2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222652

SEMI-SUPERVISED FEW-SHOT SEGMENTATION WITH NOISY SUPPORT IMAGES

Runtong Zhang[1], *Hongyuan Zhu*[2], *Hanwang Zhang*[3], *Chen Gong*[4], *Joey Tianyi Zhou*[5], *Fanman Meng*[1]*

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3] Nanyang Technological University, Singapore
[4] Nanjing University of Science and Technology, Nanjing, China
[5] Centre for Frontier AI Research (CFAR), A*STAR, Singapore

ABSTRACT

Motivated by the semi-supervised learning that uses the unlabeled data and pseudo annotations to improve the image classification, this paper proposes a new semi-supervised few-shot segmentation (FSS) framework of which the training process uses not only the annotated images, but also the unlabeled images, *e.g.* images from other available datasets, to enhance the training of the FSS model. Furthermore, in the test phase, more support images and pseudo-annotations can also be generated by the proposed framework to enrich the support set of novel classes and therefore benefit the inference. However, unlabeled images are not a free lunch. The noisy intra-class samples and inter-class samples existed in the unlabeled images as well as the interferences of the bad quality of pseudo annotations make it difficult to utilize the correct images and pseudo annotations for a certain class. To this end, we further propose a ranking algorithm consisting of an inter-class confidence term and an intra-class confidence term to efficiently utilize the pseudo annotations of the class with high quality. Extensive experiments on COCO-20^i dataset demonstrate that the proposed semi-supervised FSS framework is superior to many state-of-the-art methods.

***Index Terms*—** few-shot segmentation, semi-supervised learning, noisy images

1. INTRODUCTION

Few-shot segmentation (FSS) [1] aims to segment object regions of a new class based on a small number (N-shot) of annotated support samples. Existing few-shot segmentation methods can be categorised into prototype-based approach [2, 3, 4, 5, 6] and metric-based approach [7, 8]. Prototype-based approaches focus on generating representative prototypes from a small number of annotations that can represent the new class well. Metric-based approaches [7, 8] focus on learning a robust class-agnostic similarity metric that can successfully find the common regions with large variations among support and query images pair. However, the segmentation improvement is very limited due to the fact that learning powerful representative prototypes or class-agnostic similarity metric from a limited annotation set is also hard work.

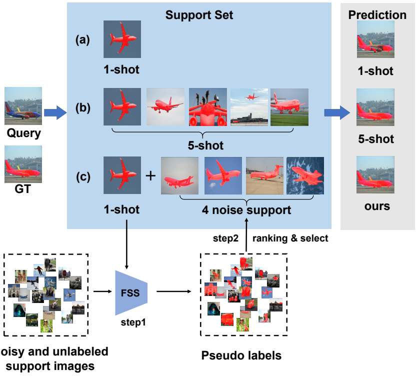

Fig. 1. (a) 1-shot setting. (b) 5-shot setting. (c) 1-shot with additional 4 noise support images with pseudo labels. Using 1-shot and 4 noise support can achieve comparable performance to 5-shot without manual annotations.

Besides, recent works [9, 10, 11, 12] demonstrate that semi-supervised learning can improve the classification via generating pseudo labels of unlabeled images. For example, the method in [9] generates and selects similar pseudo labels from unlabeled data to exploit the consistency constraint and thus increases model's generalization in classification. The method in [10] trains a teacher model to generate good pseudo labels to augment the training dataset and thus benefit the student model for object detection. The method in [11] utilizes an unlabeled set to generate and select good pseudo labels based on similar loss distribution to enhance the classification model. These semi-supervised methods provide a new solution for the limited annotation set of FSS.

In this paper, we propose a semi-supervised FSS framework utilizing the support set with a mix of annotated images and other unlabeled noisy images to enrich the annotation set. A brief pipeline is shown in Fig. 1 (c). Given the noisy and unlabeled support images, we firstly generate pseudo labels

*Corresponding author. Email: fmmeng@uestc.edu.cn

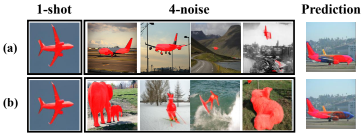

Fig. 2. Examples of two problems. (a) Noisy intra-class samples as support samples. (b) Noisy inter-class samples as support samples.

using a pretrained model. Then, the pseudo masks with high confidence are used as ground truth to expand the support set. Afterwards, the expanded support set is utilized to enhance the few-shot segmentation model in both training and test phases.

