

Inter- and Intra-image Refinement for Few Shot Segmentation

Ourui Fu, Hangzhou He, Kaiwen Li, Xinliang Zhang, Lei Zhu, Shuang Zeng, Zhaoheng Xie, Yanye Lu

Abstract—Deep neural networks for semantic segmentation rely on large-scale annotated datasets, leading to an annotation bottleneck that motivates few shot semantic segmentation (FSS) which aims to generalize to novel classes with minimal labeled exemplars. Most existing FSS methods adopt a prototype-based paradigm, which generates query prior map by extracting masked-area features from support images and then makes predictions guided by the prior map. However, they suffer from two critical limitations induced by inter- and intra-image discrepancies: 1) The intra-class gap between support and query images, caused by single-prototype representation, results in scattered and noisy prior maps; 2) The inter-class interference from visually similar but semantically distinct regions leads to inconsistent support-query feature matching and erroneous predictions. To address these issues, we propose the Inter- and Intra-image Refinement (IIR) model. The model contains an inter-image class activation mapping based method that generates two prototypes for class-consistent region matching, including core discriminative features and local specific features, and yields an accurate and robust prior map. For intra-image refinement, a directional dropout mechanism is introduced to mask inconsistent support-query feature pairs in cross attention, thereby enhancing decoder performance. Extensive experiments demonstrate that IIR achieves state-of-the-art performance on 9 benchmarks, covering standard FSS, part FSS, and cross-domain FSS. Our source code is available at <https://github.com/forypippi/IIR>.

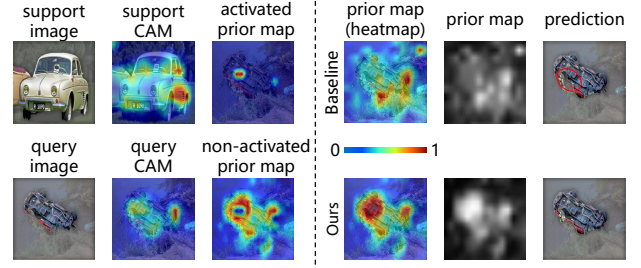
Index Terms—Few shot segmentation, Few shot learning, semantic segmentation.

I. INTRODUCTION

DEEP neural networks have achieved remarkable success in various computer vision tasks, including image classification [1], and semantic segmentation [2], [3]. However, dense prediction tasks such as semantic segmentation typically require large-scale and pixel-level annotated datasets for training, which involve substantial time and labor costs. Furthermore, conventional semantic segmentation models struggle to generalize to novel categories with only a few annotated samples. To address these limitations, researchers have extensively

This work was supported in part by the National Natural Science Foundation of China under grants 82371112, 623B2001; in part by the Science Foundation of Peking University Cancer Hospital (JC202505); in part by the China National Postdoctoral Program for Innovative Talents (BX20250368). (Corresponding authors: Yanye Lu)

Ourui Fu, Hangzhou He, Kaiwen Li, Xinliang Zhang, Lei Zhu, Shuang Zeng, Zhaoheng Xie, Yanye Lu are with the Department of Biomedical Engineering, Peking University, Beijing 100871, China; also with the Institute of Medical Technology, Peking University Health Science Center, Peking University, Beijing 100191, China; also with the National Biomedical Imaging Center, Peking University, Beijing 100871, China. (e-mail: orfu@stu.pku.edu.cn; zhuang@stu.pku.edu.cn; kaiwenli325@gmail.com; zhangxinliang_mit@163.com; zhulei@pku.edu.cn; stevezs@pku.edu.cn; xiezhaoheng@pku.edu.cn; yanye.lu@pku.edu.cn)



(a) Our model's prior map highlights the proper foreground and contains less noise, resulting in more complete predictions within the red circle.



(b) The green and blue circles represent confusing background and foreground regions in the query and support images, respectively.

Fig. 1. (a) Inter-Image Gap: capturing foreground core features and local specific features enables a more precise and robust prior map; (b) Intra-Image Gap: masking inconsistent pixel pairs mitigates false positive and false negative predictions.

investigated few shot semantic segmentation (FSS) approaches [4]–[12].

The predominant FSS methodology employs a prototype-based paradigm, where foreground features extracted from masked support images are aggregated into a class prototype to guide query segmentation via similarity matching. In practice, the feature encoder is often frozen, and only the segmentation decoder is optimized. Consequently, the final model performance critically depends on two core factors: the quality of the **Prior Map** and the efficacy of the **Decoder**. To improve prior map quality, works such as AENet [13] and TBSNet [14] focus on enhancing prototype representation by aggregating foreground and background features or introducing attention mechanisms. For decoder optimization, methods like CyCTR [6] and SCCAN [15] design cross-attention modules or feature fusion blocks to refine the prior map.

Despite these targeted efforts, most existing methods still fail to explicitly account for the inherent variations between and within support/query samples, overlooking the fundamental impact of inter- and intra-sample differences on both the prior map generation and support-query feature alignment. Our study finds that inter-support-query and intra-support/intra-

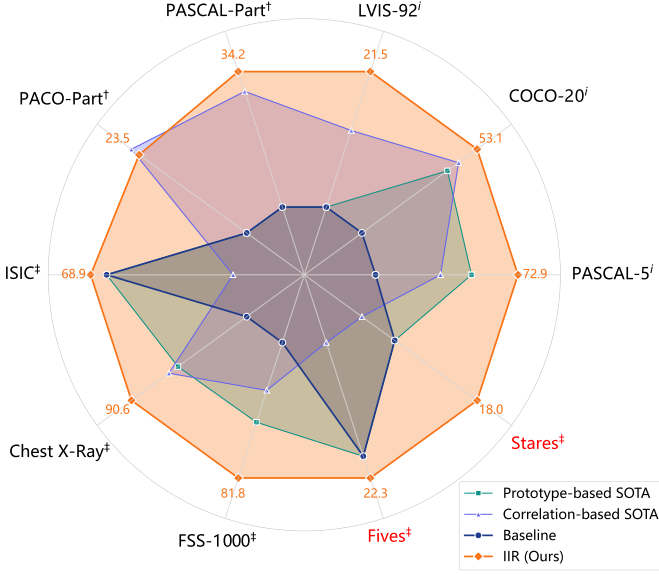


Fig. 2. Our prototype-based method outperforms the baseline across 10 datasets, including standard FSS, object part FSS (with superscript [†]), and cross domain FSS (with superscript [‡]). The new challenging fundus vessel cross domain FSS datasets are highlighted in red.

query image differences lead to inaccurate prior maps and support-query feature mismatch in decoder, respectively. These two issues collectively degrade the generalization ability of FSS models to novel categories, which motivates our research.

Inter-Image Gap, resulting from the use of a single prototype, will cause the prior map to be scattered and inaccurate (e.g., obvious background noise in the prior map of baseline models, as shown in Fig. 1(a) top-right). Specifically, using a single prototype is vulnerable to the internal heterogeneity of the foreground in two aspects:

- 1) Support prototype lacks of diversity: the foreground region of the support image contains various details, like different colors and texture changes. Simply adopting a single masked average pooling for support prototype generation will obscure such fine-grained details, thereby compromising representational capacity for local foreground details in the query image;
- 2) Query image has interference: similarity is calculated between all the query image regions and the same support foreground prototype. Consequently, if there are background regions in the query image that are similar to the foreground prototype, they will be misjudged as foreground.

To mitigate the limitations of generating a single prototype from a single area, we partition the support image and query image into core discriminative and local specific regions and extract prototypes therefrom. By performing similarity matching between each of these prototypes and the corresponding consistent regions in the query image, we generate a more accurate and robust prior map (Fig. 1(a), bottom-right).

Intra-Image Interference stems from the mutual confusion between similar foreground and background within the same image, which in turn is reflected in the decoder cross attention map. This results in the most similar support-query pairs

exhibiting inconsistent masks during training, which is referred to as support-query inconsistency. More specifically, support-query inconsistency originates from two types of Intra-Image Interference, as illustrated in the Fig. 1(b):

- 1) False Positive: the query image itself contains similar background regions (e.g., in color, texture, etc.), and the decoder misclassifies the background as the foreground, as shown in Fig. 1(b) top;
- 2) False Negative: background interference in the query image also causes the decoder to misclassify the foreground of the query image as the background, as shown in Fig. 1(b) bottom.

To tackle the issue of support-query inconsistency in the decoder, we first identify inconsistent support-query pairs via cross attention maps and then apply masking to these pairs during the training phase. By reducing support-query feature inconsistency, our method boosts decoder performance.

Building on the two aforementioned improvements, which target inter-image gaps via multi-prototype extraction and resolve intra-image interference via directional dropout, we further propose the **IIR** model (Inter- and Intra-image Refinement), which simultaneously addresses both inter-image and intra-image discrepancies in FSS. As shown in Fig. 2, comprehensive experiments on 10 datasets across three tasks (standard FSS, object part FSS and cross-domain FSS) validate the effectiveness of our proposed model in enhancing segmentation performance under diverse FSS scenarios.

