Focus on Query: Adversarial Mining Transformer for Few-Shot Segmentation Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

1 More Details for Multi-scale Object Mining Transformer.

In the object mining transformer *G*, we alternately use down-sampling layers and self-attention layers to construct hierarchical query features. And the corresponding pseudo support features are obtained by the Hadamard product of the downsampled $M_{q\tau}$, specifically,

$$\mathbf{F}_{q,l} = \mathbf{Down}(\mathcal{F}^{-1}(\mathbf{FeatAgg}(\mathcal{F}(\mathbf{F}_{q,l-1}), \mathcal{F}(\mathbf{F}_{q,l-1})))), \tag{1}$$

where the $\mathcal{F} : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{HW\times C}$ is the spatial flatten operation and $\mathbf{Down}$ denotes the down-sampling layers, which is implemented with convolutional layers of double strides. In this way, we obtain multi-scale query features $\mathbf{F}_{q,l} \in \mathbb{R}^{\frac{H}{2^{l-1}}\times\frac{W}{2^{l-1}}\times C}, l = 1, \ldots, L$ and pseudo support features $\mathbf{F}_{psd,l} = \mathbf{F}_{q,l} \odot \varphi(M_{q\tau})$. ($\varphi$ is bilinear interpolation operation). We perform feature aggregation within each scale to enable contextual information exploration, thus avoiding the spatial inconsistency in the query image. L is set to 3 in our experiments. The embedding dim is set to 64, and the number of head is set to 4 in all the attention layers.

2 Detailed Experimental Settings

To achieve a fair comparison with previous methods [1–3], we adopt the ensemble strategy following BAM [1] to filter the base categories seen during training. Specifically, a base learner is trained using the training splits in a fully supervised manner, and the learned base learner is used to explicitly predict the targets of base classes. PSPNet [4] is adopted as the base learner in all of our experiments, and we follow the BAM [1] to ensemble the prediction of the base learner and the proposed AMFormer. Our code is available at `https://github.com/****/****` (Anonymity will be lifted after publication).

We also conduct an additional ablation experiment to evaluate the influence of the ensemble strategy as shown in Table 1. We can observe that the ensemble strategy can incrementally improve performance. It should be noted that our AMFormer can also surpass previous state-of-the-art methods (IPMT [5]) without the ensemble strategy.

Table 1: Ablation of ensemble strategy on Pascal-5^i with ResNet-101 backbone and 1-shot setting.

	fold0	fold1	fold2	fold3	mean
IPMT [5]	**71.6**	73.5	58.0	61.2	66.1
w/o ensemble	69.7	**75.2**	**69.5**	**62.9**	**69.3**
w/ ensemble	71.3	76.7	70.7	63.9	70.7

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3 Dataset Settings

We respectively divided Pascal-5^i and COCO-20^i into four splits following [6] for cross-validation. In Table 2 and Table 3, we provide the detailed split settings.

Table 2: Detailed splits setting of PASCAL-5^i

Fold	Test classes
PASCAL-5^0	aeroplane, bicycle, bird, boat, bottle
PASCAL-5^1	bus, car, cat, chair, cow
PASCAL-5^2	diningtable, dog, horse, motorbike, person
PASCAL-5^3	potted plant, sheep, sofa, train, tv/monitor

4 More Experimental Results

4.1 Quantitative analysis of intra-object similarity.

We compute the average pairwise pixel similarity from the same object (intra-object) and different objects from the support and query images of the same category (inter-object) using the cosine similarity. Note that the pixel features that we used to compute the similarity are middle-level features $\mathbf{F}_s$ and $\mathbf{F}_q$ as described in the L255-L256 in the original manuscript. The quantitative results of different categories are provided in Table 4 and Table 5. From the tables, we can observe that the intra-object similarity is at least one order of magnitude higher than the inter-object similarity. This demonstrates the superiority of the query-centric approach relying on intra-object similarity over support-centric methods that rely on inter-object similarity.