However, unlabeled noisy images are not a free lunch. There are two problems that complicate pseudo-label selection (as shown in Fig. 2). 1) **Noisy Intra-Class Samples:** The noisy intra-class samples contain ambiguous objects that may strengthen the background and weaken the foreground, *e.g.*, noisy "background" dominates the image as shown in Fig. 2 (a). 2) **Noisy Inter-Class Samples:** The noisy inter-class samples introduce irrelevant features to the task, which may cause feature bias and thus confuse the FSS model, *e.g.*, the FSS model is confused by "elephant", "person" and "sheep" when segmenting "aeroplane" as shown in Fig. 2 (b).

To solve the above two problems, we propose a ranking algorithm to automatically eliminate the noisy intra-class samples and inter-class samples. Specifically, the proposed ranking algorithm consists of two terms: an intra-class confidence term R and an inter-class confidence term T based on the two types of noisy samples. The term R is calculated by two sub-terms: E_{sc} and E_{imc}, where E_{sc} measures prediction uncertainty based on binary entropy of each pixel, and E_{imc} identifies different types of errors based on the co-teaching framework [13, 14]. The term T is calculated by measuring the feature similarities between the support prototypes and the noisy unlabeled images with pseudo labels. Finally, a ranking score E is obtained based on R and T, and the top scored pseudo labels are selected as new annotations. The proposed semi-supervised FSS framework is validated on COCO-20^i and is superior to many state-of-the-art methods in Sect. 3.

2. METHOD

2.1. Semi-Supervised FSS Framework

Fig. 3 (a) shows the proposed semi-supervised FSS framework. The key innovation is in phase II, where a ranking algorithm is proposed to evaluate the pseudo labels. Specifically, an intra-class confidence term R and an inter-class confidence term T are calculated for each pseudo label. Then, a final ranking score E is obtained by simply calculating the weighted sum of R and T:

$$E = \alpha \cdot R + \beta \cdot T \tag{1}$$

where α and β are weighting coefficients. Afterwards, the top k scored pseudo labels are selected to form a new annotation set:

$$\mathcal{S}^{base}_{new} \leftarrow \mathcal{S}^{base} + \{(X_1, \hat{Y}_{X_1}), (X_2, \hat{Y}_{X_2}), ..., (X_k, \hat{Y}_{X_k})\} \tag{2}$$

where $\mathcal{S}^{base}$ indicates the initial annotation set of base classes in the training phase, $\hat{Y}_X$ indicates the pseudo label of image X.

Finally, in phase III, the new annotation set $\mathcal{S}^{base}_{new}$ is used to retrain N_θ and get better predictions. More details of the intra-class confidence term R and the inter-class confidence term T are introduced in Sect. 2.1.1 and Sect. 2.1.2, respectively.

2.1.1. *Intra-Class Confidence Term R*

The term R aims to estimate the credibility of intra-class pseudo labels, which is calculated by:

$$R = E_{sc} \times E_{imc} \tag{3}$$

where E_{sc} estimates the prediction uncertainty of pseudo labels and E_{imc} identifies different types of errors in pseudo labels.

Segmentation Confidence Term E_{sc}. This term is calculated by adopting a binary-entropy-based function to measure the prediction uncertainty:

$$E_{sc} = -\frac{1}{N}\sum_{i} H(i) + B \tag{4}$$

where i indicates a pixel position, $H(\cdot)$ is the binary entropy function, N is the total number of pixels, and B is a bias term to ensure $E_{sc} \in [0, 1]$. The formulation of $H(x)$ is shown in Eq. 5, where $p(i)$ is the logit at pixel position i.

$$H(x) = -p(i)log(p(i)) - (1 - p(i))log(1 - p(i)) \tag{5}$$

Instance Mask Consensus Term E_{imc}. This term is motivated by the co-teaching framework [13, 14], which proves that two diverged networks can filter different types of errors. Therefore, if two diverged FSS networks output similar predictions to the same unlabeled image, the predictions contain little errors and have high confidence.

The pipeline of getting E_{imc} is shown in Fig. 4. Specifically, the unlabeled image X is processed by two FSS networks $N_{\theta 1}$ and $N_{\theta 2}$ with a given support sample $\{S, Y_S\}$, where S is the support image and Y_S is the manual annotation. Then, a metric $m(\cdot, \cdot)$ is calculated between the two output $\hat{Y}^1_X$ and $\hat{Y}^2_X$. Thus, the calculation of E_{imc} is:

$$E_{imc} = m(\hat{Y}^1_X, \hat{Y}^2_X) \tag{6}$$

where $\hat{Y}^1_X$ and $\hat{Y}^2_X$ are predictions of the same unlabeled image X from two diverged networks $N_{\theta 1}$ and $N_{\theta 2}$. $m(\cdot, \cdot)$ indicates a segmentation metric score, *e.g.*, mIoU.