We summarize our key contributions as follows:

- A CAM-based prior map generation method that addresses the inter-image discrepancy, yielding a more accurate and robust segmentation prior map.
- An inconsistency masking approach that suppresses misleading feature matching pairs and intra-image interference, which contributes to refining the decoder’s feature alignment and prediction performance.
- Our method achieves state-of-the-art performance on 9 benchmarks, including two newly created fundus vessel segmentation datasets, spanning standard, object part, and cross domain FSS tasks.

The paper is organized as follows: section II reviews the related work in recent years; section III elaborates on the proposed method in detail; section IV presents the experimental results and corresponding analysis; section V provides further discussions on our motivations and results; Finally, section VI summarizes the content of this paper, its limitations, and potential future research directions for improvement.

II. RELATED WORK

A. Few Shot Learning (FSL)

Most deep learning methods require a huge amount of data for training. However, collecting a large dataset is a time-consuming and labor-intensive process. To alleviate this burden, researchers focus on few shot learning (FSL) [16]–[19], which aims to improve model generalization when trained with only a limited amount of annotated data. PrototypeNet [16] introduced the prototypical network which is widely used in FSL. Instead of comparing the features from all support

images with the query features, the model calculates the average features of the support images as a class prototypical feature, then the query image is classified by comparing its feature with the class prototype.

B. Few Shot Segmentation (FSS)

Few shot segmentation (FSS) is an application of FSL in semantic segmentation tasks [2], [20], aiming to assign each pixel to a semantic class under limited data constraints. FSS models can be categorized into two main classes: prototypical-based methods and 4D correlation-based methods. Prototypical-based methods [5], [8], [10], [12], [13], [21]–[28] concentrate on comparing the features of the query images with a prototype representing the foreground class in the support set. Instead of pixel-level comparisons, these methods leverage the concept of prototypical representation to guide the segmentation task. However, most methods employ only a single prototype as the representative feature of the foreground class in the support image. Such prototype features are susceptible to gaps between the support and query images, leading to inactivated foreground regions or false activation of background regions, and thus resulting in a prior map with substantial noise. In contrast, we obtain two prototypes by leveraging class cues, which enhances the robustness of the prototype features and yields a prior map with more accurate activations, as shown in section V-A.

4D correlation-based methods [6], [9], [11], [14], [29] can be summarized as follows: it performs pixel-wise matching between support images and query images to form a 4D tensor, and then decodes this tensor via 4D convolution to predict the mask of the query image. CyCTR [6] introduces a cyclic consistency constraint by masking the entire column of cross attention map when meets support→query→support inconsistent. However, our analysis reveals that this column-wise consistency over-masks more than 30% pixels in average, leading to severe information loss. In response to this critical observation, we develop a pixel-wise consistency verification mechanism that masks unidirectional support→query inconsistencies, effectively reducing the over-mask ratio as discussed in section V-B.

In recent years, with the widespread application of SAM [30], [31] in the field of semantic segmentation, a surge of works [32]–[35] has emerged to explore the segmentation capabilities of SAM under data-scarce scenarios. PerSAM [36] focuses on the train-free paradigm: it selects positive and negative sample points with the highest and lowest similarity values from the prior map, and uses the prior map to guide the decoding process of the SAM decoder. This urges the prompt tokens to focus on foreground target regions, thereby enhancing feature aggregation. However, a series of SAM-based methods are all trained or evaluated on the pre-trained SAM backbone, which has been previously trained on the SA-1B large-scale dataset. Its “segment anything” capability is fundamentally built on “having seen everything”, which constitutes a fundamental conflict with the requirement of “generalization to unseen classes” in few shot segmentation.

C. Cross Domain Few Shot Segmentation (CDFSS)

PATNet [37] further extends few shot segmentation to cross domain few shot segmentation (CDFSS). This task rejects the assumption of few shot segmentation that training and testing domains are the same (e.g., in PASCAL-5^t, the training set categories include aeroplane, bicycle, bird, boat, and bottle, while the testing set categories include pottedplant, sheep, sofa, train, and tvmonitor—all belonging to the natural image domain) and replaces the testing domain with domains such as remote sensing domain and medical domain. IFA [38] proposes a fine-tuning strategy, which achieves model finetuning by treating augmented support images as pseudo query images. Several models [39]–[41] address the domain shift issue in cross domain scenarios by designing customized modules or optimization strategies, thereby achieving the adaptation and decoupling of domain-related information and further enhancing the models’ cross domain generalization performance.

III. METHOD

In this section, we first summarize FSS task definition and our IIR model in III-A. Then we present the Prior Map Generation Module (PMGM) in Section III-B, followed by discussion of the Directional Dropout Module (DDM) and the decoder in Section III-C.

A. Problem and Method Overview

In general, our method follows the standard episodic training process of FSS. The dataset is divided into a training set ($\mathcal{D}_{train}$) and a test set ($\mathcal{D}_{test}$). The two sets contain non-overlapping foreground classes. An episode is extracted from the dataset, composed of a support set $\mathcal{S} = \{I^s, M^s\}$ and a query set $\mathcal{Q} = \{I^q, M^q\}$, where I and M represent the image and its binary mask, respectively. During training, the model samples an episode from $\mathcal{D}_{train}$ and learns to map (I^s, M^s, I^q) to M^q . During testing, the model is frozen and episodes are sampled from $\mathcal{D}_{test}$ to predict M^q based on (I^s, M^s, I^q) .

Given a support image-mask pair, the prototype-based paradigm first extracts a single foreground prototype $p^s \in \mathbb{R}^c$ through masked average pooling (MAP). Next, the model generates the prior map $P \in \mathbb{R}^{1 \times h \times w}$ by performing pixel-wise similarity matching between this single prototype and the whole feature map of the query image. This prior map serves as the initial guidance for subsequent segmentation. For further feature refinement, the model employs a cross attention decoder. In this decoder, flatten support features $X^s \in \mathbb{R}^{hw \times c}$ are treated as *keys/values* and flatten query features $X^q \in \mathbb{R}^{hw \times c}$ as *queries* to facilitate feature interaction. Finally, the refined feature representation is fed into a segmentation head to get the final binary mask of the query image $\hat{M}^q$.

The standard prototype paradigm has two critical limitations. First, the single prototype generated by MAP fails to capture fine-grained foreground details, leading to the inter-image gap and resulting in inaccurate prior maps. Second, the cross attention module of the model lacks effective mechanisms to suppress misleading feature correspondences, which causes intra-image inconsistency between support and query

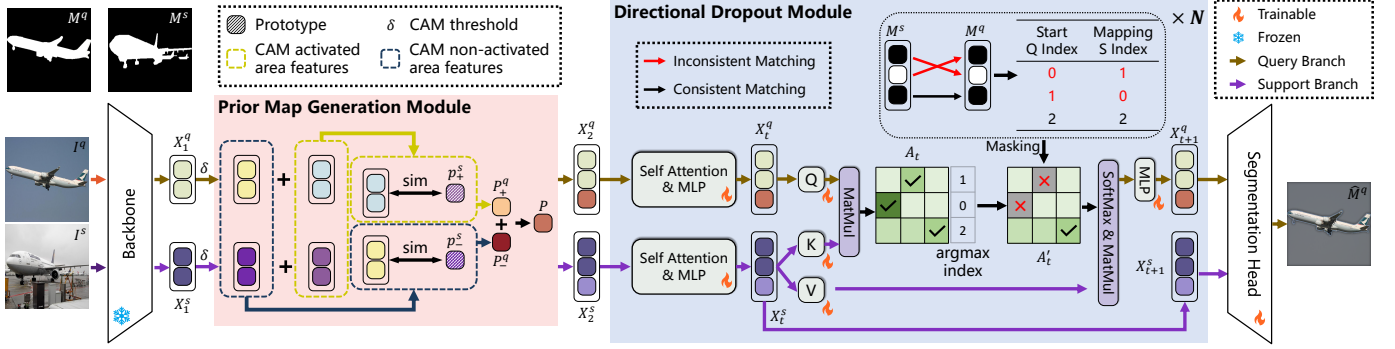


Fig. 3. **Overall architecture of the proposed model.** Support (superscript notation s) and query images (superscript notation q) are firstly processed through a frozen backbone encoder to extract preliminary features X_1^s and X_1^q , respectively. These features, combined with the support mask M^s are subsequently fed into our Prior Map Generation Module (PMGM), which generates a CAM-based support and query prior map. Next, support branch guides query branch training through Directional Dropout Module (DDM). In DDM, the query feature serves as the *Query*, while the support feature acts as the *Key* and *Value* in the cross attention. The black and red line and text in DDM means the consistent and inconsistent support $\rightarrow$ query mapping. Dashed box in DDM is exclusively guide training procedure.

