A. Other experiments

A.1. Difficult query samples

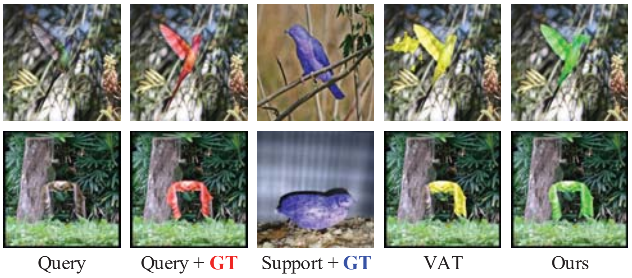

Figure 1. **Difficult query samples where query FG and query BG look similar.** Zoom in for more details.

As shown in Fig. 1, we further provide two difficult examples. We could observe that query FG objects look very similar to query BG, and it is even hard for human to perform segmentation. VAT [2] could not accurately separate query FG and BG in this case, while our model could mitigate the *FG-BG entanglement* issue well, and the segmentation results are quite good.

A.2. Weak support annotations

To further reduce annotation costs of support images, we follow existing studies [10, 12, 4] to use cheaper bounding box annotations. As shown in Tab. 1, SCCAN could outperform others well, and the performance is slightly worse than using expensive pixel-wise masks, which validates the effectiveness of SCCAN, *i.e*., query FG and BG could be well differentiated.

Method	5^0	5^1	5^2	5^3	Mean	FB-IoU
PANet† [10]	-	-	-	-	45.1	-
CANet† [12]	-	-	-	-	52.0	-
DPCN† [4]	59.8	70.5	63.2	55.5	62.3	-
SCCAN†	67.3	71.8	65.6	58.0	65.7	75.5
SCCAN‡	**67.5**	**72.6**	**67.2**	**60.5**	**67.0**	**76.4**

Table 1. **Study on weak support annotations**. Support annotations are whittled from pixel-wise masks to bounding boxes. DPCN utilizes multi-scale testing. † means using bounding boxes, ‡ means using pixel-wise masks.

A.3. GPU memory cost

We further show the GPU memory cost of our SCCAN (with 8 SCCA blocks), CyCTR [13] and VAT [2] in Tab. 2. Compared with other attention-based methods, we could observe that our SCCAN could save much GPU memory, which could demonstrate the effectiveness of our design.

Method	Input shape	GPU memory (Mb)
CyCTR (NIPS'21) [13]	$4 \times 3 \times 473 \times 473$	25,463
VAT (ECCV'22) [2]		23,553
SCCAN (Ours)		**10,667**

Table 2. **GPU memory cost of different methods**. We uniformly set the batch size as 4, and set the image size as 473×473.

A.4. More testing episodes on COCO-20^i

To make test results more reliable, we follow PFENet [8] to randomly sample 20,000 episodes from COCO-20^i to perform meta-test again, and show the results in Tab. 3. As we could observe from the table, our SCCAN could still outperform previous state-of-the-arts by large margins, including HSNet [6], DCAMA [7] and VAT [2].

B. Discussions

B.1. Support background utilization

Recall that we target on cross attention methods for FSS in this paper, *i.e*., we aim to use cross attention to combine query features with support FG features. However, existing methods suffer from two issues, namely, *BG mismatch* and *FG-BG entanglement*, both of which are raised due to the fact that query BG cannot find matched features in support FG. As a result, query BG will inevitably be fused with dissimilar features and get biased. In addition, query FG also correctly aggregate matched support FG features, and as both query FG and BG take in support FG, they get entangled, which is against the goal of FSS, *i.e*., distinguish query FG and BG.

Naturally, a naïve idea is to find matched BG features from support BG. Nevertheless, we claim that it is not appropriate to use support BG in cross attention-based FSS, which is explained as follows.

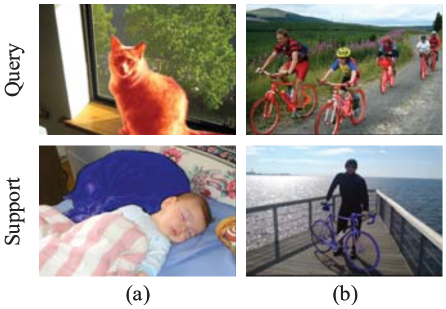

Figure 2. **Dissimilar background of query and support images**.

(1) Support BG is not always similar to query BG, *e.g*., in Fig. 2(a), two cats in query and support images are surrounded by window and bed, respectively; while in Fig. 2(b), the background objects for bicycle are mountain and sea, respectively. As query BG still cannot find matched features in support samples, the *BG mismatch* and *FG-BG entanglement* issues remain.

