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

(Supplementary Material)

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

A. 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 [8] network as the CAM backbone, with weights pre-trained over a reduced subset of ILSVRC 2012 [5] with categories in $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ removed (the number of training classes $N_{\text{CAM}} = 894$). For semantic labels, we use the word embedding vectors pre-trained on Wikipedia using fastText [1], which has a dimension of 300.

The decision rule from $\overline{T}$ to $\tilde{M}$ is defined as

$$\tilde{M}_{(x,y)} = \begin{cases} c, & \text{if } \overline{T}_{(x,y)} \geq \tau_1 \text{ and } \Phi_{(x,y)} \geq \tau_2 \\ 0, & \text{if } \overline{T}_{(x,y)} < \tau_1 \text{ and } \Phi_{(x,y)} < \tau_2 \\ \text{`}\textit{uncertain}\text{'}, & \text{otherwise} \end{cases},$$

where $\Phi \in [0,1]^{H \times W}$ is the saliency map generated by DSS [3], which is a saliency detection model pre-trained over MSRA-B [4] with only foreground/background information observed (without categorical supervision). The thresholds are selected via cross-validation (we have $\tau_1 = 0.2$ and $\tau_2 = 0.8$). For images containing multiple categories, pixels identified with more than one foreground labels are also assigned '*uncertain*'. Both the CAM heatmap outputs and the saliency maps are $1/4$ the size of the input image, while the generated pseudo labels are upsampled back to the original size.

For the segmentation model, we use a DeepLabv3+ [2] pre-trained over irrelevant categories with ResNet-101 as backbone, which remains fixed throughout the meta-learning process. The final classification layer is removed so that it outputs downscaled 304-dimensional pixel-wise feature maps of $1/2$ the input size (i.e. $d = 304$). Note that the descriptions in Section 3.3 of our main paper assumed that the feature maps have the same dimension as the input images for the sake of brevity. The encoder E is a multilayer perceptron with two hidden layers, and the output dimension (for the latent space) is set to 64. The model is trained by iterating a total of 30,000 episodes with an initial learning rate of 10^{-3}, which decreases by a factor of 0.5 every 5,000 iterations.

All experiments are implemented by PyTorch, and are run using a single NVIDIA Titan RTX graphics card with 24GB of video memory.

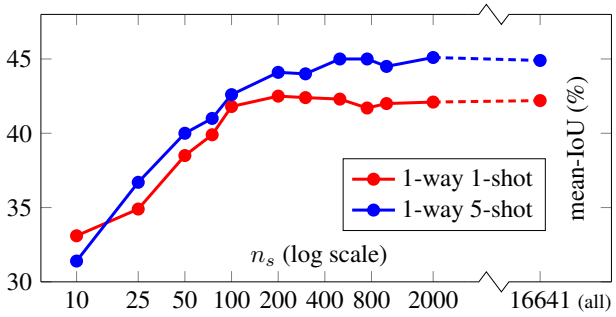

Figure 1: Performances of support pixel sampling with varying n_s on PASCAL-5^i. Note that n_s denotes the number of support pixels sampled from each category for semantic segmentation.

B. Additional Quantitative Results

B.1. Effects on Support Pixel Sampling

As introduced in Section 3.3 of the main paper, our meta-learning model produces pseudo pixel-level labels for the support images. Instead of taking all the masked pixels for learning purposes, we randomly sample a fixed number (n_s) of features from each category and form the associated support pixel features, which guide the segmentation of query inputs. We now study the effect of the sampled feature numbers n_s. As shown in Figure 1, despite the performance generally increased with larger n_s, such improvements became marginal with n_s above 200. Thus, we set n_s as 200

Method	5-shot					
	Split-0	Split-1	Split-2	Split-3	Mean	Δ
1-way						
Co-att [6]	45.9	65.7	48.6	46.6	51.7	—
AMP [7]	41.8 / 13.6	55.5 / 18.0	50.3 / 11.0	39.9 / 16.2	46.9 / 14.7	32.2
PANet [10]	51.8 / 38.7	64.6 / 44.4	59.8 / 39.5	46.5 / 28.2	55.7 / 37.7	18.0
PFENet [9]	63.1 / 51.4	70.7 / 51.8	55.8 / 43.3	57.9 / 32.5	61.9 / 44.8	17.1
Ours	41.6 / 40.2	61.1 / 55.4	56.7 / 49.4	42.9 / 37.3	50.6 / **45.5**	**5.1**
2-way						
PANet [10]	53.1† / 35.3	53.1† / 44.9	53.1† / 37.8	53.1† / 28.2	53.1 / 36.6	16.5
Ours	41.4 / 37.4	57.9 / 52.6	55.1 / 45.6	42.9 / 36.6	49.3 / **43.0**	**6.3**

Table 1: Split-wise performance of 5-shot tasks evaluated on PASCAL-5^i in terms of mean-IoU.

Method	5-shot					
	Split-0	Split-1	Split-2	Split-3	Mean	Δ
1-way						
PANet [10]	29.7† / 25.1	29.7† / 12.0	29.7† / 9.1	29.7† / 9.3	29.7 / 13.9	15.8
Ours	31.0 / 24.2	23.1 / 13.4	30.5 / 21.1	23.4 / 11.2	27.0 / **17.5**	**9.5**
2-way						
Ours	22.8 / 21.2	16.2 / 10.0	11.5 / 10.5	8.9 / 6.0	14.8 / **11.9**	**2.9**

Table 2: Split-wise performance of 5-shot tasks evaluated on MS COCO in terms of mean-IoU.

