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

Yuan Wang[1]*, **Naisong Luo**[1]*, **Tianzhu Zhang**[1]†
[1]Deep Space Exploration Laboratory/School of Information Science and Technology,
University of Science and Technology of China, Hefei 230026, China.
`{wy2016,lns6}@mail.ustc.edu.cn`, `tzzhang@ustc.edu.cn`

Abstract

Few-shot segmentation (FSS) aims to segment objects of new categories given only a handful of annotated samples. Previous works focus their efforts on exploring the support information while paying less attention to the mining of the critical query branch. In this paper, we rethink the importance of support information and propose a new query-centric FSS model Adversarial Mining Transformer (AMFormer), which achieves accurate query image segmentation with only rough support guidance or even weak support labels. The proposed AMFormer enjoys several merits. First, we design an object mining transformer (G) that can achieve the expansion of incomplete region activated by support clue, and a detail mining transformer (D) to discriminate the detailed local difference between the expanded mask and the ground truth. Second, we propose to train G and D via an adversarial process, where G is optimized to generate more accurate masks approaching ground truth to fool D. We conduct extensive experiments on commonly used Pascal-5^i and COCO-20^i benchmarks and achieve state-of-the-art results across all settings. In addition, the decent performance with weak support labels in our query-centric paradigm may inspire the development of more general FSS models. Code will be available at `https://github.com/Wyxdm/AMNet`

1 Introduction

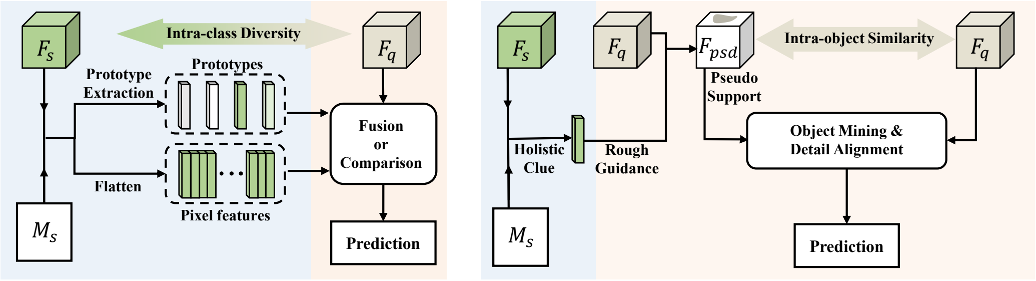
As a fundamental computer vision task, semantic segmentation [1] has drawn decades of research interest in both academia and industry. However, the underlying driving force of fully supervised segmentation models is massive, densely labeled data. The inherent category sensitivity of these models leads to severe performance degradation when dealing with previously unseen categories. To equip segmentation models with generalization to novel classes, few-shot segmentation (FSS) has been proposed [2], which achieves segmentation of new category images (called query images) with only a handful of annotated samples (called support images) without retraining the model.

The current top-performing FSS methods focus on adequately exploring the information contained in the support samples to guide the segmentation of the query images. Among them, prototypical learning methods [3–6] concentrate support features into one or several representative prototypes to direct the classification of query pixels, while the affinity learning approaches [7–9] try to equip the query features with support semantic clue of pixel granularity. Although considerable progress has been made in the direction of probing support information, we noticed that significant intra-class diversities (as illustrated in Figure 2) between support and query images are common occurrences. In these cases, the target feature from the query image may be quite distinct from the support one.

*Equal contribution.

†Corresponding Author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Support-centric approaches **(b) Our query-centric approach**

Figure 1: Comparison between the previous methods and ours. (a) Most of the previous methods focus their efforts on extracting more support information. They condense support information into prototypes [3, 4, 6, 12] or directly explore pixel-level support features [7–9, 13]. (b) The proposed AMFormer focus on the query features and exploits intra-object similarity to mining the complete target, forming a query-centric FSS method. Only rough category guidance from support is needed in our approach.

Hence there raises a natural question: is exhaustive support information extraction indispensable for query image segmentation?

We revisit the role of support information in few-shot segmentation through a pilot study. As shown in Table 1, we randomly erode support foreground features with varying proportions to explore how the support information completeness affects query segmentation accuracy. It can be observed that the performances do not degrade significantly or even surpass the complete one. We deem two main reasons contribute to the phenomenon: 1) Rough category guidance from support is sufficient to guide the accurate query image segmentation. 2) Forcing fine-grained support information may introduce redundancy, or even bias, which confuses the prediction of query images, particularly when intra-class diversities are prevalent. Motivated by the above insights, we adjust the direction and focus on the query branch. Building on the observation of intra-object similarity, *i.e.*, pixels within the same object are more similar than those across different objects [10, 11] (as shown in Figure 2 and Table 8, we argue that partially activated object regions can serve as cues to infer whole objects. In this paper, we construct a pseudo support image that has low intra-class variation with the query image by activating query image with a rough category-level support clue. Then we exploit the pseudo support feature to guide the query segmentation, forming a novel query-centric exploration strategy as illustrated in Figure 1.

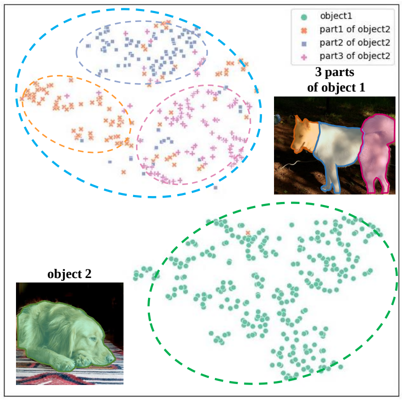


Figure 2: t-SNE visualization of the intra-class diversity and the intra-object similarity. As shown by the blue and green circles, different instances of the same category may be far apart in the feature space, *e.g.*, dog. While pixels from the same object share high feature similarity, as shown by the three circles within the blue one.

We tackle query-centric FSS by applying three intuitive procedures. **(1) Discriminative region localization.** The holistic semantic clue from the labeled support image is utilized to roughly locate the most discriminative target area on the query image. **(2) Local to global expansion.** We utilize intra-object feature similarity to explore contextual information and then highlight the less discriminative object parts. **(3) Coarse to fine alignment.** The coarse activated region is further refined by eliminating the detailed difference with ground truth. Different from previous methods that focus on extracting fine-grained support information, Our method is less susceptible to intra-class diversity as only a rough holistic support guidance is needed in the first procedure to generate the pseudo support. More attention is paid to the last two query-focused procedures that enable growing incomplete local regions into accurate segmentations.

We propose an end-to-end Adversarial Mining Transformer(AMFormer) to couple procedures (2) and (3) and mutually enhance each other, which optimizes an object mining transformer G and a detail mining transformer D via an adversarial process. AMFormer aims at fully excavating the target area and aligning it in detail under the guidance of the pseudo support features. **To dig out the entire object**

region, in the G, we conduct multi-scale pixel-level correlation between the query features and pseudo support features. The exploration of the intra- and inter-scale contextual information in G progressively highlights the latent target parts that share high intra-object similarities. **To effectively align the prediction details**, *e.g.*, object boundaries, D models multiple detail proxies to accurately distinguish the subtle local differences between the ground truth and the prediction generated by G. By means of adversarial training, G is optimized to predict more accurate segmentations to fool D and only the well-learned G is required to produce accurate predictions in the inference stage.