4.2 More visualization results.

Visualization of segmentation at different stages. We tackle query-centric FSS by applying three intuitive procedures. (1) Discriminative region localization. (2) Local to global expansion. (3) Coarse to fine alignment. To illustrate the effects of the above three steps, in Figure 1, we visualize the outcomes of different stages. Procedure (1) can only roughly local the discriminative region of the target (2^{nd} column of Figure 1). In procedure (2), the object mining transformer *G* exploits the intra-object similarity to explore multi-scale contextual information, thus highlighting the whole object (3^{rd} column of Figure 1). Segmentation from *G* can roughly cover the entire target but there still exist misalignments as shown in the yellow boxes in the 3^{rd} column of Figure 1. The detail mining transformer *D* is responsible for discriminating those detailed misalignments, *i.e.*, procedure (3). The proposed AMFormer couples procedures (2) and (3) via adversarial training. In this way, the G can be optimized to generate more accurate segmentations(4^{th} column of Figure 1) to fool *D*.

Visualization of activation maps. We visualize the attention weight of query features between the support target and pseudo support. Specifically, the attention matrix $\mathbf{S} \in \mathbb{R}^{H_q W_q \times H_s W_s}$ is computed according to the Eqn (1) of the original manuscript. Then we compute the average activation of each query pixel over all support foreground pixels:

$$\mathbf{Act}(i) = \frac{\sum_{j=1}^{H_s W_s} \mathbf{S}(i,j) \cdot [\mathcal{F}(\mathbf{M}_s)(j) > 0]}{\sum_{j=1}^{H_s W_s} [\mathcal{F}(\mathbf{M}_s)(j) > 0]}, \tag{2}$$

where the $\mathbf{M}_s$ is the (pseudo) support mask. In the baseline, the $\mathbf{S}$ is oriented from the cross attention between the query features and the support features (support-centric). While in our query-centric AMFormer, the $\mathbf{S}$ is computed from the pseudo support and the query features. As shown in Figure 2, the support targets not only cannot fully activate the target in the query image, but also frequently activates the background categories. While the pseudo support can well excavate the full object attribute to intra-object similarity.

Visualization of local proxies. To explore the regions of interest for learnable local proxies, Figure 3 visualizes the activation maps of a part of proxies. It can be observed that different proxies tend to focus on different local regions, and most proxies attend to the boundaries of the object, which

Table 3: Detailed splits setting of COCO-20^i

Fold	Test classes
COCO-20^0	Person, Airplane, Boat, Park meter, Dog, Elephant, Backpack, Suitcase, Sports ball, Skateboard, W. glass, Spoon, Sandwich, Hot dog, Chair, D. table, Mouse, Microwave, Fridge, Scissors
COCO-20^1	Bicycle, Bus, T.light, Bench, Horse, Bear, Umbrella, Frisbee, Kite, Surfboard, Cup, Bowl, Orange, Pizza, Couch, Toilet, Remote, Oven, Book, Teddy
COCO-20^2	Car, Train, Fire H., Bird, Sheep, Zebra, Handbag, Skis, B. bat, T. racket, Fork, Banana, Broccoli, Donut, P. plant, TV, Keyboard, Toaster, Clock, Hairdrier
COCO-20^3	Motorcycle, Truck, Stop, Cat, Cow, Giraffe, Tie, Snowboard, B. glove, Bottle, Knife, Apple, Carrot, Cake, Bed, Laptop, Cellphone, Sink, Vase, Toothbrush

Table 4: Intra- and inter-object similarity of each class within Pascal-5^i,

Pascal-5^0			Pascal-5^1			Pascal-5^2			Pascal-5^3		
class	Intra-	Inter-	class	Intra-	Inter-	class	Intra-	Inter-	class	Intra-	Inter-
aeroplane	0.449	0.008	bus	0.453	0.011	diningtable	0.496	0.010	potted plant	0.515	0.012
bicycle	0.460	0.009	bus	0.453	0.011	dog	0.571	0.009	sheep	0.536	0.007
bird	0.493	0.010	cat	0.643	0.021	horse	0.489	0.012	sofa	0.535	0.006
boat	0.483	0.009	chair	0.519	0.015	motorbike	0.445	0.008	train	0.509	0.008
bottle	0.504	0.120	cow	0.573	0.016	person	0.529	0.039	tv/monitor	0.513	0.024