mask alignments. To address these limitations, our method designs two targeted modules with an overview presented in Fig. 3:

- The Prior Map Generation Module (PMGM) aims to resolve the inter-image gap. Different from existing approaches that use MAP to produce a single prototype, the PMGM decomposes the support feature space into two complementary regions via Class Activation Mapping (CAM) heatmap. These regions are CAM activated regions that highlight discriminative cores and CAM non-activated regions that capture local details. From these two regions combined with mask constraints, PMGM extracts three specialized prototypes instead of one. These prototypes include an activated prototype $p_+^s \in \mathbb{R}^c$ which contains foreground core information, a non-activated prototype $p_-^s \in \mathbb{R}^c$ which integrates foreground local details, as well as the original prototype p^s that retains the initial feature representation. By performing region-consistent similarity matching between each prototype and the corresponding CAM-aligned regions in the query image, PMGM outputs a more accurate and robust refined prior map, effectively mitigating the prior map inaccuracy of previous model.
- The Directional Dropout Module (DDM) targets the intra-image inconsistency of the previous model. For the decoder stage, our method adopts a dual-branch interaction mechanism. Support and query features are fed into each decoder branches. The support branch first refines the prior map generated by the PMGM to align with the support’s ground-truth mask. This refined representation then guides the query branch to optimize mask prediction. During this guidance process, the DDM dynamically identifies and masks inconsistent support-query feature pairs in the cross attention map. The DDM prevents intra-image inconsistency from propagating to the final prediction, further enhancing decoder’s performance.

During the testing phase, due to the absence of M_q , the model does not perform masking on the cross attention map; to avoid the inconsistency between the training and testing

distributions, the model dynamically increases the values of unmasked pairs according to the masking ratio during the training phase. For the K-shot setting, we input K support images into the model. PMGM average K prior maps and prototypes while decoders share a unified support branch.

B. Prior Map Generation Module (PMGM)

To address the inter-image feature gap between support and query images and mitigate the inaccuracy of prior maps, the PMGM is specifically designed to refine prior map generation by leveraging class-discriminative region information. Given a specific class o^s from the support image, CAM method generates a heatmap that highlights the regions contributing significantly to classifying the image into that class. PMGM leverages CAM to localize core discriminative (activated) and local specific (non-activated) regions, extracting two additional prototypes from these regions and catching class information to fill the gap between support and query features, thus reducing the noise and resulting in a more accurate prior map.

1) *Prototype Extraction*: To capture both core discriminative regions and local detail regions of the target class, we generate CAM heatmaps for support and query images. Firstly, the query feature $X_1^q \in \mathbb{R}^{c \times h \times w}$ and the support feature $X_1^s \in \mathbb{R}^{c \times h \times w}$ are extracted from the frozen backbone. Since the support and query foreground share the same class, the class o^s can be inferred by the high-level support foreground features X_1^s and support mask $M^s \in \mathbb{R}^{1 \times h \times w}$ with a classification head:

$$\begin{aligned} X_1^s &= \mathcal{F}_{en}(I^s), X_1^q = \mathcal{F}_{en}(I^q), \\ o^s &= \arg \max(\mathcal{F}_{cls}(X_1^s \odot M^s)), \end{aligned} \quad (1)$$

where $\mathcal{F}_{en}$ and $\mathcal{F}_{cls}$ refer to the encoder and classification head, and $\odot$ represents element-wise multiplication. Based on the implicit o^s , support heatmap C^s and query heatmap C^q are generated to highlight the most discriminative object regions.

Based on C^s and C^q , we further explicitly distinguish core discriminative regions and local detail regions. Four binary masks are generated based on a hyperparameter threshold δ

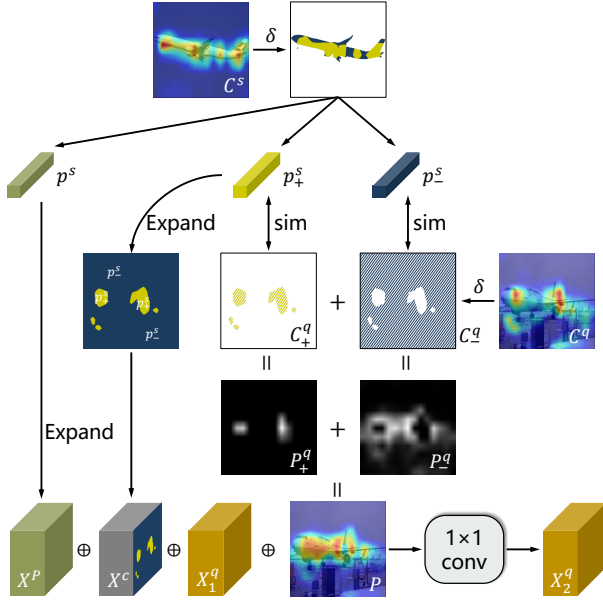


Fig. 4. PMGM workflow: support/query CAM heatmaps are thresholded by δ into C_+^s , C_-^s , C_+^q , C_-^q ; MAP on support foreground and its two sub-regions generates prototypes p_+^s , p_-^s ; similarities between p_+^s/C_+^q and p_-^s/C_-^q yield prior map P ; expanded p_+^s , expanded padded p_+^q/p_-^q , X_1^q , and P are concatenated and fed into 1×1 convolution to obtain new X_2^q .

and the indicator function $\mathbb{1}(\cdot)$ (i.e., $\mathbb{1}(\text{true}) = 1$ and $\mathbb{1}(\text{false}) = 0$). Specifically:

- The support activated mask C_+^s highlights core discriminative regions: $C_+^s = \mathbb{1}(C^s \geq \delta)$;
- The query activated mask C_+^q emphasizes core distinguishing regions: $C_+^q = \mathbb{1}(C^q \geq \delta)$;
- The support non-activated mask C_-^s capture local detail regions: $C_-^s = 1 - C_+^s$;
- The query non-activated mask C_-^q identifies fine-grained local regions: $C_-^q = 1 - C_+^q$.

For each foreground region, PMGM derives three prototypes p_+^s , p_-^s , and p^s using MAP:

$$p_i^s = \frac{\sum_{x,y} X_1^s(x,y) \mathbb{1}(M^s(x,y) \wedge C_i^s(x,y))}{\sum_{x,y} \mathbb{1}(M^s(x,y) \wedge C_i^s(x,y))} \quad (i \in \{+, -\}) \quad (2)$$

$$p^s = \frac{\sum_{x,y} X_1^s(x,y) \mathbb{1}(M^s(x,y))}{\sum_{x,y} \mathbb{1}(M^s(x,y))} \quad (3)$$

where $X_1^s(x,y)$ indexes the support feature vector at spatial locations (x,y) , and $\mathbb{1}(\cdot)$ is an indicator function.

2) *Prior Map Generation*: To generate the prior maps for the core discriminative region (P_+^q for C_+^q) and local detail region (P_-^q for C_-^q) in the query image, we build upon the baseline's single prior map generation pipeline and compute the feature similarity between query region-specific features and the support region prototypes (p_+^s and p_-^s), respectively. Specifically, the foreground probability of each pixel in the query regions is quantified via cosine similarity and min-max normalization:

$$P_i^q(x,y) = \frac{(X_1^q(x,y))^T q_i^s}{\|X^q(x,y)\| \|q_i^s\|} \quad ((x,y) \in C_i^q, i \in \{+, -\}), \quad (4)$$

where $\|\cdot\|$ is the euclidean distance. $P_+^q(x,y)$ and $P_-^q(x,y)$ represent the foreground probability of pixel (x,y) in the area C_+^q and C_-^q , respectively.

Since the core discriminative region C_+^q and local detail region C_-^q of the query image are mutually complementary, we sum their respective prior maps to generate the full-image prior map. The resulting prior map P is denoted as

$$P = P_+^q + P_-^q. \quad (5)$$

Building on the aforementioned prototypes and the full-image prior map P derived above, the model fuses the core discriminative prototype p_+^s and local detail prototype p_-^s into the query feature space. Specifically, p_+^s and p_-^s are filled into their corresponding response areas C_+^q and C_-^q , respectively:

$$X^c = X_+^s \odot C_+^q + X_-^s \odot C_-^q, \quad (6)$$

where X_+^s and X_-^s denote the spatially expanded versions of p_+^s and p_-^s , respectively.

Finally, the new feature X_2^q of the query branch is generated by a 1×1 convolution $\mathcal{F}_{1 \times 1}$ that integrates the original feature, the foreground prototype, the core discriminative prototype, the local detail prototype, and the prior map. The entire pipeline is illustrated in Fig. 4. The support branch's new feature X_2^s is generated in the same way.