Table 1: The performance (mIoU) under different support feature erosion ratios. We conduct this pilot study on 1^{st} split of Pascal-5^i using ResNet-50 and 1-shot setting. * means reproduced results.

Erosion Ratio (%)	80	65	50	35	20	0*
PFENet [14]	60.6	61.0	**61.5**	61.2	61.0	61.2
HDMNet [13]	69.3	70.0	**70.5**	70.4	69.9	70.2

We evaluate our AMFormer on two widely used benchmarks COCO-20^i and Pascal-5^i with different backbones and the AMFormer consistently sets new state-of-the-art on different settings. Moreover, our query-centric approach does not reckon on elaborate support annotations. It achieves remarkable performance even with weak support labels such as scribbles or bounding boxes, formulating a more practical FSS model. We also hope that our work can inspire more research on query-centric FSS methods. **To recap**, our contributions are concluded as follows: (i) We re-evaluate the importance of support information in FSS and demonstrate that a coarse category hint suffices for accurate query segmentation. This motivated us to put forward a novel query-centric FSS method. (ii) We propose a novel Adversarial Mining Transformer (AMFormer) that optimizes an object mining transformer G and a detail mining transformer D for region expansion and detail alignment, respectively. (iii) Extensive experiments show that AMFormer significantly outperforms previous SOTAs. The conspicuous performance with weak support label also sheds light for future research of more general FSS models.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation has been widely applied to autonomous driving, medical image processing [15–17] and so on. The aim of semantic segmentation is to assign each pixel within the given image to a predefined category label. The seminal FCN [1] sparked a wave of remarkable advances in semantic segmentation [18–20] based on convolutional neural networks (CNN). Various networks focus on better context exploration by enlarging the receptive field of CNN via dilated convolutions [21, 22], global pooling [23] and pyramid pooling [21, 24]. In addition to CNN-based models, the success of Vision Transformer (ViT) [25] encourages a series of transformer-based segmentation models [26–29]. For instance, maskformer [30] adopts the transformer decoder [31] to conduct the mask classification based on a set prediction objective. Many subsequent works improve this framework [32–35] and gradually formed a unified segmentation model that can address different image segmentation tasks. Despite the success, these methods cannot generalize to novel classes in the low-data regime.

2.2 Few-Shot Semantic Segmentation

Few-shot segmentation (FSS) [2] is established to segment new category images (query images) with only a few labeled samples (support images). Owing to the reliable annotations in the support set, current FSS methods mainly focus on effectively excavating support information, which can be roughly divided into two categories: prototypical learning methods and affinity learning methods. Motivated by PrototypicalNet [36], many previous works condense the support information into single [37–39, 6, 40] or multiple prototypes [41–43, 5, 4, 44–46] and then conduct feature comparison or aggregation. For example, ASGNet [3] adopts superpixel-guided clustering to adaptively construct multiple prototypes, which are concatenated with the most relevant query features to guide the pixel classification. Differently, for the seek of fine-grained support guidance, affinity learning methods [13, 47, 8, 48–50] are designed to establish pixel-level associations between support and query features via attention mechanism [7, 9] or cost volume aggregation [8, 48]. For instance, CyCTR [7] introduces a cycle-consistent attention mechanism to equip query features with relevant support information. Though achieving promising results, these methods depend heavily on support information and are prone to segmentation errors in the presence of large intra-class variations. Some methods[11, 6] try to solve this problem by mining the class-specific representation from the query branch but remain at relatively coarse prototype granularity. In this paper, we propose a novel query-centric approach that exploits intra-object similarity to probe query targets with only overall category-level guidance from support.

3 Method

3.1 Problem Definition

Few-shot segmentation (FSS) tackles novel class object segmentation with only a few densely-annotated samples. Episodic meta-training [51] is widely used to enhance the generalization of FSS models. Specifically, the dataset is divided into the training set $\mathcal{D}_{train}$ and the testing set $\mathcal{D}_{test}$. The category sets of $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are disjoint, *i.e.*, $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. A series of episodes are sampled from $\mathcal{D}_{train}$ to train the model, each of which is composed of a support set $\mathcal{S} = \{I_s^k, M_s^k\}_{k=1}^K$ and a query set $\mathcal{Q} = \{I_q, M_q\}$ in the K-shot setting, where I and M denote the RGB image and corresponding binary masks, respectively. Under the supervision of M_q, the model is trained to predict the query mask conditioned on the $\mathcal{S}$ and I_q. After that, the trained model is evaluated on the episodes sampled from $\mathcal{D}_{test}$ without further optimization.

3.2 Revisiting of Transformer-based Feature Aggregation

Transformer layers [52] are widely used in computer vision tasks for feature aggregation. The critical component of transformer layer is the attention mechanism that enables the long-range modeling capability. Specifically, the attention layer is first applied to compute the attention weight of the source feature sequence $\mathbf{S} \in \mathbb{R}^{N_1 \times C}$ and the target sequence $\mathbf{T} \in \mathbb{R}^{N_2 \times C}$, where the N_1, N_2 denote the length of sequences and C is the embedding dimension, formulaic as:

$$\mathbf{S} = \mathrm{Softmax}(\frac{\mathbf{Q}(\mathbf{K})^{\mathsf{T}}}{\sqrt{\mathrm{d}}}), \quad \mathbf{Q} = \mathbf{T}\mathbf{W}^{\mathrm{Q}}, \quad \mathbf{K} = \mathbf{S}\mathbf{W}^{\mathrm{K}}, \tag{1}$$

in which $\mathbf{W}^{\mathrm{Q}}$ and $\mathbf{W}^{\mathrm{K}} \in \mathbb{R}^{C \times d}$ are learnable linear projections and $\sqrt{d}$ is the scaling factor. The source information is adaptively transported to the target sequence according to attention weight, and a feed-forward network (FFN) is further applied to transform the fused features:

$$\widehat{\mathbf{T}} = \mathbf{FFN}(\mathbf{S}\mathbf{V}), \quad \mathbf{V} = \mathbf{S}\mathbf{W}^{\mathrm{V}}, \tag{2}$$

where the $\mathbf{W}^{\mathrm{V}}$ is the linear linear projection and $\widehat{\mathbf{T}}$ is the enhanced target feature sequence. We can abbreviate the feature aggregation process as:

$$\widehat{\mathbf{T}} = \mathbf{FeatAgg}[\mathbf{T}, \mathbf{S}]. \tag{3}$$

Note that when $\mathbf{T} = \mathbf{S}$, the $\mathbf{FeatAgg}(,)$ explores contextual information within the feature sequence, *i.e.*, acts as the self-attention mechanism. The above processes are implemented with the multi-head mechanism to enhance performance further.

