakly supervised setting. We also observe that the performance drop between the two different settings of our model is significantly less than the others. That is, when only image-level labels (instead of ground truth pixel-level masks) are observed during both training and testing, both PFENet [42] and PANet [46] suffered from a >20% performance drop while only 5.1% was reported by our model.

MS COCO. As shown in Table 2b, despite the increased difficulty in few-shot segmentation on MS COCO, our model is able to achieve satisfactory performances under both fully and weakly supervised settings when comparing to [46]. Specifically, we only observe a 2.7% performance drop between the two settings on the 1-way 1-shot task, while that of [46] is 12.9%. The above quantitative results support the use of our propose framework for solving few-shot semantic segmentation, especially when only image-level labels can be observed during both training and testing (i.e., the weakly supervised setting).

Multi-way Segmentation. We now show that our model is applicable to the cases when there is more than one foreground object category in an image, which is more challenging since it requires more information to be learned in each episode. In the lower parts of Table 2, we list the performances of different methods under the 2-way setting (i.e., two types of foreground objects exist in an image). It is worth noting that, while most existing methods [36, 38, 54, 27, 17, 42] are designed to tackle only 1-way segmentation, they typically claim such extension can be realized by forward passing K times with additional decision rules. On the contrary, our method can directly produce output labels of a multi-category image as the final classification by a k-NN search. As shown in Table 2a, our model performed favorably against previous methods by a margin of 11.5% on the PASCAL-5^i 2-way 1-shot task. To

Figure 6: Example failure cases for our weakly supervised model and the fully supervised PFENet [42].

the best of our knowledge, we are the first to report results of 2-way tasks for the MS COCO dataset, as shown in the last row of Table 2b.

4.2. Analysis of Our Proposed Method

Ablation Study. As our meta-learning process is mainly achieved by the episodic learning of encoder E, we now design a baseline version such that it directly predicts the output mask by a k-NN search on the pixel features encoded by the DeepLabv3+ backbone (i.e., without embedding into the latent space via E). As shown in the first row of Table 3, the results were severely degraded with a drop of up to 18%, if the model was not learned via the meta-learning objectives. This further confirms that, the promising performance achieved by our proposed method is not a direct result of using particular strong backbones. Instead, it leverages spatial and structural details captured by the backbone, upon which semantic and category-wise information is reinforced via the meta-learning process.

Additionally, we provide results that are trained using pseudo masks generated without saliency gating (as mentioned in the last step of Section 3.2). As shown in the second row of Table 3, a slight decrease in mean-IoU (less than 5%) was observed, while still outperforming the baseline version by a large margin. Thus, the use of saliency gating as the post-processing step for pseudo pixel-level masks would be preferable but not critical.

Qualitative Analysis. For visual comparisons, we present qualitative results of 1-way 1-shot segmentation on PASCAL-5^i dataset in Figure 5. As detailed in Section 3.1, our model is realized in the weakly supervised setting and does not observe ground truth masks of support images during training (i.e., column 1 in Figure 5). Instead, given each image and its image-level label (i.e., class name), we first generate its pseudo mask, which serves as the guidance for our meta-learning framework. As evident in column 2 of Figure 5, our generated pseudo masks do not cover the entire foreground object, but instead contain only discriminative regions that are informative in classifying image-level labels (e.g., muzzle of a horse, or wheels and pedals of a motorcycle). Nevertheless, from this figure, we see that our proposed model is able to predict masks for query images with satisfactory performances (e.g., columns 4 and 6 in Figure 5). It is also worth noting that no post processing is performed on our predicted results.

Method	1-shot		5-shot	
	1-way	2-way	1-way	2-way
Ours w/o E	30.0	19.7	31.5	25.4
Ours w/o Saliency	41.0	35.4	42.1	39.4
Full Version	42.4	37.7	45.5	43.0

Table 3: Ablation study of our model on PASCAL-5^i in mean-IoUs. Ours w/o E denotes our framework without meta-learner encoder E, while w/o Saliency indicates our model without applying saliency gating to process the generated pseudo masks.