C. Directional Dropout Module (DDM)

The DDM is designed to address intra-image interference caused by directionally masking two incorrect support $\rightarrow$ query mapping: query interference (false positive) and support interference (false negative). Our model incorporates support branch decoder to learn to refine the support prior map to support ground truth, guiding query branch decoder to segment query mask. At each decoder block with the index i , the inputs of both query and support branches are flattened as $X_t^q \in \mathbb{R}^{h_q w_q \times c}$ and $X_t^s \in \mathbb{R}^{h_s w_s \times c}$ respectively. Query features X_t^q are denoted as *Query* and support features X_t^s are denoted as *Key* and *Value* in cross attention:

$$A_t = \frac{QK^T}{\sqrt{c}}, Q = X_t^q W_t^Q, K = X_t^s W_t^K, V = X_t^s W_t^V, \quad (7)$$

where $W_t^Q, W_t^K, W_t^V \in \mathbb{R}^{c \times c}$ are the learnable weights to generate *Query*, *Key* and *Value* in the t -th cross attention block.

The mispredicted regions in the query image can be categorized into two types of cases:

- False positive: the GT query mask of pixel i corresponds to the background, whereas the most similar feature from the support pixel j corresponds to the foreground (Fig. 3 DDM, (0, 1));
- False negative: the GT query mask i corresponds to the foreground, whereas the most similar feature from the support features j corresponds to the background (Fig. 3 DDM, (1, 0)).

Since the values in the cross attention map A_t reflect the similarity between the corresponding query features and support features, based on the above mispredicted region attribution, DDM iterates through each query pixel i , retrieves the support pixel j corresponding to the maximum value in row $A_t[i]$,

checks their masks, and disrupts the inconsistent support-query mask pairs ($M_i^q \neq M_j^s$) to prevent the propagation of confusing information.

During the test phase, the DDM reverts to the traditional cross-attention mechanism due to the absence of the query GT mask. To address the potential performance degradation caused by the training-test distribution mismatch, we propose a dynamic dropout strategy. DDM takes the mask frequency as the dynamic dropout rate r , and subsequently multiplies the remaining unmasked elements by $\frac{1}{1-r}$. This ensures consistent feature expectations between the training and testing phases, thereby guaranteeing distribution consistency across the two phases.

In short, DDM applies fine-grained pixel-wise masking, its overall pipeline is illustrated in Algorithm 1.

Algorithm 1: Directional Dropout Module

```

1 Input:
2 cross attention map  $A \in \mathbb{R}^{h_q w_q \times h_s w_s}$ ,
3 flatten support mask  $M^s \in \mathbb{R}^{h_s w_s}$ ,
4 flatten query mask  $M^q \in \mathbb{R}^{h_q w_q}$ 
5 Output: masked cross attention map  $A' \in \mathbb{R}^{h_q w_q \times h_s w_s}$ 
6 Initialize cross attention mask  $A^s \in \mathbb{R}^{h_q w_q \times h_s w_s}$  and
  mask ratio  $r$  with 0;
7 for  $i = 1$  to  $h_q w_q$  do
8    $j = \arg \max_{k \in \{1, 2, \dots, h_s w_s\}} A[i, k]$ ;
9   if  $M^q[i] \neq M^s[j]$  then
10     $A^s[i, j] = -\infty$ ;
11     $r = r + \frac{1}{h_q w_q \times h_s w_s}$ ;
12   end
13 end
14  $A' = A + A^s$ ;
15  $A' = A' \times \frac{1}{1-r}$ ;
```

The query branch $(t+1)$ -th block input X_{t+1}^q is:

$$\begin{aligned} A'_t &= \text{DDM}(A_t, M^s, M^q), \\ X_{t+1}^q &= \text{MLP}(\text{softmax}(A'_t) V). \end{aligned} \quad (8)$$

Finally, the overall loss can be computed using

$$\mathcal{L} = \text{BCE}(\hat{M}^q, M^q) + \frac{1}{N} \sum_{t=1}^N \text{BCE}(\hat{M}_t^q, M^q), \quad (9)$$

where N denotes the number of cross blocks, and $\hat{M}^q$ as well as $\hat{M}_t^q$ represent the final prediction and intermediate prediction of the t -th decoder block, respectively.

IV. EXPERIMENT

We evaluate the performance of our IIR model on 11 datasets across three tasks: **standard few shot segmentation (FSS)**, **object part few shot segmentation (OPFSS)**, and **cross domain few shot segmentation (CDFSS)**. These three tasks are respectively defined as pixel-wise segmentation of unseen novel categories based on a small number of annotated samples, fine-grained pixel-wise segmentation of specific parts of objects, and segmentation of novel categories in new domains. Additionally, we conduct ablation experiments

to validate the effectiveness of the two modules, PMGM and DDM. Experiments demonstrate that our model achieves excellent performance across different tasks and backbones.

A. Datasets

- Datasets includes 9 existing FSS datasets across 3 tasks:
- 1) FSS: PASCAL-5ⁱ [4], COCO-20ⁱ [4], LVIS-92ⁱ [32];
 - 2) OPFSS: PASCAL-Part [32], PACO-Part [32];
 - 3) CDFSS: FSS-1000 [46], Deepglobe [47], ISIC2018 [48], Chest X-Ray [49], [50].

More details are as shown in Appendix Section B.

Existing CDFSS comprises two medical datasets (ISIC-2018 for skin diseases and Chest X-Ray for chest diseases), failing to fully reflect CDFSS's advantages. Their mask regions have clear boundaries, with annotation merely requiring boundary delineation, undermining the necessity of utilizing FSS (columns 2 and 3 of Fig. 8). In contrast, fundus blood vessel segmentation faces small public datasets and high annotation difficulty: image contains dense small blood vessels, and mask annotation demands precise vascular trajectory delineation. Each image takes hours and substantial labor for repeated verification. To better evaluate current CDFSS methods on complex topological and real-world medical tasks, we build two fundus blood vessel CDFSS benchmarks (Fives, Stare) based on public datasets; see Appendix Section B for details.

B. Implementation Details

For PASCAL-5ⁱ and COCO-20ⁱ, our method adopts BAM [43] as its baseline. For LVIS-92ⁱ, PASCAL-part and PACO-Part, we follow the setting of Matcher [32]. For CDFSS datasets, our model takes IFA [38] as its baseline. More details are shown in Appendix section C. We adopted mean intersection over union (mIoU) and foreground-background intersection over union (FB-IoU) as evaluation metrics. All tests of the proposed model on all datasets are conducted using the PyTorch and run on an NVIDIA RTX 4090 GPU.

C. Comparison With State-of-the-Arts

1) *FSS*: In Table I, Table II and Table III, we compare our method with other SOTA approaches on PASCAL-5ⁱ, COCO-20ⁱ and LVIS-92ⁱ, respectively.

For PASCAL-5ⁱ, our IIR achieves 72.9% mIoU with ResNet-50 backbone and 73.1% mIoU with ResNet-101 backbone for 1-shot segmentation, representing improvements of 4.6% and 4.5% in mIoU over the baseline, respectively. It outperforms previous SOTA results by 1.5% and 1.2%, respectively. For 5-shot segmentation, our method attains mIoU gains of 3.8% and 3.1% compared to the baseline when using ResNet-50 and ResNet-101 backbones, respectively, while surpassing SOTA methods by 1.1% and 0.8% in mIoU.

For COCO-20ⁱ results in Table II, our method achieves 53.1% mIoU with ResNet-50 backbone and 54.8% mIoU with ResNet-101 backbone for 1-shot segmentation, representing improvements of 6.2% and 6.3% in mIoU over the baseline, respectively. For 5-shot segmentation, our IIR attains mIoU

TABLE I
QUANTITATIVE COMPARISON RESULTS ON THE PASCAL-5ⁱ DATASET. THE BEST AND THE SECOND BEST mIoU (IN %) ARE HIGHLIGHTED WITH **BOLD** AND UNDERLINE, RESPECTIVELY.