3.3 Adversarial Mining Transformer

3.3.1 Overview

The proposed framework tailored for query-centric FSS consists of three procedures, *i.e.*, 1) discriminative region localization, 2) local to global region expansion, 3) coarse to fine mask alignment. We roughly locate the object of novel class under a rough support guidance via nonparametric cosine similarity in procedure 1). The resulting incompleted region is expanded by the object miner *G* to cover the whole target in procedure 2). The detail miner *D* in procedure 3) is tasked with identifying local discrepancies between the ground truth and the expanded mask generated by *G*, thereby aligning the coarse prediction in detail. We optimize the *G* and *D* via an adversarial process and forming the end-to-end AMNet as illustrated in Fig. 3. The details are as follows.

3.3.2 Discriminative region localization

Considering the significant intra-class diversity between samples, we contend that support information is more suitable as overall guidance indicating the novel class rather than detailed reference. Given

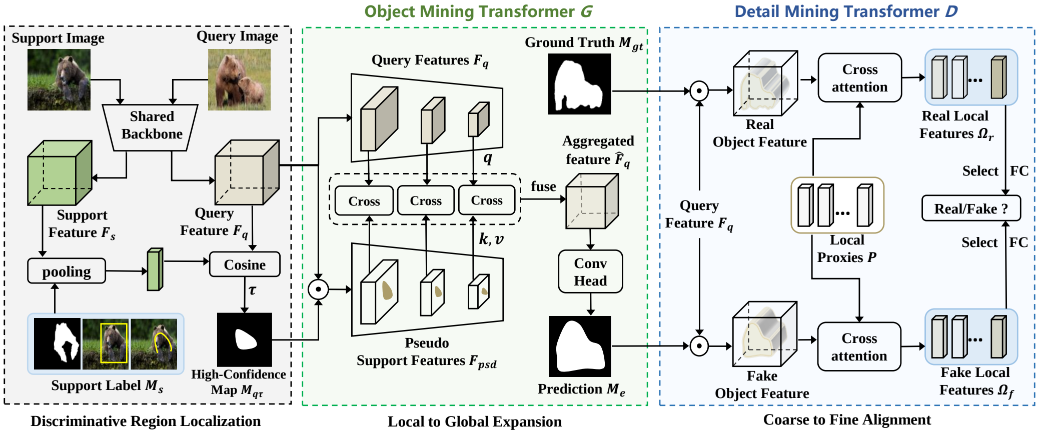


Figure 3: Illustration of the proposed AMFormer. We approach the query-centric FSS with three procedures, *i.e.*, discriminative region localization, local to global region expansion, and coarse to fine mask alignment. AMFormer optimizes an object mining transformer and a detail mining transformer via an adversarial process to couple these procedures.

the support image features $\mathbf{F}_s \in \mathbb{R}^{H\times W\times C}$ and corresponding mask $\mathbf{M}_s \in \mathbb{R}^{H\times W}$, where H and W describe the feature size and C is the feature dim, we apply mask average pooling (MAP) to obtain the mean target feature $\mathbf{F}_h \in \mathbb{R}^{1\times C}$ that serves as the holistic support guidance, formally:

$$\mathbf{F}_h = \mathbf{MAP}(\mathbf{F}_s, \mathbf{M}_s). \tag{4}$$

We exploit the $\mathbf{F}_h$ to indicate the target region in the $\mathbf{F}_q$ based on the cosine similarity:

$$\mathbf{M}_{q\tau} = \mathbb{1}^\tau(\mathrm{Softmax}(\cos(\mathbf{F}_\mathrm{h}, \mathbf{F}_\mathrm{q}))), \quad \mathbb{1}^\tau(\mathrm{x}) = \begin{cases} 1, & x > \tau \\ 0, & \mathrm{otherwise} \end{cases} \tag{5}$$

A relatively high threshold τ is set to locate the most class-relevant regions and suppress activation to backgrounds. In our experiment, τ is set to 0.7.

3.3.3 Local to global region expansion

Adaptive feature aggregation. The high-confidence object regions $\mathbf{M}_{q\tau} \in \mathbb{R}^{H\times W}$ obtained by Eqn (5) are incomplete. We introduce the object mining transformer G to expand the local region to the entire target area based on the intra-object similarity. Specifically, G aims at aggregating the representative target features from the pseudo support to less discriminative object parts in the query, formally,

$$\widehat{\mathbf{F}}_q = \mathcal{F}^{-1}(\mathbf{FeatAgg}[\mathcal{F}(\mathbf{F_q}), \mathcal{F}(\mathbf{F}_{psd})]), \tag{6}$$

where the $\mathcal{F}: \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{HW\times C}$ is the spatial flatten operation. $\mathbf{F}_{psd}$, which serve as the source sequence in Eqn (1), are essentially query features filtered by $\mathbf{M}_{q\tau}$:

$$\mathbf{F}_{psd} = \mathbf{F}_q \odot \mathbf{M}_{q\tau}. \tag{7}$$

Homogeneous feature guidance enables our aggregation to circumvent the effects of intra-class variance. Note that the original attention mechanism applies *Softmax* activation along the source sequence dimension. Directly adopting this scheme in our aggregation makes the query background features inevitably attend to foreground features since all the non-zero features in the pseudo support belong to the foreground. Inspired by [53, 13], we adjust to implement *Softmax* along the axis of $\mathbf{F}_q$. In this way, the discriminative object features are more likely to be aggregated target areas that share high intra-object similarities with $\mathbf{F}_{psd}$.

Cross-scale information transportation. We observe that in many cases, the query image contains different objects with varying scales, which are to be segmented in other episodes. To handle this spatial inconsistency, we construct the G in a multi-scale form. Concretely, we follow [13] to establish the hierarchical query features $\{\mathbf{F}_{q,l}\}_{l=1}^{L}$ with down-sampling and self-attention layers, in which

$$\mathbf{F}_q^l \in \mathbb{R}^{\frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}} \times C}, \quad l = 1, \ldots, L. \tag{8}$$

The corresponding pseudo support features$\{\mathbf{F}_{psd,l}\}_{l=1}^{L}$ is obtained by the Hadamard product of $\{\mathbf{F}_{q,l}\}_{l=1}^{L}$ and downsampled $\mathbf{M}_{q\tau}$. Then Eqn (6) is implemented independently in each scale to obtain the aggregated query features $\{\widehat{\mathbf{F}}_{q,l}\}_{l=1}^{L}$. Finally, we fuse the multi-scale aggregated query features in a top-down manner as down in [14], specifically,

$$\widehat{\mathbf{F}}'_{q,l} = \mathbf{Conv}_{3\times3}(\mathbf{Conv}_{1\times1}(\widehat{\mathbf{F}}_{q,l} + \mathcal{R}(\widehat{\mathbf{F}}_{q,l+1})) + \mathcal{R}(\widehat{\mathbf{F}}_{q,l+1})). \tag{9}$$

The fused features from the last stage $\widehat{\mathbf{F}}'_{q,1} \in \mathbb{R}^{H\times W\times C}$ are exploited to predict the expanded target area $\mathbf{M}_e \in \mathbb{R}^{H\times W}$ via a small convolution head. Benefiting from the adaptive feature aggregation and cross-scale information transportation in the object mining transformer *G*, less discriminative target parts can be highlighted in the query image, forming the expanded object mask $\mathbf{M}_e$.