Table 5: Intra- and inter-object similarity of each class within COCO-20^i

COCO-20^0			COCO-20^1			COCO-20^2			COCO-20^3		
class	Intra-	Inter-	class	Intra-	Inter-	class	Intra-	Inter-	class	Intra-	Inter-
Person	0.553	0.002	Bicycle	0.398	0.017	Car	0.514	0.012	Motorcycle	0.485	0.007
Airplane	0.582	0.020	Bus	0.440	0.019	Train	0.544	0.017	Truck	0.552	0.021
Boat	0.585	0.013	T.light	0.474	0.013	Fire H.	0.467	0.007	Stop	0.544	0.008
Park meter	0.476	0.014	Bench	0.459	0.044	Bird	0.530	0.016	Cat	0.549	0.010
Dog	0.482	0.010	Horse	0.574	0.018	Sheep	0.487	0.010	Cow	0.584	0.015
Elephant	0.467	0.012	Bear	0.460	0.012	Zebra	0.510	0.012	Giraffe	0.556	0.013
Backpack	0.479	0.013	Umbrella	0.571	0.020	Handbag	0.519	0.032	Tie	0.559	0.006
Suitcase	0.542	0.007	Frisbee	0.384	0.005	Skis	0.432	0.007	Snowboard	0.509	0.014
Sports ball	0.538	0.009	Kite	0.470	0.016	B.bat	0.532	0.023	B.glove	0.576	0.006
Skateboard	0.537	0.018	Surfboard	0.589	0.024	T.racket	0.373	0.012	Bottle	0.545	0.007
W.glass	0.434	0.011	Cup	0.606	0.009	Fork	0.451	0.017	Knife	0.589	0.013
Spoon	0.562	0.027	Bowl	0.563	0.010	Banana	0.511	0.015	Apple	0.656	0.010
Sandwich	0.469	0.013	Orange	0.474	0.028	Broccoli	0.471	0.026	Carrot	0.539	0.007
Hot dog	0.558	0.008	Pizza	0.515	0.015	Donut	0.477	0.300	Cake	0.597	0.006
Chair	0.442	0.015	Couch	0.542	0.013	P.plant	0.433	0.032	Bed	0.560	0.007
D.table	0.474	0.046	Toilet	0.503	0.006	TV	0.456	0.014	Laptop	0.480	0.021
Mouse	0.447	0.021	Remote	0.583	0.011	Keyboard	0.542	0.007	Cellphone	0.498	0.018
Microwave	0.454	0.019	Oven	0.567	0.011	Toaster	0.433	0.007	Sink	0.375	0.027
Fridge	0.476	0.023	Book	0.586	0.025	Clock	0.526	0.018	Vase	0.567	0.013
Scissors	0.456	0.027	Teddy	0.583	0.016	Hairdrier	0.373	0.007	Toothbrush	0.479	0.013

is usually the most ambiguous region. In addition, a particular proxy consistently focuses on the boundaries in a specific direction, *e.g.*, *proxy 1* always activates the right border (5^{th} column). Through the cooperation of multiple proxies, our detail mining transformer *G* can effectively detect the detailed local differences between the prediction of the object mining transformer *G* and the ground truth. By means of adversarial training, G will produce more accurate segmentations, especially in ambiguous regions, by adjusting itself to fool D.