Methods	1-shot						5-shot					
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
ResNet 50 Backbone												
CyCTR ₂₀₂₁ [6]	65.7	71.0	59.5	59.7	64.0	-	69.3	73.5	63.8	63.5	67.5	-
PFENet++ ₂₀₂₄ [25]	63.3	71.0	65.9	59.6	64.9	76.8	66.1	75.0	<u>74.1</u>	64.3	69.9	81.1
TBS ₂₀₂₄ [14]	68.5	72.0	63.8	59.5	65.9	77.7	72.3	74.1	68.4	67.2	70.5	81.3
CGMGM ₂₀₂₄ [5]	71.1	75.0	69.6	63.7	69.9	80.5	71.8	<u>78.9</u>	69.1	68.6	72.1	83.1
ABCB ₂₀₂₄ [22]	<u>72.9</u>	<u>76.0</u>	69.5	64.0	70.6	-	<u>74.4</u>	78.0	73.9	68.3	73.6	-
AENet ₂₀₂₄ [13]	71.3	<u>75.9</u>	68.6	65.4	70.3	-	73.9	77.8	73.3	<u>72.0</u>	74.2	-
HMNet ₂₀₂₄ [42]	72.2	75.4	70.0	63.9	70.4	<u>81.6</u>	74.2	77.3	<u>74.1</u>	70.9	74.1	<u>84.4</u>
OCNet ₂₀₂₅ [26]	73.5	75.9	<u>71.1</u>	<u>64.9</u>	71.4	-	75.9	77.1	<u>74.1</u>	70.9	<u>74.5</u>	-
BAM ₂₀₂₃ [43]	69.2	74.7	67.8	61.7	68.3	-	71.8	75.7	72.0	67.5	71.8	-
IIR(Ours)	72.0	77.5	74.1	68.1	72.9	83.3	73.6	79.8	76.2	72.6	75.6	85.1
ResNet 101 Backbone												
CyCTR ₂₀₂₁ [6]	67.2	71.1	57.6	59.0	63.7	-	71.0	75.0	58.5	65.0	67.4	-
DBMNet ₂₀₂₄ [23]	64.2	72.1	64.8	58.9	65.0	-	68.6	73.3	69.0	64.2	68.8	-
PFENet++ ₂₀₂₄ [25]	63.1	72.4	63.4	62.2	65.3	75.5	67.2	76.1	75.5	67.2	71.5	82.7
ABCB ₂₀₂₄ [22]	73.0	<u>76.0</u>	<u>69.7</u>	<u>69.2</u>	<u>71.9</u>	-	74.8	<u>78.5</u>	73.6	<u>72.6</u>	<u>74.8</u>	-
BAM ₂₀₂₃ [43]	69.9	75.4	67.1	62.1	68.6	80.2	72.6	77.1	70.7	69.8	72.5	84.1
IIR(Ours)	<u>72.8</u>	77.9	71.5	70.0	73.1	82.1	<u>74.1</u>	79.6	<u>75.3</u>	73.3	75.6	84.7

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE COCO-20ⁱ DATASET. THE BEST AND THE SECOND BEST mIoU (IN %) ARE HIGHLIGHTED WITH **BOLD** AND UNDERLINE, RESPECTIVELY.

Methods	1-shot						5-shot					
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
ResNet 50 Backbone												
CyCTR ₂₀₂₁ [6]	38.9	43.0	39.6	39.8	40.3	-	41.1	48.9	45.2	47.0	45.6	-
CGMGM ₂₀₂₄ [5]	<u>47.1</u>	49.3	48.8	44.4	47.4	-	50.3	54.6	51.3	51.8	52.0	-
ABCB ₂₀₂₄ [22]	44.2	54.0	52.1	49.8	50.0	-	50.5	59.1	57.0	53.6	55.1	-
HMNet ₂₀₂₄ [42]	45.5	58.7	52.9	51.4	<u>52.1</u>	<u>74.5</u>	<u>53.4</u>	<u>64.6</u>	60.8	<u>56.8</u>	<u>58.9</u>	77.6
OCNet ₂₀₂₅ [26]	45.9	56.9	52.9	50.4	51.5	73.7	52.7	63.1	57.4	54.8	57.0	76.8
BAM ₂₀₂₃ [43]	43.8	51.4	47.9	44.5	46.9	-	49.8	55.4	52.3	50.2	51.9	-
IIR(Ours)	49.7	<u>58.1</u>	52.3	51.4	53.1	75.4	53.7	65.2	<u>59.6</u>	57.3	59.0	<u>77.3</u>
ResNet 101 Backbone												
HPA ₂₀₂₃ [28]	43.2	50.5	46.2	46.3	48.8	-	49.4	58.4	52.5	50.9	52.8	-
DBMNet ₂₀₂₄ [23]	41.8	45.6	43.2	41.3	43.0	-	46.7	50.5	48.8	44.7	47.7	-
ABCB ₂₀₂₄ [22]	46.0	56.3	<u>54.3</u>	51.3	51.5	-	51.6	63.5	62.8	57.2	58.8	-
BAM ₂₀₂₃ [43]	45.2	55.1	48.7	45.0	48.5	-	48.3	58.4	52.7	51.4	52.7	-
IIR(Ours)	50.4	58.6	56.1	54.0	54.8	74.8	52.8	64.7	63.0	57.4	59.5	77.9

TABLE III
QUANTITATIVE COMPARISON RESULTS ON THE LVIS-92ⁱ DATASET IN TERMS OF mIoU (%) UNDER 1-SHOT AND 5-SHOT SETTINGS.

Method	Type	1-shot	5-shot
HSNet ₂₀₂₁ [9]	correlation-based	17.4	22.9
VAT ₂₀₂₂ [29]	correlation-based	18.5	22.7
Painter ₂₀₂₃ [44]	VLM-based	10.5	10.9
SegGPT ₂₀₂₃ [45]	VLM-based	18.6	25.4
baseline	Prototype-based	14.6	20.9
IIR (Ours)	Prototype-based	21.5	26.3

gains of 7.1% and 6.8% compared to the baseline when using ResNet-50 and ResNet-101 backbones, respectively, while surpassing SOTA methods by 0.1% and 0.7% in mIoU.

For LVIS-92ⁱ results in Table III, compared with the baseline, our model achieves improvements of 6.9% and 5.4% under the 1-shot and 5-shot settings, respectively. In comparison with correlation-based methods, we obtain gains of 3.0% and 3.4% under the 1-shot and 5-shot scenarios, respectively.

Additionally, we list various SAM-based methods, among which Matcher [32] exhibit stronger performance. This may be attributed to the fact that the SAM-1B dataset [30] contains a diverse range of image-mask pairs, thereby endowing methods based on it with robust FSS capabilities.

Based on the performance across these FSS datasets, we observe that 5-shot settings achieve slightly inferior performance compared to 1-shot settings under all configurations. We attribute this phenomenon to the fact that 5-shot settings integrate five prior maps, resulting in higher robustness of the prior map than that in 1-shot settings. As the PMGM is designed to enhance the accuracy and robustness of the prior map, the performance improvement under 5-shot settings is relatively limited.

2) *OPFSS*: In Table IV, we present additional quantitative results on the PASCAL-Part and PACO-Part datasets. Among these results, our proposed IIR model outperforms the baseline method by a significant margin in all folds, providing a robust baseline result for future research. Specifically, our IIR surpasses the baseline by 4.1% on PASCAL-Part and

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON THE PASCAL-PART AND PACO-PART DATASET IN TERMS OF mIoU (%).

Method	Type	PASCAL-Part					PACO-Part				
		animal	indoor	person	vehicles	mean	Fold-0	Fold-1	Fold-2	Fold-3	mean
HSNet ₂₀₂₁ [9]	correlation-based	21.2	53.0	20.2	35.1	32.4	22.0	22.9	26.0	23.1	23.5
VAT ₂₀₂₂ [29]	correlation-based	21.5	55.9	20.7	36.1	33.6	13.7	12.5	15.0	15.1	14.1
Painter ₂₀₂₃ [44]	VLM-based	20.2	49.5	17.6	34.4	30.4	13.9	12.6	14.8	12.7	13.5
SegGPT ₂₀₂₃ [45]	VLM-based	22.8	50.9	31.3	38.0	35.8	19.4	20.5	23.8	21.2	21.2
baseline	Prototype-based	18.8	50.6	21.1	29.8	30.1	19.8	20.8	24.4	23.1	22.0
IIR (Ours)	Prototype-based	24.5	54.2	25.8	32.1	34.2	21.7	22.3	25.7	23.9	23.4