3.3.4 Coarse to fine mask alignment

Detail mining transformer *D* is designed to discriminate subtle differences between the ground truth $\mathbf{M}_{gt}$ and the expanded object mask $\mathbf{M}_e$ generated by *G*, since $\mathbf{M}_e$ adequately covers the object but still exhibits misalignments in fine-grained details, e.g., boundaries. By training *G* and *D* via an adversarial process, *G* is optimized to generate more accurate target masks approaching ground truth to fool *D*, thus achieving coarse to fine prediction alignment.

To capture comprehensive details, *D* models multiple local proxies, each of which is tasked with exploring the object features respectively specified by $\mathbf{M}_e$ and $\mathbf{M}_{gt}$ to obtain a pair of local features. The most different pair is selected to output real/fake results. Concretely, we concatenate the learnable proxies as a feature sequence $\mathbf{P} \in \mathbb{R}^{N\times C}$, where N denotes the number of proxies. We perform feature adaptive aggregation through the attention mechanism to construct local features, In specific,

$$\mathbf{\Omega}_f = \mathbf{FeatAgg}[\mathbf{P}, \mathcal{F}(\mathbf{F_q} \odot \mathbf{M}_e)], \tag{10}$$

where the “fake” object features specified by predicted mask $(\mathbf{F_q}\odot\mathbf{M}_e)$ serves as the source sequence of aggregation process. Note that the $\mathbf{M}_e$ is not binarized. Similarly, local features originate from “real” object features are obtained by:

$$\mathbf{\Omega}_r = \mathbf{FeatAgg}[\mathbf{P}, \mathcal{F}(\mathbf{F_q} \odot \mathbf{M}_{gt})]. \tag{11}$$

Finally, we calculate the cosine similarity among the real local features $\mathbf{\Omega}_r = \{\boldsymbol{\omega}_f^i\}_{i=1}^{N}$ and the fake ones $\mathbf{\Omega}_f = \{\boldsymbol{\omega}_f^i\}_{i=1}^{N}$. The most different pair $(\boldsymbol{\omega}_f^k, \boldsymbol{\omega}_r^k)$ is selected and fed into a fully-connected layer to predict the fake/real results for adversarial training. By this way, *D* is optimized to discriminate detailed local differences, while the *G* will generate more precise masks to fool *D* by adjusting itself.

Since there is no explicit supervision, different proxies may focus on the same object part. We impose a diversity loss to avoid this degradation by expanding the discrepancy among local features:

$$\mathcal{L}_{div} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1,i\neq j}^{N} \left(\frac{\langle \boldsymbol{\omega}_f^i, \boldsymbol{\omega}_f^j \rangle}{\parallel \boldsymbol{\omega}_f^i \parallel_2 \parallel \boldsymbol{\omega}_f^j \parallel_2} + \frac{\langle \boldsymbol{\omega}_r^i, \boldsymbol{\omega}_r^j \rangle}{\parallel \boldsymbol{\omega}_r^i \parallel_2 \parallel \boldsymbol{\omega}_r^j \parallel_2} \right). \tag{12}$$

The intuition of Eqn (12) is trivial. If different proxies focus on the same object region, $\mathcal{L}_{div}$ will be large and adjust the learning of proxies.

3.3.5 Training and Inference

Training loss. Our architecture consists of a generative part, *i.e.*, object mining transformer *G* and a discriminative part, *i.e.*, detail mining transformer *D*. The *D* judges whether the mask is real (ground truth) or fake (generated by *G*) by mining the detailed features of the object framed by the masks. To fool *D*, *G* is supposed to predict a more accurate mask approaching the ground truth. We alternately train two parts to achieve mutual promotion. When training *D*, the parameters of *G* are frozen, and the loss function for *D* is formulated as:

$$\mathcal{L}_d = -\log(\mathbf{FC}(\boldsymbol{\omega}_r^k)) - \log(1 - \mathbf{FC}(\boldsymbol{\omega}_f^k)) + \lambda_{div}\mathcal{L}_{div}, \tag{13}$$

where the $\mathbf{FC} : \mathbf{\Omega} \rightarrow [0, 1]$ denotes the fully-connected layers that output the real/fake results, and the $\boldsymbol{\omega}_r^k, \boldsymbol{\omega}_f^k \in \mathbf{\Omega}$ denote the most different pair of local features from the object specified by ground truth $\mathbf{M}_{gt}$ and the predicted mask $\mathbf{M}_e$, respectively. λ_{div} denotes the weight of diversity loss and we

Table 2: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on PASCAL-5^i using the mIoU (%) evaluation metric. Best results are shown in bold.

Method	Backbone	1-shot					5-shot				
		fold0	fold1	fold2	fold3	Mean	fold0	fold1	fold2	fold3	Mean
PFENet[TPAMI2020] [14]	Res-50	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
RePRI[CVPR2021] [54]		59.8	68.3	62.1	48.5	59.7	64.6	71.4	71.1	59.3	66.6
HSNet[ICCV2021] [8]		64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	67.1	69.5
CyCTR[NIPS2021] [7]		65.7	71.0	59.5	59.7	64.0	69.3	73.5	63.8	63.5	67.5
NERTNet[CVPR2022] [43]		65.4	72.3	59.4	59.8	64.2	66.2	72.8	61.7	62.2	65.7
DCAMA[ECCV2022][9]		67.5	72.3	59.6	59.0	64.6	70.5	73.9	63.7	65.8	68.5
SSP[ECCV2022] [11]		60.5	67.8	66.4	51.0	61.4	67.5	72.3	**75.2**	62.1	69.3
IPMT[NIPS2022] [6]		**72.8**	73.7	59.2	61.6	66.8	73.1	74.7	61.6	63.4	68.2
VAT[ECCV2022] [48]		67.6	72.0	62.3	60.1	65.5	72.4	73.6	68.6	65.7	70.1
BAM[CVPR2022] [55]		69.0	73.6	67.6	61.1	67.8	70.6	75.1	70.8	67.2	70.9
HDMNet[CVPR2023] [13]		71.0	75.4	68.9	62.1	69.4	71.3	76.2	71.3	68.5	71.8
AMFomer (ours)		71.1	**75.9**	**69.7**	**63.7**	**70.1**	**73.2**	**77.8**	73.2	**68.7**	**73.2**
PFENet[TPAMI2020] [14]	Res-101	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
CyCTR[NIPS2021] [7]		67.2	71.1	57.6	59.0	63.7	71.0	75.0	58.5	65.0	67.4
NERTNet[CVPR2022] [43]		65.5	71.8	59.1	58.3	63.7	67.9	73.2	60.1	66.8	67.0
DCAMA[ECCV2022] [9]		65.4	71.4	63.2	58.3	64.6	70.7	73.7	66.8	61.9	68.3
SSP[ECCV2022] [11]		63.2	70.4	68.5	56.3	64.6	70.5	76.4	**79.0**	66.4	73.1
IPMT[NIPS2022] [6]		**71.6**	73.5	58.0	61.2	66.1	**75.3**	76.9	59.6	65.1	69.2
VAT[ECCV2022] [48]		70.0	72.5	64.8	**64.2**	67.9	75.0	75.2	68.4	69.5	72.0
AMFormer (ours)		71.3	**76.7**	**70.7**	63.9	**70.7**	74.4	**78.5**	74.3	**67.2**	**73.6**

Table 3: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on COCO-20^i using the mIoU (%) evaluation metric. Best results are shown in bold.