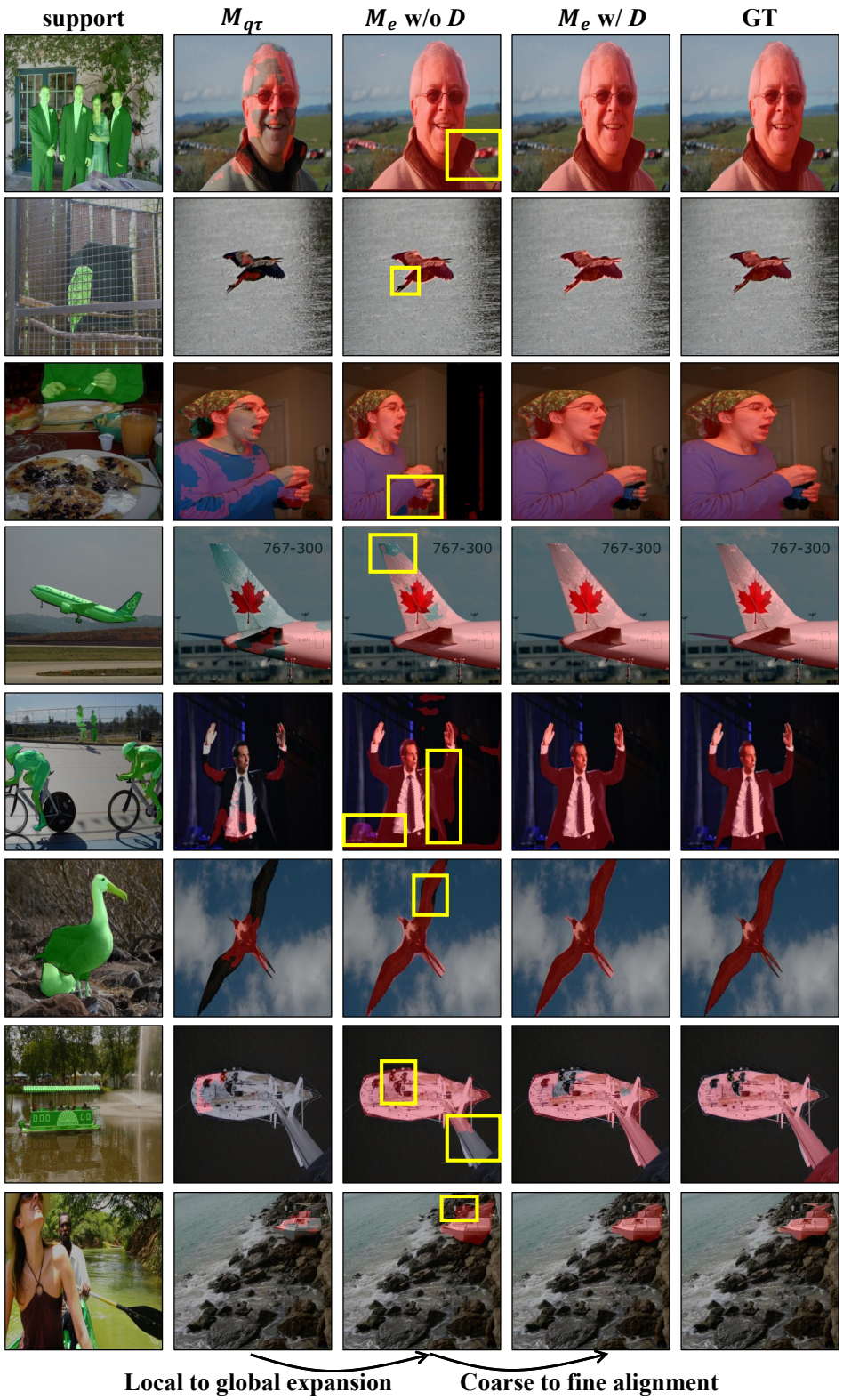

Figure 1: Visualization of the segmentation of the proposed AMFormer at different stages