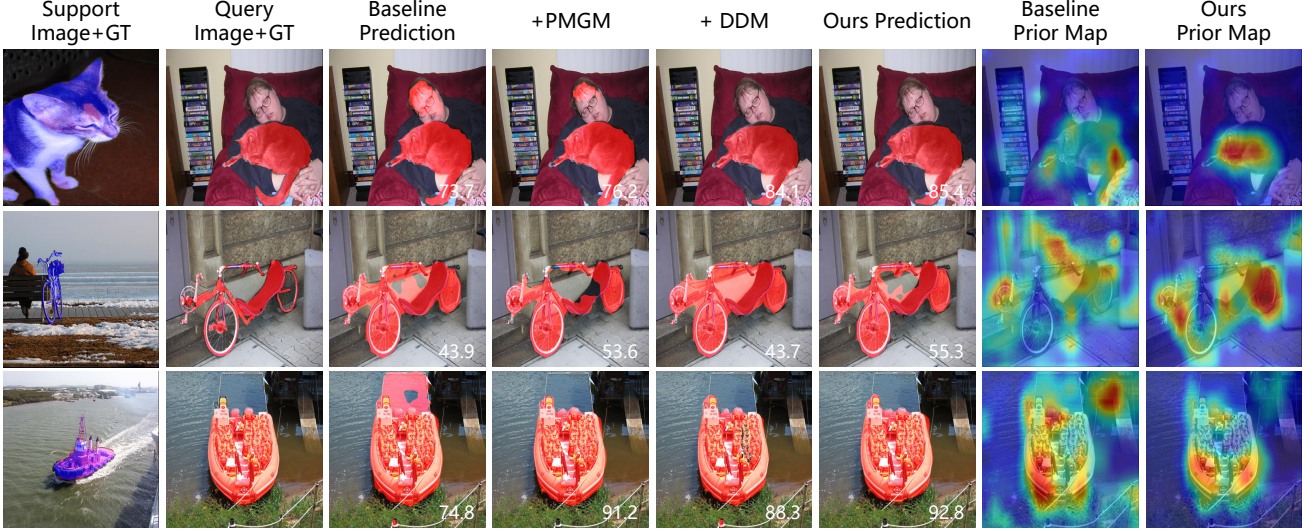


Fig. 5. Visualization on the PASCAL-5ⁱ dataset. The first row is segmentation in similar backgrounds, the second row addresses segmentation within small foregrounds and the third row focuses on segmentation in salient foregrounds. The third column is the prediction of our baseline model BAM [43]. The last two columns show our PMGM generates more accurate foreground cues. More visualization results are presented in the supplementary materials.

by 1.4% on PACO-Part. Furthermore, our model outperforms correlation-based methods in most cases, and even exceeds Painter and SegGPT.

3) *CDFSS*: Table V presents the quantitative results of our IIR model for cross domain few shot segmentation under the 1-shot and 5-shot settings. It can be observed from the results that our model achieves significant performance improvements on most datasets. Compared with the baseline IFA: on the ISIC dataset, the improvements are 2.6% and 13.4% under the 1-shot and 5-shot settings, respectively; on the Chest X-ray dataset, the improvements are 16.6% and 17.4% under the 1-shot and 5-shot settings, respectively; on the fives dataset, the improvements are 1.7% and 1.5% under the 1-shot and 5-shot settings, respectively; on the stare dataset, the improvements are 4.0% and 4.8% under the 1-shot and 5-shot settings, respectively; on average, the improvements over the baseline are 3.8% and 6.1% under the 1-shot and 5-shot settings, respectively. However, our IIR method exhibits a significant performance drop on the Deepglobe dataset—particularly a 4.4% decrease under the 1-shot setting. Through visualization analysis, we attribute this phenomenon to the significant domain gap between the training and test data: our model is trained on natural images, while Deepglobe is a remote sensing image dataset. The visual features and semantic patterns for prior map generation differ drastically between natural and

remote sensing domains. This discrepancy weakens CAM’s ability to capture effective discriminative features for remote sensing targets, leading to reduced quality of the prior map and ultimately causing the performance drop on Deepglobe. Furthermore, all CDFSS methods achieve relatively low performance on the fundus vessel segmentation dataset, which indicates that existing CDFSS approaches fail to accomplish the segmentation of complex topological structures. In this regard, FSS algorithms still require further improvement.

D. Qualitative Results

To more intuitively demonstrate the effectiveness of our algorithm, Fig. 5 presents the final prediction performance and prior map prompting capability of a single support-query image pair on the PASCAL-5ⁱ dataset under the ResNet50 1-shot setting, from the baseline to the IIR model with the incremental addition of each module. As illustrated in Fig. 5, our model consistently delivers more accurate segmentation outcomes compared with the baseline across diverse challenging scenarios including query images with mixed foreground-background (row 1), complex topological structures (row 2) and significant dissimilarity between support and query foregrounds (row 3), by effectively concentrating the prior map on the query’s foreground. Further demonstrations are provided in Appendix Fig. S3.

TABLE V
QUANTITATIVE COMPARISON RESULTS ON THE CROSS DOMAIN FEW SHOT SEGMENTATION DATASETS. THE BEST AND THE SECOND BEST MIOU (IN %) ARE HIGHLIGHTED WITH **BOLD** AND *underline*, RESPECTIVELY.

Source Domain: Pascal VOC 2012 → Target Domain: Below														
method	Deepglobe		ISIC		Chest X-Ray		FSS-1000		Fives		Stare		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PATNet ₂₀₂₂ [37]	37.9	43.0	41.2	53.6	66.6	70.2	78.6	81.2	11.7	11.9	10.9	10.9	41.2	45.1
ABCD _{FSS} ₂₀₂₄ [51]	42.6	49.0	45.7	53.3	79.8	81.4	74.6	76.2	11.3	12.5	12.4	12.8	44.4	47.5
APM ₂₀₂₄ [39]	40.9	44.9	41.7	51.2	78.3	82.8	79.3	81.8	10.0	10.1	10.4	10.5	43.4	46.9
IFA ₂₀₂₄ [38]	50.6	58.8	66.3	69.8	74.0	74.6	80.1	82.4	20.6	21.0	14.0	14.0	50.9	53.4
IIR (Ours)	46.2	58.6	68.9	83.2	90.6	92.0	81.8	81.6	22.3	22.5	18.0	18.8	54.7	59.5

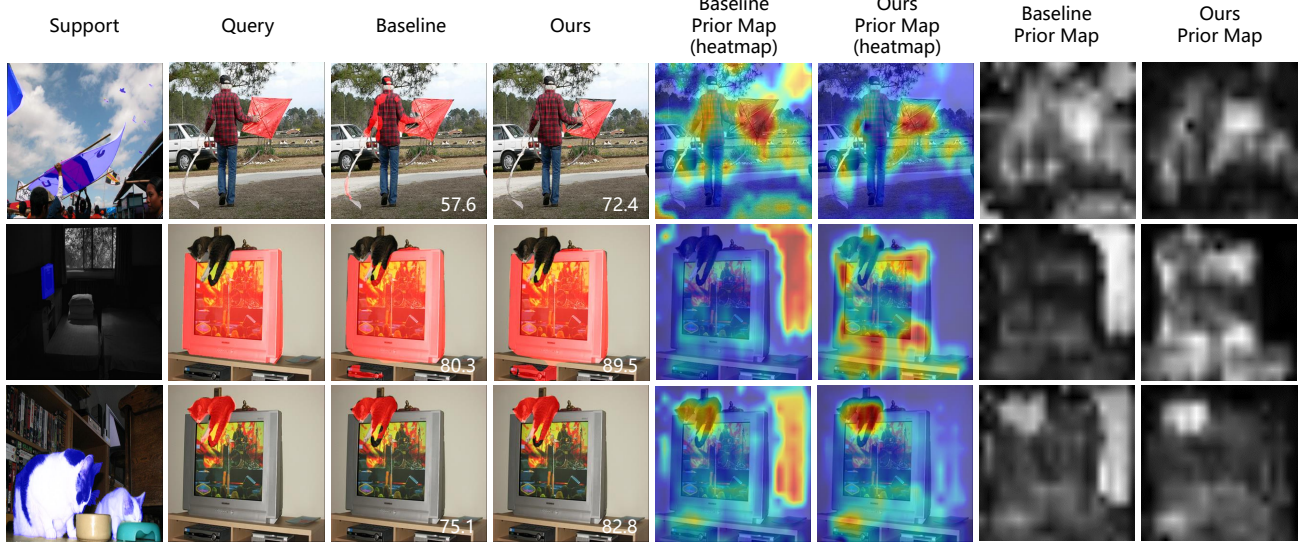


Fig. 6. Visualization on COCO-20ⁱ dataset. Last two columns is the prior map on grayscale images. Our prior map outperforms baseline in foreground focusing and noise elimination. Last two rows have same query’s foreground distinction and we accurately targets corresponding foregrounds for better segmentation.

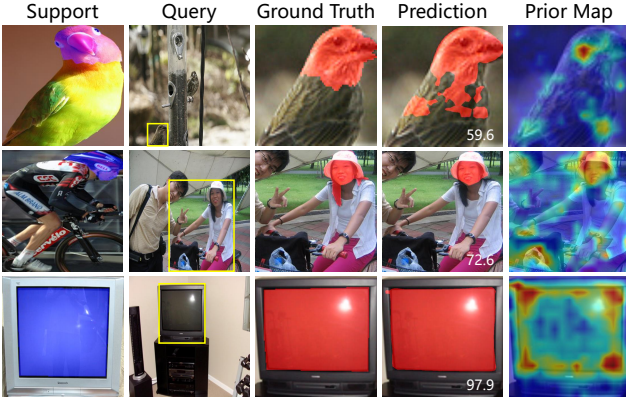


Fig. 7. Visualization on PASCAL-Part dataset. The yellow box in the query image (second column) serves as model input (third column).