Method	Backbone	1-shot					5-shot				
		fold0	fold1	fold2	fold3	Mean	fold0	fold1	fold2	fold3	Mean
NERTNet[CVPR2022] [43]	Res-101	38.3	40.4	39.5	38.1	39.1	42.3	44.4	44.2	41.7	43.2
SSP[ECCV2022] [11]		39.1	45.1	42.7	41.2	42.0	47.4	54.5	50.4	49.6	50.2
HSNet[ICCV2022] [8]		37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
DCAMA[ECCV2022] [9]		41.5	46.2	45.2	41.3	43.5	48.0	58.0	54.3	47.1	51.9
IPMT[NIPS2022] [6]		40.5	45.7	44.8	39.3	42.6	45.1	50.3	49.3	46.8	47.9
PFENet[TPAMI2020] [14]	Res-50	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	37.4
RePRI[CVPR2021] [54]		32.0	38.7	32.7	33.1	34.1	39.3	45.4	39.7	41.8	41.6
HSNet[ICCV2021] [8]		36.3	43.1	38.7	38.7	39.2	43.3	51.3	48.2	45.0	46.9
CyCTR[NIPS2021] [7]		38.9	43.0	39.6	39.8	40.3	41.1	48.9	45.2	47.0	45.6
BAM[CVPR2022] [55]		43.4	50.6	47.5	43.4	46.2	49.3	54.2	51.6	49.6	51.2
IPMT[NIPS2022] [6]		41.4	45.1	45.6	40.0	43.0	43.5	49.7	48.7	47.9	47.5
DCAMA[ECCV2022] [9]		41.9	45.1	44.4	41.7	43.3	45.9	50.5	50.7	46.0	48.3
VAT[ECCV2022][48]		39.0	43.8	42.6	39.7	41.3	44.1	51.1	50.2	46.1	47.9
HDMNet[CVPR2023] [13]		43.8	55.3	51.6	49.4	50.0	50.6	61.6	55.7	56.0	56.0
AMFomer (ours)		**44.9**	**55.8**	**52.7**	**50.6**	**51.0**	**52.0**	**61.9**	**57.4**	**57.9**	**57.3**

set it to 0.1 in experiments. Similarly, *D* is not optimized when training *G*. Following [13], we also include a KL (Kullback-Leibler) divergence loss between the correlation maps of adjacent stages to distill informative semantic cues of earlier stages to refine the segmentation quality. Please refer to [13] for more details. The overall loss function for *G* is:

$$\mathcal{L}_g = -\log(\mathbf{FC}(\boldsymbol{\omega}_f^k)) + \mathcal{L}_{KL} + BCE(\mathbf{M}_e, \mathbf{M}_{gt}). \tag{14}$$

Where the BCE is the Cross-entropy loss. After training, only the well-learned *G* is needed in the inference stage.

K-shot inference. When K support samples $\{I_s^k, M_s^k\}_{k=1}^K$ are available, different discriminative target regions $\{\mathbf{M}_{q\tau}^i\}_{i=1}^K$ can be obtained from different support images via Eqn (4) and Eqn (5). We adopt the union of all the activated regions as the initial object region, *i.e.*, $\mathbf{M}_{q\tau} = \mathbf{M}_{q\tau}^1 \cap \mathbf{M}_{q\tau}^2 \cap \cdots \cap \mathbf{M}_{q\tau}^K$. The comprehensive $\mathbf{M}_{q\tau}$ originating from multiple supports is more robust to pose differences, occlusion, etc., than that from a single support image, thereby boosting the performance.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the proposed AMFormer on two commonly used few-shot segmentation benchmarks, Pascal-5^i [2] and COCO-20^i [56]. Pascal-5^i is constructed based on the PASCAL VOC 2012

Table 4: Comparison of FB-IoU and the number of learnable parameters on COCO-20^i.

Backbone	Methods	FB-IoU (%) 1-shot	FB-IoU (%) 5-shot	#learnable params
Res-50	HSNet [8]	60.4	67.0	10.4M
	BAM [55]	68.2	70.7	**2.6M**
	DCAMA [9]	69.5	71.7	47.7M
	HDMNet [6]	72.2	77.7	4.2M
	AMFormer (ours)	**72.9**	**78.8**	5.1M

Table 5: Ablation studies for different components and architecture design in AMFormer.

OM		DM		mIoU	Δ
Single	Multi	w/o $\mathcal{L}_{div}$	w/ $\mathcal{L}_{div}$		
				65.0	0.0
✓				66.5	+1.5
	✓			69.5	+4.5
	✓	✓		69.7	+4.7
	✓		✓	**70.7**	**+5.7**

Table 6: Performance comparison on 1^{st} split of varying the number of local proxies.

#Part	4	6	8	10	12	14
mIoU	70.7	71.0	71.2	**71.3**	71.2	71.2

Table 7: mIoU of different kinds of labels (1^{st} split).

mask	bbox	scribbles
71.3	70.5	70.0

Table 8: Quantitative measurement of intra- and inter-object similarity.

Intra-object	Inter-object
0.015	0.546

dataset [57] and additional annotations from SBD [58]. Following previous works [14, 55], we equally divide the 20 categories into four splits, three of which for training and the rest one for testing. COCO-20^i is a larger benchmark based on MSCOCO dataset [59], the 80 categories of which are partitioned into four splits for cross-validation as down in [14]. We randomly sampled 1000 episodes from the testing split for evaluation. Following common prectices [37, 14, 55, 38], we adopt mean intersection-over-union (mIoU) and foreground-background intersection-over-union (FBIoU) as evaluation metrics.

4.2 Implementation Details

We adopt ResNet-50 and ResNet-101 [60] as the backbone network in our experiment. Following previous works [14, 7, 13], we concatenate the features from the 3^{rd} and 4^{th} blocks of backbone and exploit a 1×1 convolution layer to generate $\mathbf{F}_s$ and $\mathbf{F}_q$ of middle-level to avoid overfitting. The number of attention layers in the *G* and *D* are set to 1 and 2, respectively. We employ the same data augmentation setting as [14]. Since the *G* and *D* of AMFormer are trained alternately, we increase the number of training epochs to 300 for Pascal-5^i and 75 for COCO-20^i, and set the batch sizes as 8 and 4, respectively. AdamW [61] optimizer with poly learning rate decay is used to train both the *G* and *D*. The initial learning rate is set to $1e^{-4}$ and the weight decay is $1e^{-2}$. It should be noted that we adopt a base learner to filter the categories that appear during the training process for a fair comparison with previous works [55, 13]. For more details please refer to the **Supplementary Material**. Our approach is implemented using PyTorch and all experiments are conducted on 4 NVIDIA GeForce RTX 3090 GPUs.