From the split-wise mean-IoUs listed in Table 2, we see that some data splits would generally suffer from performance drops across different models including ours. In Figure 6, we show failure segmentation example results by our weakly supervised model and the fully supervised PFENet [42]. As evident in this figure, the ground truth images of such categories generally contain small foreground regions (e.g., bottle), or those that cannot be easily distinguished from the background (e.g., chair). For such image categories, training with few-shot samples would *not* be expected to address semantic segmentation tasks well. This would be the limitation of developing solutions for few-shot semantic segmentation.

5. Conclusion

In this paper, we proposed a unique learning framework for few-shot semantic segmentation in a weakly supervised manner, which only observes image-level labels during both training and testing. Under such weak supervision, our proposed model serves as a pixel-level meta-learner, which produces pseudo segmentation masks for guiding pixel-wise classification in a meta-learning fashion. Our experiments confirmed the effectiveness of our model, which achieved satisfactory results on two benchmark datasets under fully supervised settings, while surpassing the state-of-the-art methods in weakly supervised settings. Finally, we point out the challenge and limitation of the task of few-shot semantic segmentation.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 110-2634-F-002-036 and 110-2221-E-002-121.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1, 2

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. 2

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 6

[4] Maxime Bucher, VU Tuan-Hung, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 466–477, 2019. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1, 2, 3

[6] 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. 1, 2

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 4, 6

[9] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. 2, 5

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 2

[12] Michael Fink. Object classification from a single example utilizing class relevance metrics. In *Advances in neural information processing systems*, pages 449–456, 2005. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 1, 2

[14] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. *arXiv preprint arXiv:2004.15014*, 2020. 2

[15] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 6

[16] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 6

[17] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8441–8448, 2019. 7

[18] 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. 4

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021. 1, 2

[21] 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. Springer, 2014. 6

[22] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010. 6

[23] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4173, 2020. 2

[24] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. *arXiv preprint arXiv:2007.06309*, 2020. 2

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[27] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 622–631, 2019. 1, 2, 3, 6, 7

[28] Niloufar Pourian, Sreejith Karthikeyan, and Bangalore S Manjunath. Weakly supervised graph based semantic segmentation by learning communities of image-parts. In *Proceedings of the IEEE international conference on computer vision*, pages 1359–1367, 2015. 2

[29] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation, 2018. 2

[30] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017. 1, 2

[31] Hasnain Raza, Mahdyar Ravanbakhsh, Tassilo Klein, and Moin Nabi. Weakly supervised one shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4, 6

[34] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 1, 2

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[36] 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*. BMVA Press, 2017. 2, 6, 7

[37] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic inputs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 2, 3, 6, 7

[38] Mennatullah Siam, Boris N. Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6, 7

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[40] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2, 5

[41] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2

[42] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Annals of the History of Computing*, (01):1–1, 2020. 1, 2, 3, 5, 6, 7, 8

[43] Alexander Vezhnevets and Joachim M Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3249–3256. IEEE, 2010. 2

[44] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann. Weakly supervised structured output learning for semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 845–852. IEEE, 2012. 2

[45] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. ECCV, 2020. 2

[46] 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*, pages 9197–9206, 2019. 1, 2, 3, 4, 6, 7

[47] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*, 2019. 2

[48] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021. 2, 3

[49] Wenxuan Xie, Yuxin Peng, and Jianguo Xiao. Weakly-supervised image parsing via constructing semantic graphs and hypergraphs. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 277–286, 2014. 2

[50] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Tell me what you see and i will show you where it is. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3190–3197, 2014. 2

[51] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3781–3790, 2015. 2

[52] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. *arXiv preprint arXiv:2008.03898*, 2020. 2

[53] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[54] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 2, 3, 4, 6, 7

[55] Luming Zhang, Mingli Song, Zicheng Liu, Xiao Liu, Jiajun Bu, and Chun Chen. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1908–1915, 2013. 2

[56] Wei Zhang, Sheng Zeng, Dequan Wang, and Xiangyang Xue. Weakly supervised semantic segmentation for social images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2015. 2

[57] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *arXiv preprint arXiv:1810.09091*, 2018. 6

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[59] 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, 2016. 2, 4