(2) Even though support and query BG could be simi-

Backbone	Method	1-shot						5-shot					
		20^0	20^1	20^2	20^3	Mean	FB-IoU	20^0	20^1	20^2	20^3	Mean	FB-IoU
ResNet50	HSNet‡ (ICCV'21) [6]	35.1	41.8	38.8	38.4	38.5	65.4	41.9	49.8	46.7	44.3	45.7	69.0
	DCAMA‡ (ECCV'22) [7]	38.3	41.9	43.3	40.0	40.9	63.5	43.5	49.2	49.6	46.3	47.2	66.2
	VAT‡ (ECCV'22) [2]	36.5	43.2	41.3	38.9	40.0	66.2	41.9	49.7	48.3	44.4	46.1	69.4
	SCCAN† (Ours)	38.5	49.1	45.4	43.7	44.2	68.1	45.4	54.7	52.7	50.7	50.9	71.8
	SCCAN‡ (Ours)	**39.8**	**50.2**	**47.7**	**45.7**	**45.8**	**69.7**	**47.6**	**57.7**	**57.5**	**53.0**	**59.9**	**74.2**
ResNet101	HSNet‡ (ICCV'21) [6]	35.9	44.8	41.4	40.9	40.8	66.0	43.2	51.4	48.9	47.3	47.7	69.9
	DCAMA‡ (ECCV'22) [7]	39.7	46.0	45.2	40.2	42.8	64.2	45.7	54.7	52.7	46.4	49.9	67.5
	SCCAN† (Ours)	41.0	51.5	46.9	46.1	46.4	68.3	47.8	58.5	56.6	53.4	54.1	73.2
	SCCAN‡ (Ours)	**42.3**	**52.2**	**49.5**	**47.9**	**48.0**	**69.8**	**49.4**	**61.4**	**60.2**	**55.8**	**56.7**	**74.8**

Table 3. **Testing results of 20,000 episodes on COCO-20^i. Bold** results represent the best performance, while the underlined results indicate the second best. † and ‡ indicate that the resize methods from PFENet [8] and HSNet [6] are used during testing, respectively.

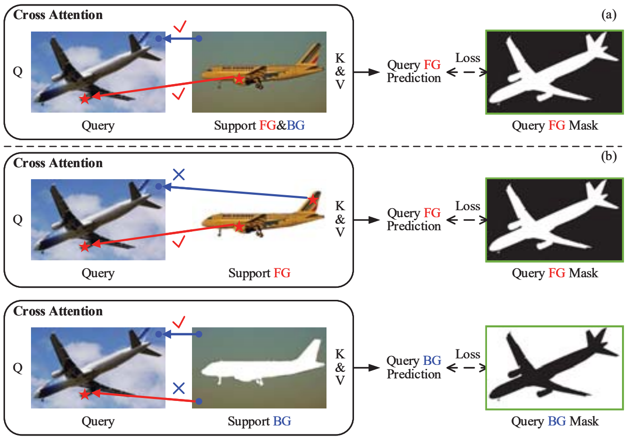

Figure 3. **Two possible ways of using support BG. (a) Single cross attention. (b) Double cross attentions.**

lar, cross attention will still fail. There are roughly two attempts: (a) Use the complete support features as $K\&V$. Although cross attention works normally now (as in Fig. 3(a)), this way is against the expectation of FSS, *i.e.*, the model could not know which part in the support sample is the region of interest (ROI). Specifically, the sky in query is similar to that in support, and the planes in two images are also similar. Recall that in FSS, different objects in query image will be considered as FG, with the changing of support information, *e.g.*, if only sky is provided in support, then the model is expected to extract the sky-related pixels from query and consider them as FG. Thus, we cannot use the whole support features as $K\&V$. (b) Jointly train two cross attentions. As shown in Fig. 3(b), two rows aim at extracting plane and sky, respectively. Although the problem in (a) is solved, the aforementioned *BG mismatch* and *FG-BG entanglement* come back, *e.g.*, in two rows, query BG could not find matched BG in support FG, and query FG cannot find its matched FG in support BG, respectively.

In summary, support BG cannot be effectively used in cross attention-based FSS to mitigate our proposed issues, and our solution is absolutely effective and novel.

B.2. Pseudo mask aggregation and PFENet

Pseudo/Prior masks are recently popular in FSS, which could roughly locate query FG objects without learnable parameters, by measuring similarity between high-level query and support features (with annotations) that are directly obtained from the pretrained backbone.

PFENet [8] firstly propose such cheap but effective mechanism. Specifically, it firstly measures the similarity between each pair of query pixels and support FG pixels. Then, for each query pixel, its largest similarity score is normalized and taken as its probability of being query FG.