4.3 Comparison with State-of-the-Art Methods

We present the comparison of our method with previous FSS methods on Pascal-5^i and COCO-20^i datasets in Table 2 and Table 3. It can be observed that the proposed AMFormer significantly outperforms previous advanced approaches and achieves new state-of-the-art results under all settings. Specifically, on Pascal-5^i, our AMFormer achieves 70.4% and 73.2% mIoU when using ResNet-50 as the backbone for 1-shot and 5-shot settings, surpassing the most competitive HDMNet [13] by 1.0% and 1.4%, respectively. With ResNet-101 backbone, our method outperforms the previous best results [6] by 2.8% (1-shot) and 1.6% (5-shot). We can obtain additional improvement when using larger backbone network, which demonstrates the scalability of the AMFormer. We attribute this performance gain to better intra-object similarity within the more informative features. As for the more complicated COCO-20^i, our approach also exhibits superior performances compared to other methods, demonstrating its competitiveness on complex data. Besides, Table 4 gives the comparison with previous methods in terms of FBIoU on Pascal-5^i using ResNet-50 backbone. AMFormer also outperforms all of the previous works by a considerable margin. Qualitative results are shown in Figure 4, please refer to **Supplementary Material** for analysis.

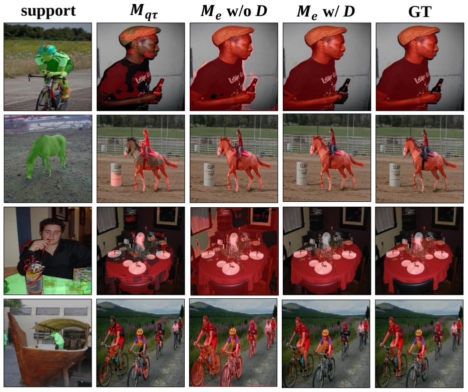


Figure 4: Qualitative results from different stages.

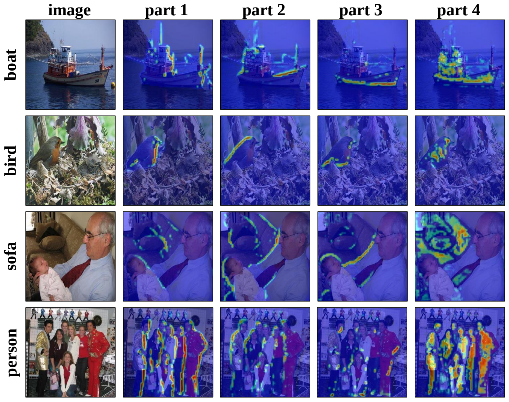


Figure 5: Visualization of regions activated by different local proxies.

4.4 Ablation Study

As shown in Table 5, a series of ablation studies are conducted on the first split of Pascal-5^i with ResNet-101 backbone to analyze each component of the proposed AMFormer. Note that the first line of Table 5 is the result of our ablation baseline. The baseline adopts self-attention within the query features for feature parsing, and cross-attention to aggregate support information into query features as done in [7].

Effectiveness of object mining transformer *G*. We first construct a naive single-scale G as the 2^nd row of Table 5. We can observe a significant performance lift, *i.e.*, 1.5% in mIoU. This improvement demonstrates the effectiveness of the proposed query-centric strategy, which is built upon intra-object similarity and thus is less affected by intra-class diversity. The multi-scale implementation of G further brings a 3.4% improvement in mIoU, and it already achieves a decent performance. It shows the importance of multi-scale feature aggregation in dealing with objects with dramatically changing scales in the query image.

Effectiveness of detail mining transformer *D*. The comparison between the 3^{rd} and the 5^{th} row of Table 5 shows that the combination of D improves the performance by 1.2% mIoU on the basis of G. We attribute this performance gain to the exploration and distinction of local features in the D, which encourages G to pay more attention to these ambiguous regions, *e.g.*, object boundaries. Compared with the pixel-level supervision given by Cross-entropy loss, the proposed part-level adversarial training can incorporate local region context information to guide accurate segmentation. Note that without $\mathcal{L}_{div}$, the performance improvement brought by D drops to 0.2%. This phenomenon is reasonable because in the absence of $\mathcal{L}_{div}$, local proxies tend to degrade to focus on the same region, leaving some details unexplored.

Investigation of the local proxies. We first visualize the region activated by the learnable proxies to analyze what they mainly focus on. As shown in Figure 5, we observe that the highlighted area mainly lies on the boundary of the target and different proxies correspond to different directions. It confirms that the local proxies can well capture the local misalignment of the relatively coarse prediction from G, and then G is optimized to pay more attention to these ambiguous areas under the direction of D. In addition, we investigate the impact of the number of proxies on performance. As reported in Table 6, using more proxies can gradually improve the performance until the number achieves 10. More proxies no longer bring additional improvement. This result is expected, as too many proxies would learn redundant patterns.

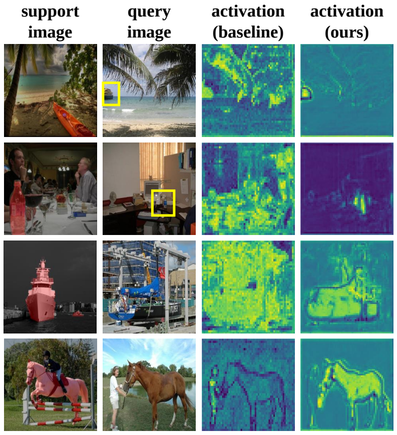


Figure 6: Visualization of the activation of query images..

Discussion on the query-centric FSS paradigm. Since the FSS models are trained in the extremely low data regime with the backbone not optimized, the feature space is not well aligned for novel classes. In this situation, the novel class data distribution usually has low coverage in the feature space, *i.e.*, object features of the same category but different image instances may be far apart. On the contrary, as we can observe in Table 8, pixel features within the same object are closer to each other than those across objects. In a nutshell, intra-object similarity is more reliable than intra-category but inter-object similarity. As shown in Figure 6, we visualize the attention weight of query features between support target and pseudo support, respectively. It shows that support targets tend to mistakenly activate background categories, while the pseudo support can well excavate the full object attribute to intra-object similarity.

Owing to the reduced reliance on support information, query-centric FSS methods could achieve remarkable performance with more weak support annotations. Table 7 shows the results of the proposed AMFormer on the first split of Pascal-5^i with bounding boxes or scribbles as support labels. The promising results demonstrate the feasibility of more general FSS segmentation models.

4.5 Broader Impact and Limitations.

We proposed a novel query-centric FSS paradigm that shifts the research focus from support to query features. This is a new perspective that may inspire the development of more general FSS models that can be adopted in different tasks such as video object segmentation [62–64] or open-vocabulary segmentation [65, 66]. Although our AMFormer achieves remarkable performance, the number of training epochs is larger than some of previous approaches since the G and D in the AMFormer are trained alternately.