We visualize the performance of our model on the COCO-20ⁱ dataset in Fig. 6. The last two columns present the effect of the prior map on grayscale images, from which it can be clearly observed that our prior map exhibits stronger foreground focusing ability and noise elimination ability compared with the baseline prior map. Additionally, in the last two rows, we compare the ability to distinguish different foregrounds for the same query image; the results show that our prior map can well focus on the corresponding foreground targets and

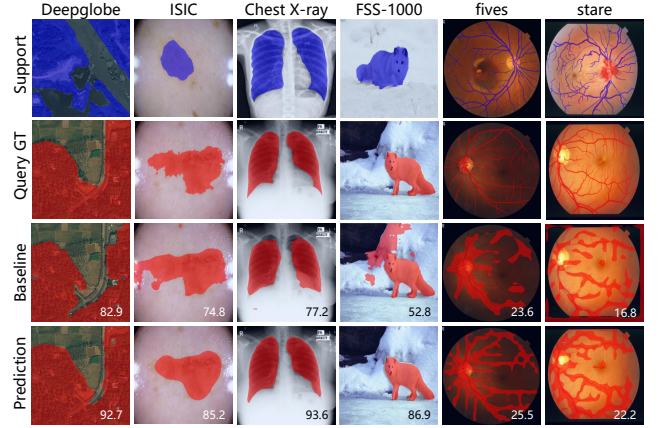


Fig. 8. Visualization on cross domain datasets, including remote sensing domain, medical domain and natural domain.

ultimately achieve better segmentation results.

We also conducted corresponding visualizations on the LVIS-92ⁱ dataset, as shown in Appendix Fig. S4. The visualization results demonstrate that our method exhibits excellent segmentation performance for multiple targets, size-varying targets, and irregularly shaped targets. Fig. 7 and Appendix Fig. S5 presents several representative segmentation results to illustrate typical scenarios in the object part few shot

TABLE VI
BCE BETWEEN THE PRIOR MAP AND THE GROUND TRUTH. LOWER BCE INDICATES GREATER SIMILARITY BETWEEN THE PRIOR MAP AND THE GROUND TRUTH. THE SUBSCRIPT INDEX IS THE STANDARD DEVIATION ACROSS TEN RANDOM VALIDATION SEED.

	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Baseline	0.2875 _{0.0021}	0.3550 _{0.0149}	0.3908 _{0.0335}	0.3698 _{0.0284}	0.3508 _{0.0089}
IIR	0.2457 _{0.0018}	0.2848 _{0.0023}	0.3029 _{0.0023}	0.2902 _{0.0018}	0.2809 _{0.0007}

TABLE VII
ABLATION STUDIES. ‘COL -inf’ MEANS MASKING THE WHOLE COLUMN TO -inf WHEN MEETING INCONSISTENCY.

PMGM	supp branch	CyCTR	DDM		mIoU
			col -inf	cell -inf	
					70.5
	✓			✓	71.4
✓					71.3
✓	✓				71.6
✓	✓	✓			71.8
✓	✓		✓		71.8
✓	✓			✓	72.9

segmentation task, where the model is required to perform segmentation on cropped images as shown in the second column of figures. The visualization results across cross domain datasets are shown in Fig. 8.

E. Ablation Study

We conduct ablation studies on the PASCAL-5ⁱ dataset with the ResNet50 backbone to evaluate the effectiveness of PMGM and DDM. Table VII reveals incremental impact through row-wise comparisons. The baseline (BAM with equivalent decoders) presents mIoU results of 70.5% (row 1). The mIoU improves 2.2% by merely adding decoder number, as shown in Fig. S1, demonstrating our architecture’s inherent performance enhancement capabilities. According to the third row of Table VII, the PMGM module, when used alone, improves the performance to 71.3%. The DDM module, when used alone (as shown in the second row), enhances the performance by 0.9%.

To further investigate and compare the performance details of the DDM and similar models, we first test the performance when only the support branch is introduced, which leads to a 0.3% improvement (row 4 versus row 3). This result indicates that introducing support information can convey certain effective information to aid in segmentation. Secondly, we introduce CyCTR and DDM for comparison. Compared with the baseline (fourth row), the introduction of CyCTR improves the performance by 0.2% (fifth row). The last row and the fifth row of Table VII, together with Table VIII, present a performance comparison between our DDM and CyCTR. Our DDM can reduce the number of support-query inconsistent pairs, thereby achieving a 1.1% improvement in mIoU. The reason lies in that our DDM can directionally mask only inconsistent support-query pairs, avoiding excessive column-wise masking. We also conducted comparative tests on the internal parameter settings of DDM. Specifically, we compared the effects of the dropout strategy and the train-test consistent strategy of CyCTR, as shown in the 5-th and 6-th rows of Table VII. The results indicate that the dropout

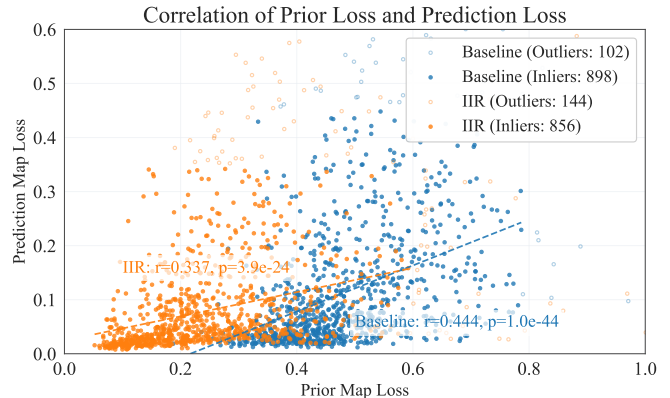


Fig. 9. The BCE loss comparison on the PASCAL-5ⁱ datasets. Blue and orange points denote the baseline and our model’s samples, respectively. The results indicate that the BCE loss of the prior map and that of the prediction map exhibit a correlation and our model is far more accurate than the Baseline model, thereby improving the model performance.

strategy can effectively align the gap between training and testing, and there is little difference in performance between this strategy and CyCTR’s train-test consistent strategy. By comparing the 6th row with the last row, it can be seen that our masking strategy significantly reduces the mask ratio of the cross attention module (by $1 - \frac{1}{900} = 99.89\%$), thereby avoiding the drawback of information loss caused by excessive masking in CyCTR and improving the mIoU by 1.1%.

Additional ablation experiments regarding the Decoder number, CAM method, and CAM threshold are provided in Appendix Section D.

V. DISCUSSION

A. Prior Map

Inter-Image Gap with the use of a single prototype causes the prior map to be scattered and inaccurate. To quantify this issue, we calculated two sets of Binary Cross Entropy (BCE) losses for the baseline model [43]: one between the prior map and ground truth (GT) results, and the other between the final prediction map and GT results. The Pearson Correlation Coefficient r is used to quantify the direction and strength of the linear correlation between two variables. The correlation coefficient r in our experiment is 0.444, confirming that *the lower the loss of the prior map, the lower the loss of the final prediction map*, as illustrated in Fig. 9.

Additionally, we performed quantitative analysis. Table VI demonstrates that the BCE loss of the prior map for IIR decreases by 19.93% compared with the baseline, with a smaller standard deviation (the BCE loss of IIR is 0.2809 ± 0.0007 , while that of the baseline is 0.3508 ± 0.00089). This indicates

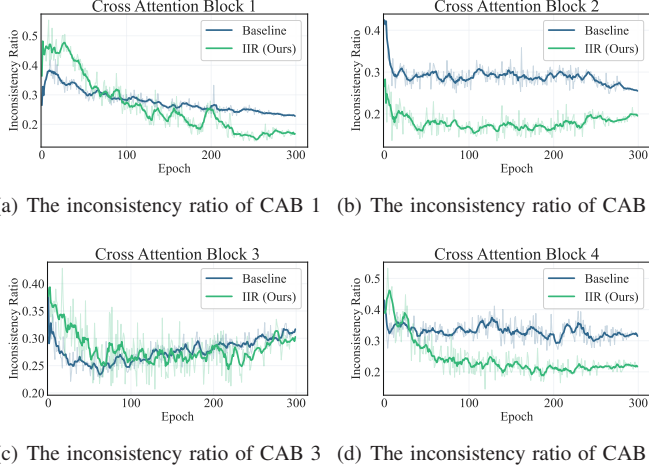


Fig. 10. Comparison of the inconsistency ratio of four Cross Attention Blocks (CAB) during the training process.

that the dual prototypes of “foreground core features + local specific features”, generated by PMGM via CAM, effectively mitigate the background noise and foreground dispersion issues of the prior map. The prior map of IIR is more focused on foreground regions, thereby improves the model performance.