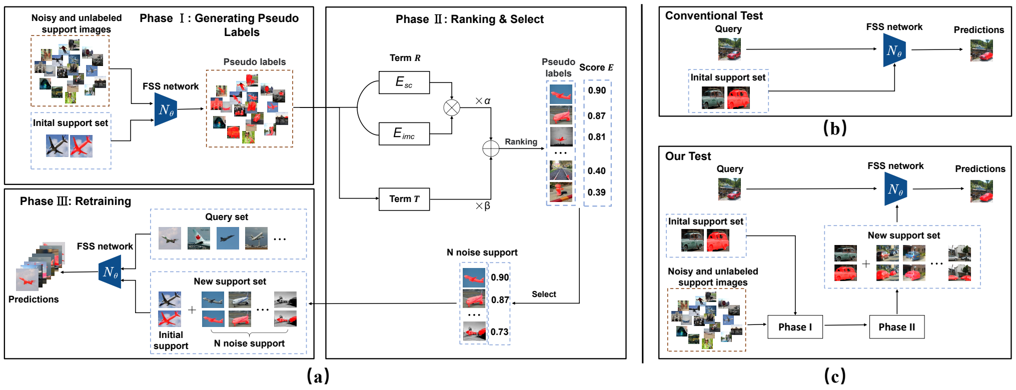

Fig. 3. (a) The proposed semi-supervised FSS framework consisting of three phases. In phase I, a pretrained FSS network N_θ is used to obtain the pseudo labels. In phase II, a ranking algorithm is proposed to calculate quality scores E of pseudo labels. In phase III, top scored pseudo labels are selected as new support samples to retrain N_θ. (b) Conventional FSS test. N_θ is tested on novel classes with an annotated initial support set. (c) Our FSS test based on the proposed semi-supervised framework. N_θ is tested on novel classes with a new support set, which is expanded following phase I and phase II.

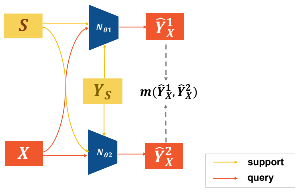

Fig. 4. The pipeline of E_{imc}. The unlabeled image X is processed by two FSS networks $N_{\theta 1}$ and $N_{\theta 2}$ with a given support sample $\{S, Y_S\}$. Then, a metric $m(\cdot,\cdot)$ is calculated between the two output $\hat{Y}_X^1$ and $\hat{Y}_X^2$.

2.1.2. *Inter-Class Confidence Term T*

The term T aims to identify the noisy inter-class samples based on the feature similarities between the support prototypes and the pseudo labels. First, the prototype of class c of the initial support set $\mathcal{S}^c = \{S_1^c, S_2^c, ..., S_n^c\}$ are calculated by:

$$\mathcal{P}^c = \frac{1}{n}\sum_{i=1}^{n} \sigma(\mathcal{F}_{S_i^c}, Y_{S_i^c}) \tag{7}$$

where $\mathcal{F}_{S_i^c} \in \mathbb{R}^{C\times H\times W}$ is the feature map of support S_i^c of class c, $Y_{S_i^c}$ is the manual annotation, $\sigma(\cdot)$ is the masked global average pooling, and $\mathcal{P}^c \in \mathbb{R}^C$ is the prototype of class c. Then, the term T can be calculated by:

$$T = s(\mathcal{P}^c, \sigma(\mathcal{F}_X, \hat{Y}_X)) \tag{8}$$

where $\mathcal{F}_X \in \mathbb{R}^{C\times H\times X}$ is the feature map of X, $\hat{Y}_X$ is the generated pseudo label, $s(\cdot,\cdot)$ is a similarity metric, *e.g.*, cosine similarity.