However, there exist two issues: (1) PFENet only uses support FG for comparison, which will be limited when the gap between query FG and support FG is large, *e.g.*, human head (query FG), and human arms (support FG). Consequently, the similarity scores between support FG and query FG/BG are both small, and thus, the locating function of pseudo masks are weakened. For instance, in the forth column of Fig. 4, the ships in query and support are not so similar. Although the generated mask could still locate the ship in query, there are many wrongly activated BG pixels, and the model is likely to also take them as FG. (2) As PFENet only takes each query pixel's largest similarity score for reference, the generated pseudo mask is not robust, *e.g.*, it will be heavily affected by noises. As we could observe from the first column of Fig. 4, the pseudo mask of PFENet cannot locate the car in query image well.

As introduced in the paper, PMA could mitigate these issues by: (1) We also take support BG into consideration, *e.g.*, human head (query FG) is more similar to human arms (support FG), compared with room (support BG). (2) For each query pixel, we calculate its similarity scores with all support pixels, including FG and BG. Then, the normalized scores are used to aggregate the support mask values (FG is 1, BG is 0) and generate the pseudo mask. In this way, the side-effect of single largest value could be suppressed.

Some training-agnostic pseudo masks obtained from PFENet [8] and our PMA module are visualized in Fig. 4. Besides, we further use threshold 0.75 to binarize the pseudo masks and directly measure their mIoU scores (av-

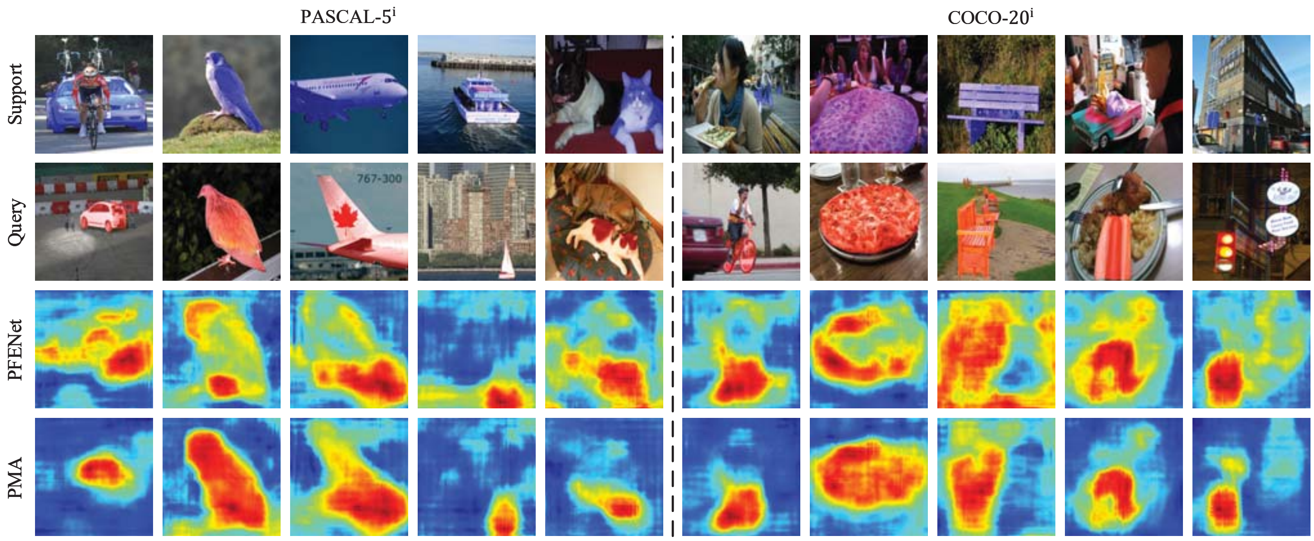

Figure 4. **Comparisons of training-agnostic pseudo mask generation methods between PFENet [8] and our proposed pseudo mask aggregation (PMA) module.**

eraged from 4 folds) on PASCAL-5^i, with ResNet50 as the backbone. The scores of pseudo masks from PFENet and PMA are 22.8% and 38.7%, respectively. In conclusion, PMA could consistently outperform PFENet in two aspects: (1) PMA is better at locating query FG. (2) There are less wrongly activated BG pixels in the generated pseudo masks.

B.3. Comparison with CyCTR [13]

It is confusing that CyCTR [13] seems to be a solution to the *BG mismatch* issue, but it is actually not, and we explain the reasons as follows. (1) It is more appropriate to claim that CyCTR is likely to not have the above issue, when *query and support* **BG** belong to the *same class*. However, we mention the side-effects of using support BG in cross attention-like methods in Appendix B.1. (2) When **BG** *classes differ*, CyCTR consistently suffers from the above issue. Its cycle-consistent attention does not solve this issue, which starts from a support BG pixel P_S, finds its most similar query pixel P_Q, and find P_Q's most similar support pixel P'_S. P_S is preserved as long as P'_S belongs to BG, regardless of the difference between **BG** classes, *i.e.*, its cross attention still fuses *dissimilar* support BG to query BG.