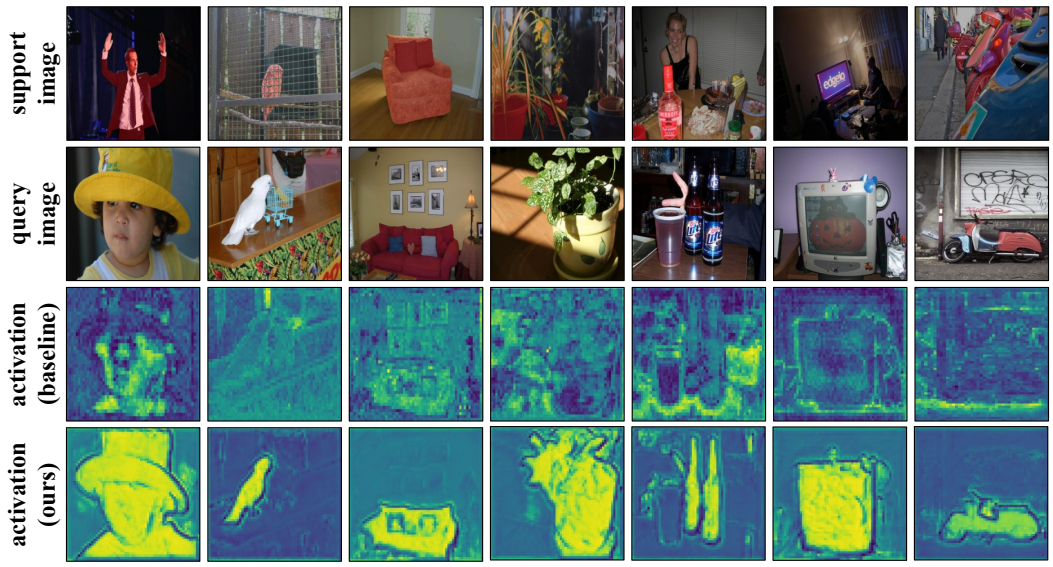

Figure 2: Visualizations of the attention weight of query features between the support target (support-centric baseline) and pseudo support (ours).

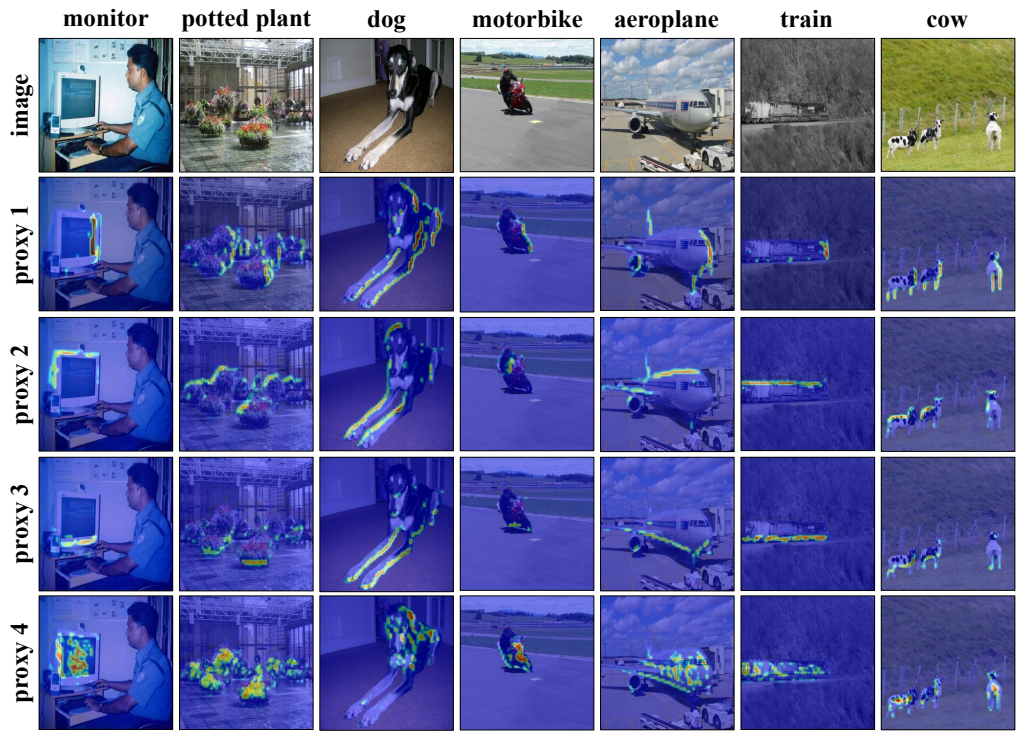

Figure 3: Visualizations of the activated regions of local proxies.

References

[1] 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.

[2] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chenyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. *arXiv preprint*

arXiv:2303.14652, 2023.

[3] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*, 2022.

[4] 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.

[5] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. *arXiv preprint arXiv:2210.06780*, 2022.

[6] 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*, (01):1–1, 2020.