5 Conclusion

In this paper, we propose a novel query-centric FSS method, *i.e.*, Adverasial Mining Transformer (AMFormer), which can achieve accurate query segmentation with only rough support guidance. Extensive experimental results demonstrate the superiority of our method. The decent performance with weak support labels also demonstrates the potential of the query-centric FSS paradigm.

6 Acknowledgments

This work was partially supported by the National Defense Basic Scientific Research Program of China (Grant JCKY2021601B013).

References

[1] 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.

[2] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[3] 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.

[4] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8312–8321, 2021.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985.

[11] 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.

[12] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020.

[13] 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.

[14] 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.

[15] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021.

[16] Li Wangkai, Li Zhaoyang, Sun Rui, Mai Huayu, Luo Naisong, Yuan Wang, Pan Yuwen, Xiong Guoxin, Lai Huakai, Xiong Zhiwei, et al. Maunet: Modality-aware anti-ambiguity u-net for multi-modality cell segmentation. In *Competitions in Neural Information Processing Systems*, pages 1–12. PMLR, 2023.

[17] Rui Sun, Huayu Mai, Naisong Luo, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2023.

[18] 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.

[19] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[20] 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.

[21] 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.

[22] 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.

[23] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[24] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[27] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[28] Yifan Zhang, Bo Pang, and Cewu Lu. Semantic segmentation by early region proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1258–1268, 2022.

[29] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

[30] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[32] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021.

[33] Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M Ni, and Lei Zhang. Mp-former: Mask-piloted transformer for image segmentation. *arXiv preprint arXiv:2303.07336*, 2023.

[34] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023.

[35] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2023.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[37] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020.

[38] 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.

[39] Leilei Cao, Yibo Guo, Ye Yuan, and Qiangguo Jin. Prototype as query for few shot semantic segmentation. *arXiv preprint arXiv:2211.14764*, 2022.

[40] Siyu Jiao, Gengwei Zhang, Shant Navasardyan, Ling Chen, Yao Zhao, Yunchao Wei, and Humphrey Shi. Mask matching transformer for few-shot segmentation. *arXiv preprint arXiv:2301.01208*, 2022.

[41] Chunbo Lang, Binfei Tu, Gong Cheng, and Junwei Han. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation. *arXiv preprint arXiv:2204.09903*, 2022.

[42] 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*, pages 763–778. Springer, 2020.

[43] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11573–11582, 2022.

[44] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *arXiv preprint arXiv:2210.06908*, 2022.

[45] Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 362–378. Springer, 2022.

[46] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022.

[47] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 730–746. Springer, 2020.

[48] 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*, pages 108–126. Springer, 2022.

[49] Zhitong Xiong, Haopeng Li, and Xiao Xiang Zhu. Doubly deformable aggregation of covariance matrices for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 133–150. Springer, 2022.

[50] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023.

[51] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

[54] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021.

[55] 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.

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

[57] 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.

[58] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014.

[59] 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.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[62] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021.

[63] Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, and Feng Wu. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1218–1228, 2023.

[64] Rui Sun, Naisong Luo, Yuan Wang, Yuwen Pan, Huayu Mai, Zhe Zhang, and Tianzhu Zhang. 1st place solution for youtubevos challenge 2022: Video object segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[65] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017.

[66] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.

7 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,

$$\mathbf{F}_{q,l} = \mathbf{Down}(\mathcal{F}^{-1}(\mathbf{FeatAgg}(\mathcal{F}(\mathbf{F}_{q,l-1}), \mathcal{F}(\mathbf{F}_{q,l-1})))), \tag{15}$$

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.

8 Detailed Experimental Settings

To achieve a fair comparison with previous methods [55, 13], we adopt the ensemble strategy following BAM [55] 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 [18] is adopted as the base learner in all of our experiments, and we follow the BAM [55] to ensemble the prediction of the base learner and the proposed AMFormer. Our code is available at `https://github.com/Wyxdm/AMNet`

We also conduct an additional ablation experiment to evaluate the influence of the ensemble strategy as shown in Table 9. 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 [6]) without the ensemble strategy.

Table 9: Ablation of ensemble strategy on Pascal-5^i with ResNet-101 backbone and 1-shot setting.

	fold0	fold1	fold2	fold3	mean
IPMT [6]	**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

9 Dataset Settings

We respectively divided Pascal-5^i and COCO-20^i into four splits following [14] for cross-validation. In Table 10 and Table 11, we provide the detailed split settings.

Table 10: 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

10 More Experimental Results

10.1 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

Table 11: 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

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 12 and Table 13. 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.

Table 12: 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

10.2 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 7, we visualize the outcomes of different stages. Procedure (1) can only roughly local the discriminative region of the target (2^{nd} column of Figure 7). 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 7). 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 7. 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 7) 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 $\mathbf{S} \in \mathbb{R}^{H_qW_q \times H_sW_s}$ 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:

$$\mathbf{Act}(i) = \frac{\sum_{j=1}^{H_sW_s} \mathbf{S}(i,j) \cdot [\mathcal{F}(\mathbf{M}_s)(j) > 0]}{\sum_{j=1}^{H_sW_s} [\mathcal{F}(\mathbf{M}_s)(j) > 0]}, \tag{16}$$

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

Table 13: 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

AMFormer, the $\mathbf{S}$ is computed from the pseudo support and the query features. As shown in Figure 8, 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 9 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 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.