2.2. Our FSS Test with Semi-Supervised Framework

Based on the proposed semi-supervised FSS framework, we can further expand the initial support set of novel classes simply in the test phase to improve the segmentation performance, of which the pipeline is shown in Fig. 3 (c). Specifically, different from the conventional FSS test (Fig. 3 (b)) where only a small annotated support set $\mathcal{S}^{novel}$ of novel classes is utilized, our test enriches $\mathcal{S}^{novel}$ following the pipeline of phase I and phase II in the semi-supervised framework to obtain the new support set $\mathcal{S}^{novel}_{new}$:

$$\mathcal{S}^{novel}_{new} \leftarrow \mathcal{S}^{novel} + \{(X_1, \hat{Y}_{X_1}), (X_2, \hat{Y}_{X_2}), ..., (X_k, \hat{Y}_{X_k})\} \tag{9}$$

Then, the query images will be segmented with the new support set $\mathcal{S}^{novel}_{new}$ to get better predictions.

3. EXPERIMENT

3.1. Dataset, Metrics and Training Details

We train and validate our method on COCO-20^i [1] following the dataset settings of existing FSS methods [15, 8, 16, 2]. Besides, we use 123,403 unlabeled images in COCO2017 [17] as the noisy unlabeled image dataset. The mean intersection over union (mIoU) and foreground-background IoU (FB-IoU) are adopted as the evaluation metrics.

Our experiments are conducted on two NVIDIA Titan XP GPUs and Intel Core i9-9900k CPU @ 3.60GHz× 16. Our code is constructed on PyTorch. We adopt HSNet [16] as the base model N_θ in Fig. 3. In Sect. 2.2, the two diverged networks $N_{\theta 1}$, $N_{\theta 2}$ of E_{imc} are set to the base model with different backbones: ResNet50 and ResNet101. The publicly released pretrained models of HSNet are used directly in phase I and phase II. We set $m(\cdot,\cdot)$ to mIoU score in Sect. 2.1.1 and set $s(\cdot,\cdot)$ to cosine similarity in Sect. 2.1.2. The feature maps $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$ in Sect. 2.1.2 are extracted from the

Table 1. IoU and FB-IoU scores on COCO-20^i. "†" is the results of the conventional test. "‡" is the results of our test based on the proposed semi supervised framework. "Oracle" is the 5-shot performance of base model. "±0.1" is the standard deviation of repeating 5 times.

Backbone	Method	1-shot	
		mean	FB-IoU
ResNet50	HSNet [16]	39.2	68.2
	CyCTR [18]	40.3	-
	VAT [19]	41.3	68.8
	ASNet [20]	42.2	68.8
	Ours (HSNet) †	40.5 ($\pm$0.3)	69.0 ($\pm$0.3)
	Ours (HSNet) ‡	**49.5** ($\pm$0.3)	**71.8** ($\pm$0.4)
	Oracle	46.9	70.7
ResNet101	MLC [6]	36.4	-
	PFENet [8]	38.5	63.0
	HSNet [16]	41.2	69.1
	ASNet [20]	43.1	69.4
	Ours (HSNet) †	42.7 ($\pm$0.3)	69.7 ($\pm$0.2)
	Ours (HSNet) ‡	**51.0** ($\pm$0.5)	**72.5** ($\pm$0.6)
	Oracle	49.5	72.4

Table 2. Computational complexity of our with the baseline.

Method	Learnable Params	FPS	FLOPs (G)
HSNet	2.6 M	16.33	20.56
Ours (HSNet)	2.6 M	16.45	20.52

last convolutional layer of the backbone. The coefficients α and β in Eq. 1 are set to 0.3 and 0.7, respectively. In the training phase, pseudo labels with $E \geq 0.65$ are selected as new annotations of base classes. In the test phase, the top 4 scored pseudo labels are introduced into the support set of novel classes. In phase III, the retraining setting strictly follows the base model [16].

3.2. Results

We evaluate the proposed method on COCO-20^i dataset and compare it with existing FSS methods [8, 6, 18, 19, 20, 16] in Table 1. "Ours (HSNet) †" achieves mIoU improvements of 1.3% and 1.5% on ResNet50 and ResNet101, respectively, compared with HSNet (*baseline*). Besides, "Ours (HSNet) ‡" achieves larger mIoU improvements of 10.3% and 9.8% on ResNet50 and ResNet101, respectively. Meanwhile, the proposed method surprisingly surpasses the "Oracle" in some cases. This can be contributed to the introduced unlabeled images, which enriches the support image set of novel classes and thus enhances the inference of the FSS model. Most notably, "Ours (HSNet) ‡" on ResNet101 obtains 1.5% mIoU gains and 0.1% FB-IoU gains compared with the "oracle", which is a new state-of-the-art to our best knowledge.