B.4. Comparison with VAT [2]

VAT [2] looks similar to our SCCAN in the following two aspects: (1) It is also built upon swin transformer [5]. (2) VAT converts FSS task to semantic correspondence task which focuses on features matching and fusion.

However, VAT and existing cross attention FSS methods [9, 11, 13] only target at the matching of **query FG**, and suffer from the *BG mismatch* and *FG-BG entanglement* issues. Instead, our main purpose is to match and fuse **query BG** with appropriate BG features to mitigate these issues.

B.5. Comparison with BAM [3]

BAM [3] is a latest baseline, however, we do not include it in the main results because BAM adopts a special setting, which is a bit different from the standard one. That is, BAM extends standard FSS methods by using base classes' segmentation results, and its meta learner could be any standard FSS model (including ours).

In spite of the setting difference, with ResNet-50, our model (70.3%) could still achieve comparable 5-shot results with BAM (70.9%) on PASCAL-5^i, and our model (52.3%) surpasses BAM (51.2%) on COCO-20^i.

B.6. Comparison with SSP [1]

SSP [1] also focuses on the matching issues, and we explain the differences, as well as the superiority of our model, as follows:

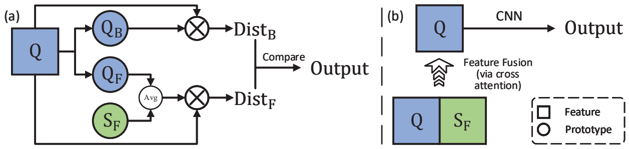

Figure 5. **Frameworks (a) SSP and (b) SCCAN.**

(1) As illustrated in Fig. 5(a), SSP coarsely separates query FG and BG first, based on which FG and BG prototypes are then constructed. Finally, similarities are calculated between query features and two prototypes for segmentation. Instead, our model (as shown in Fig. 5(b)) en-

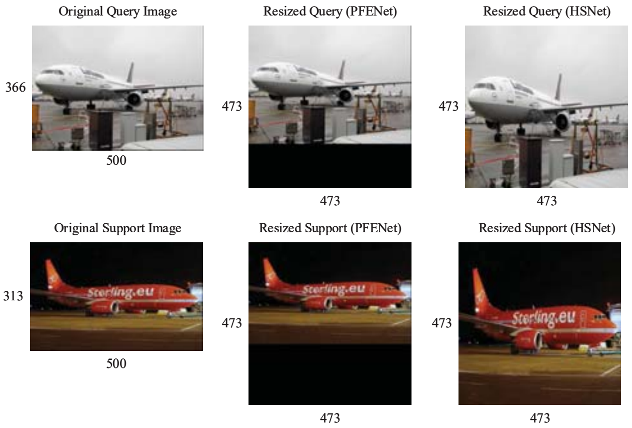

Figure 6. **Different resize methods adopted in existing baselines. (a) Original aspect ratio. (b) Different aspect ratio.**

hances query features with the concatenation of query features and support FG, and then use CNN for segmentation.

(2) SSP's BG prototype is unreliable. (a) With the coarse separation in the first phase, we cannot guarantee if the obtained BG prototype contains pure BG information or not; (b) Query BG usually contains multiple objects, *e.g.*, tree and grass in the second row of Fig. 1. As BG prototype only contains features that are most dissimilar to support FG, it may only include grass features in this case. Hence, tree features and BG prototype still mismatch.

(3) Our model outperforms SSP by a large margin, *e.g.*, 1-shot mIoU on PASCAL-5^i is 6.1% better than that of SSP.

C. Different resize methods

As mentioned in the tables of quantitative results, there are two resize methods in existing state-of-the-arts methods, and we illustrate their difference in Fig. 6. In short summary, the resize method used in PFENet [8] will resize the input query and input images, while keeping the original aspect ratio, and the shorter edge will be padded. Instead, HSNet [6] directly resizes the images to the specific shape, *e.g.*, 473×473. As a result, the resized objects in HSNet will be larger than those in PFENet. Particularly, HSNet could access to better support information, and the obtained segmentation results could be better.

References

[1] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 701–719. Springer, 2022. 3

[2] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1, 2, 3

[3] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022. 3

[4] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022. 1

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[6] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021. 1, 2, 4

[7] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022. 1, 2

[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 and machine intelligence*, 2020. 1, 2, 3, 4

[9] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *European Conference on Computer Vision*, pages 730–746. Springer, 2020. 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/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 1

[11] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019. 3

[12] 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*, pages 5217–5226, 2019. 1

[13] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021. 1, 3