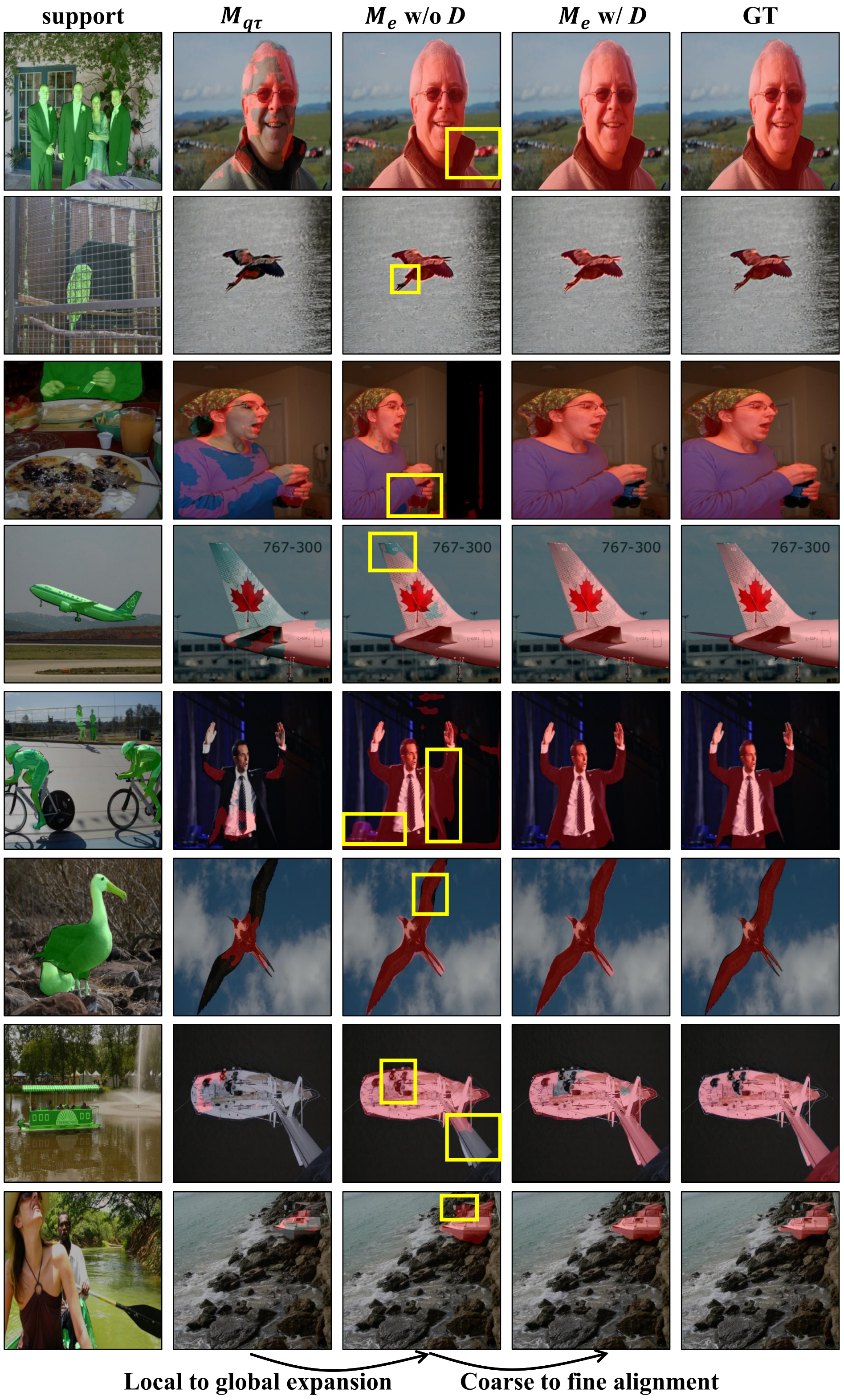


Figure 7: Visualization of the segmentation of the proposed AMFormer at different stages

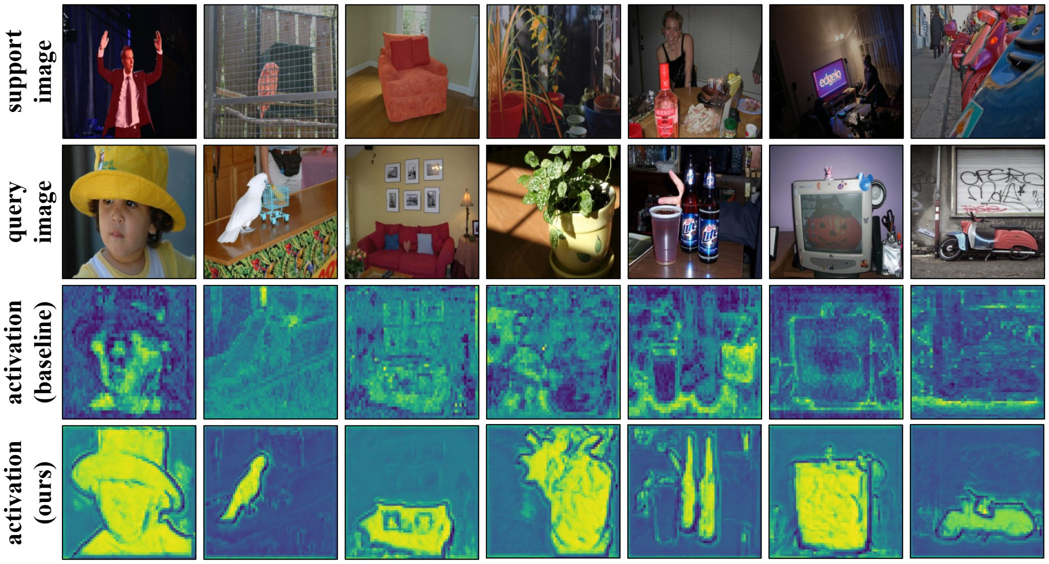


Figure 8: Visualizations of the attention weight of query features between the support target (support-centric baseline) and pseudo support (ours).

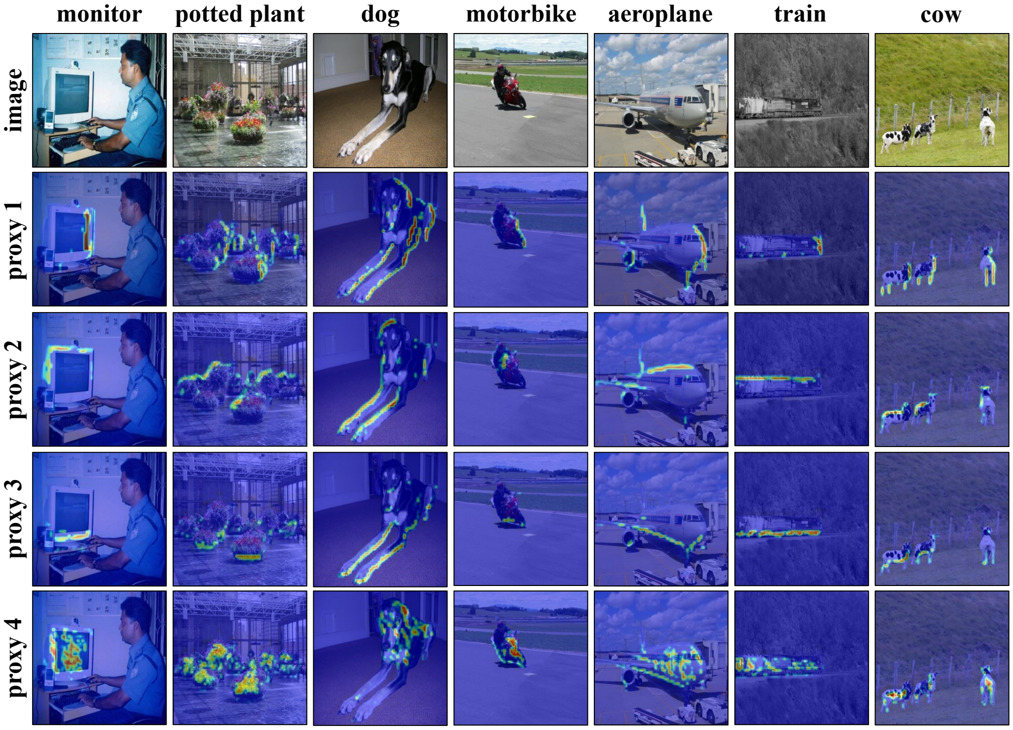


Figure 9: Visualizations of the activated regions of local proxies.