In addition, we have compared the learnable params, FPS, and FLOPs of our method with the baseline in Table 2. One can observe that our method does not introduce additional learnable params and our FPS and FLOPs are also close to the baseline. The reason is that the pseudo label generation in phase I and ranking process in phase II are completed once before the training phase and does not affect the computational complexity of both the training and test phases.

Table 3. Ablation study of the proposed method with different design choices. "±0.1" is the standard deviation of repeating 5 times.

R		T	mIoU	FB-IoU
E_{sc}	E_{imc}			
			41.2	69.1
✓			41.6 ($\pm$0.4)	69.3 ($\pm$0.3)
	✓		41.9 ($\pm$0.3)	69.8 ($\pm$0.4)
✓	✓		43.4 ($\pm$0.3)	70.3 ($\pm$0.4)
		✓	50.1 ($\pm$0.4)	71.7 ($\pm$0.5)
✓		✓	50.6 ($\pm$0.8)	72.0 ($\pm$0.4)
	✓	✓	50.5 ($\pm$0.6)	71.9 ($\pm$0.3)
✓	✓	✓	**51.0** ($\pm$0.6)	**72.5** ($\pm$0.6)

3.3. Ablation Study

We conduct a series of ablation experiments in Table 3. Without loss of generality, the ablation study experiments are performed on "Ours (HSNet) ‡" with ResNet101 backbone on COCO-20^i. First, when only with the E_{sc} or E_{imc}, the proposed method achieves mIoU improvement of 0.4% and 0.7% respectively and their combination leads to 2.2% mIoU improvement. Then, when only using the inter-class confidence term T, the proposed method achieves mIoU improvements of 8.9%, and FB-IoU improvements of 2.6%. Next, with T, E_{sc} and E_{imc} of the intra-class confidence term R contributes further mIoU improvements to different extents, which are shown in the 6_{th} and 7_{th} rows. Finally, the full combination of R and T achieves the best mIoU of 51.0% and FB-IoU of 72.5%. The ablation study proves the effectiveness of both R and T in the proposed method.

4. CONCLUSION

We have presented a novel semi-supervised FSS framework. The core idea is to expand the initial support set by introducing pseudo labeled images in both training and test phase. A ranking algorithm is proposed in the framework to eliminate the noisy intra-class samples and inter-class samples. Then, the pseudo labels with high ranking scores are kept and utilized to expand the support set. Extensive experiments demonstrate the effectiveness of the proposed method and new state-of-the-arts are achieved on COCO-20^i.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0112000, the National Natural Science Foundation of China under Grant 62271119, the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC1972, the A*STAR AME Programmatic Funding A18A2b0046, the RobotHTPO Seed Fund under Project C211518008, the EDB Space Technology Development Grant under Project S22-19016-STDP, and the Natural Science Foundation of Jiangsu Province under Grant BZ2021013.

6. REFERENCES

[1] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots, "One-shot learning for semantic segmentation," *British Machine Vision Conference*, 2017.

[2] 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/CVF International Conference on Computer Vision*, 2019, pp. 9197–9206.

[3] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He, "Part-aware prototype network for few-shot semantic segmentation," in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 142–158, Springer International Publishing.

[4] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye, "Prototype mixture models for few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 763–778.

[5] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8330–8339.

[6] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao, "Mining latent classes for few-shot segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8701–8710.

[7] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[8] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 01, pp. 1–1, 2020.

[9] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia, "Simple: Similar pseudo label exploitation for semi-supervised classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15099–15108.

[10] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3060–3069.

[11] Kai Huang, Jie Geng, Wen Jiang, Xinyang Deng, and Zhe Xu, "Pseudo-loss confidence metric for semi-supervised few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8671–8680.

[12] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu, "Cross-domain adaptive clustering for semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2505–2514.

[13] Junnan Li, Richard Socher, and Steven CH Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations*, 2019.

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Neural Information Processing Systems(NeurIPS)*, 2018.

[15] Khoi Nguyen and Sinisa Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 622–631.

[16] Juhong Min, Dahyun Kang, and Minsu Cho, "Hypercorrelation squeeze for few-shot segmenation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6921–6932.

[17] 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 *ECCV*, 2014, pp. 740–755.

[18] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei, "Few-shot segmentation via cycle-consistent transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21984–21996, 2021.

[19] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim, "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 2022, pp. 108–126.

[20] Dahyun Kang and Minsu Cho, "Integrative few-shot learning for classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9979–9990.