$\mathcal{C}_{\text{pre}}$ \ $\mathcal{C}_{\text{novel}}$	**1-way 1-shot**				**1-way 5-shot**			
	Split-0	Split-1	Split-2	Split-3	Split-0	Split-1	Split-2	Split-3
Split-0	—	47.5	42.4	31.1	—	54.6	47.2	34.0
Split-1	36.5	—	45.9	35.6	40.2	—	49.4	37.3
Split-2	34.8	51.5	—	34.7	38.1	55.0	—	31.6
Split-3	35.6	51.7	42.5	—	38.5	55.4	45.8	—
Max	36.5	51.7	45.9	35.6	40.2	55.4	49.4	37.3
μ	35.6	50.2	43.6	33.8	38.9	55.0	47.4	34.3
σ	**0.84**	**2.39**	**1.99**	**2.40**	**1.13**	**0.41**	**1.79**	**2.86**

Table 3: Performance evaluation with different pre-training categories for DeepLabv3+ on PASCAL-5^i in terms of mean-IoU under the proposed *weakly supervised* setting. Note that $\mathcal{C}_{\text{pre}}$ and $\mathcal{C}_{\text{novel}}$ denote the pre-training split for DeepLabv3+ and the testing split, respectively.

$\mathcal{C}_{\text{pre}}$ \ $\mathcal{C}_{\text{novel}}$	**1-way 1-shot**				**1-way 5-shot**			
	Split-0	Split-1	Split-2	Split-3	Split-0	Split-1	Split-2	Split-3
Split-0	—	53.7	49.3	35.7	—	58.5	55.5	39.2
Split-1	38.3	—	54.0	40.1	41.6	—	56.7	42.9
Split-2	38.2	55.4	—	33.8	41.3	58.6	—	36.6
Split-3	37.1	57.6	50.0	—	40.6	61.1	53.4	—
Max	38.3	57.6	54.0	40.1	41.6	61.1	56.7	42.9
μ	37.9	55.6	51.1	36.6	41.1	59.4	55.2	39.6
σ	**0.68**	**1.93**	**2.55**	**3.20**	**0.52**	**1.48**	**1.70**	**3.19**

Table 4: Performance evaluation with different pre-training categories for DeepLabv3+ on PASCAL-5^i in terms of mean-IoU under the *fully supervised* setting. Note that $\mathcal{C}_{\text{pre}}$ and $\mathcal{C}_{\text{novel}}$ denote the pre-training split for DeepLabv3+ and the testing split, respectively.

in our work. It is worth pointing out that this is evident on both 1 and 5-shot tasks on PASCAL-5^i. This indicates that our model is able to exploit representative information from the pseudo masks of few images, guiding the segmentation of query images in the weakly supervised settings. We note that in the case where the number of total support pixels from a category is less than n_s, we would use all support pixels from that category.

B.2. Split-wise Evaluation

In our main paper, we presented quantitative results evaluated on two datasets (PASCAL-5^i and MS COCO) in terms of the average of mean-IoUs for 4 different splits. We now provide split-wise mean-IoUs for 5-shot tasks (which were not included in the main paper due to space limits), as also reported in previous works in the few-shot semantic segmentation literature. The results for PASCAL-5^i are shown in Table 1, while those for MS COCO are shown in Table 2. The numbers before and after '/' indicate results under fully and weakly supervised settings, respectively. Δ denotes the performance drop between the two settings. Note that [6] considers a loosely weakly supervised setting and requires ground truth pixel-level masks during training, while [9] utilizes a stronger backbone (ResNet-50) compared to others (VGG-16).

B.3. Pre-training Categories for DeepLabv3+

As mentioned in our main paper, our proposed weakly supervised setting remains valid as long as the DeepLabv3+ backbone in our network is pre-trained using irrelevant categories (i.e., classes that are neither present in $\mathcal{C}_{\text{base}}$ nor $\mathcal{C}_{\text{novel}}$). We have already shown in the ablation study that the performance achieved by our framework is not a direct result of a strong backbone. Here, we provide more insight into the effect of the backbone by minimizing the number of pre-training categories used by our DeepLabv3+. Specifically, in each experiment of the PASCAL-5^i dataset, we use one out of the three training splits as the pre-training categories for DeepLabv3+ (denoted as $\mathcal{C}_{\text{pre}}$), while the remaining two are utilized as $\mathcal{C}_{\text{base}}$. This means that for each testing split, we are able to construct three different experimental setups and thus obtain three separate mean-IoU results.

In Tables 3 and 4, we report the mean-IoU of each testing split using different pre-training splits for DeepLabv3+ under weakly and fully supervised settings, respectively. As evident in the results, the performance does not exhibit drastic fluctuations despite using different sets of $\mathcal{C}_{\text{pre}}$. We specifically point this out by highlighting the column-wise standard deviations (last row of Table 3 and 4), which are reasonably low and negligible compared to the performance gap between our proposed method and previous works, as reported in Tables 1 and 2. This further shows that our model is robust against backbones pre-trained with different or even a minimal number of categories ($|\mathcal{C}_{\text{pre}}| = 5$ in our case).

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

References

[1] 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. 1

[2] 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

[3] 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. 1

[4] 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. 1

[5] 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. 1

[6] 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. *arXiv preprint arXiv:2001.09540*, 2020. 2, 3

[7] 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. 2

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] 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. 2, 3

[10] 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. 2