B. Directional Dropout

Intra-Image Interference causes support-query inconsistency during training. The performance of the cross attention decoder can be evaluated by the inconsistency ratio between support and query features, which is defined as the proportion of support-query pairs in the support-query cross attention map where query feature i and its most similar support feature j do not satisfy $M_i = M_j$. In our model, the inconsistency ratio is mask ratio r times $h_q \times w_q$. The mask pairs of query feature i and its most similar support feature j should be consistent, meaning that *the inconsistency ratio should decrease during training*, as shown in Fig. 10. Comparison between our method and the baseline reveals that our inconsistency ratio is lower than that of the baseline, demonstrating that our model can more effectively suppress the support-query feature inconsistency.

We further analyzed the inconsistency ratios of different methods. Table VIII demonstrates that the average support-query inconsistency ratio of DDM is $(54.31 + 16.8 + 19.54 + 30.27 + 21.68)/5 = 28.52\%$, which represents a reduction of 2.4% compared with CyCTR’s $(48.87 + 27.76 + 22.26 + 32.36 + 23.38)/5 = 30.93\%$. One potential reason is that, compared to CyCTR’s “column masking” strategy, DDM eliminates approximately $30.93\% \times 899/900 = 30.90\%$ of information loss. This verifies that DDM’s “directional masked inconsistent feature pair” strategy can accurately suppress false positives and false negatives, rather than excessively masking valid information as CyCTR does. Consequently, DDM alone contributes a 0.9% mIoU improvement, and achieves further performance enhancement when synergized with PMGM.

TABLE VIII
SUPPORT-QUERY INCONSISTENCY RATIO OF BASELINE, CYCTR AND DDM ACROSS CROSS ATTENTION BLOCKS (CAB) AND FOLDS DURING TESTING, IN %.

		Fold-0	Fold-1	Fold-2	Fold-3	Mean
Baseline	CAB 0	29.37	27.84	27.95	27.78	28.24
	CAB 1	29.78	26.38	17.71	17.38	22.81
	CAB 2	20.98	29.51	39.85	11.60	25.49
	CAB 3	38.27	28.31	21.14	38.98	31.68
	CAB 4	37.71	22.82	55.58	9.83	31.49
CyCTR	CAB 0	22.37	24.57	91.13	57.39	48.87
	CAB 1	44.51	24.12	24.17	18.23	27.76
	CAB 2	19.20	23.32	21.26	25.27	22.26
	CAB 3	20.22	23.62	24.54	61.05	32.36
	CAB 4	20.59	23.33	24.59	25.01	23.38
IIR	CAB 0	54.90	48.20	58.91	55.24	54.31
	CAB 1	7.35	11.46	31.09	17.30	16.80
	CAB 2	14.57	24.44	14.27	24.88	19.54
	CAB 3	23.56	30.74	55.55	11.22	30.27
	CAB 4	32.02	31.53	12.20	10.98	21.68

C. Limitation

The first key improvement of IIR is extending a single prototype to multiple ones, thereby more accurately comparing the support foreground with query images and obtaining a more precise query prior map. CAM is a method that is easy to implement and proven to excel at localizing the most salient foreground regions in images, and it has successfully validated the effectiveness of our method’s theory. However, CAM also introduces the burden of additionally training a classification head as well as computational overhead from backpropagation; future work may consider developing more efficient and reliable methods for foreground region localization. Besides, existing FSS methods, including our proposed IIR, fail to achieve foreground segmentation with clinical utility when confront with real-world medical scenarios such as fundus vessels segmentation by scarce data and complex foreground topological structures. Future research needs to further explore approaches to enhance the performance of FSS methods on such targets.

VI. CONCLUSION

In this work, we present IIR model, which refines the prior map and cross attention map generation processes to solve Inter- and Intra-image Gap. The first key component is the PMGM, which generates more accurate prior map by using CAM and categorical consistency assumption, bridging the inter support-query gap. The second key component is the DDM, an architecture that is designed to effectively identify specific inconsistencies in feature extraction, enabling the model to mask specific erroneous support-query pixel pairs, benefiting decoder blocks during training. Extensive experiments and ablation studies demonstrate that IIR achieves new SOTA results on 9 benchmarks, surpassing most of previous methods in both 1-shot and 5-shot segmentation settings. By integrating PMGM and DDM, our model effectively mitigates both inter- and intra-image discrepancies, resulting in more robust segmentation across diverse object categories.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 1, 2
- [2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241. 1, 3
- [3] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017. 1
- [4] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020. 1, 6
- [5] J. Shen, K. Kuang, X. Wang, T. Feng, and W. Zhang, "Cgmgm: A cross-gaussian mixture generative model for few-shot semantic segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4784–4792, 2024. 1, 3, 7
- [6] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 21984–21996. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/b8b12f949378552c21f28deff8ba8eb6-Paper.pdf 1, 3, 7
- [7] X. Shi, D. Wei, Y. Zhang, D. Lu, M. Ning, J. Chen, K. Ma, and Y. Zheng, "Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation," *European Conference on Computer Vision*, 2022. [Online]. Available: <https://www.semanticscholar.org/paper/Dense-Cross-Query-and-Support-Attention-Weighted-Shi-Wei/c876957fc8075969d63f2c488cbdc459dcdb20e> 1
- [8] C. Lang, G. Cheng, B. Tu, and J. Han, "Few-shot segmentation via divide-and-conquer proxies," *International Journal of Computer Vision*, 2023. [Online]. Available: <https://doi.org/10.1007/s11263-023-01886-8> 1, 3
- [9] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 7, 8
- [10] A. Okazawa, "Interclass prototype relation for few-shot segmentation," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 362–378. 1, 3
- [11] S. Moon, S. S. Sohn, H. Zhou, S. Yoon, V. Pavlovic, M. H. Khan, and M. Kapadia, "Hm: Hybrid masking for few-shot segmentation," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 506–523. 1, 3
- [12] Y. Yang, Q. Chen, Y. Feng, and T. Huang, "Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation," p. arXiv:2305.13864, May 01, 2023 2023, accepted to CVPR 2023. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2023arXiv230513864Y> 1, 3
- [13] Q. Xu, G. Lin, C. Loy, C. Long, Z. Li, and R. Zhao, "Eliminating feature ambiguity for few-shot segmentation," in *European Conference on Computer Vision*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271212290> 1, 3, 7
- [14] S. Park, S. Lee, S. Hyun, H. S. Seong, and J.-P. Heo, "Task-disruptive background suppression for few-shot segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4442–4449, Mar. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/28242> 1, 3, 7
- [15] Q. Xu, W. Zhao, G. Lin, and C. Long, "Self-calibrated cross attention network for few-shot segmentation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 655–665. 1
- [16] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4080–4090. 2
- [17] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3637–3645. 2
- [18] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *In International Conference on Learning Representations (ICLR)*, 2017. 2
- [19] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1. 2
- [20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. 3
- [21] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Retain and recover: Delying into information loss for few-shot segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 5353–5365, 2023. 3
- [22] L. Zhu, T. Chen, J. Yin, S. See, and J. Liu, "Addressing background context bias in few-shot segmentation through iterative modulation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3370–3379. 3, 7
- [23] Y. Chen, R. Jiang, Y. Zheng, B. Sheng, Z.-X. Yang, and E. Wu, "Dual branch multi-level semantic learning for few-shot segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1432–1447, 2024. 3, 7
- [24] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, 2022. 3
- [25] X. Luo, Z. Tian, T. Zhang, B. Yu, Y. Y. Tang, and J. Jia, "Pfenet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1273–1289, 2024. 3, 7
- [26] C. Wen, Y. Zhang, J. Fan, H. Zhu, X.-S. Wei, Y. Wang, Z. Kou, and S. Sun, "Object-level correlation for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 23 689–23 699. 3, 7
- [27] T. Zou, S. Xiong, R. Yao, and Y. Rong, "Balancing conservatism and aggressiveness: Prototype-affinity hybrid network for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 20 561–20 571. 3
- [28] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4650–4666, 2023. 3, 7
- [29] S. Hong, S. Cho, J. Nam, and S. Kim, "Cost Aggregation Is All You Need for Few-Shot Segmentation," *arXiv e-prints*, p. arXiv:2112.11685, Dec. 2021. 3, 7, 8
- [30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003. 3, 7
- [31] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: <https://arxiv.org/abs/2408.00714> 3
- [32] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=yzRXdhk2he> 3, 6, 7, 1, 2
- [33] A. Zhang, G. Gao, J. Jiao, C. H. Liu, and Y. Wei, "Bridge the points: Graph-based few-shot segment anything semantically," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 33 232–33 261. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/3a2ef31a1e45908901adc0ca853a8faf-Paper-Conference.pdf 3
- [34] A. Karimi and C. Poullis, "Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 4584–4594. 3
- [35] Q. Xu, L. Zhu, X. Liu, G. Lin, C. Long, Z. Li, and R. Zhao, "Unlocking the power of sam 2 for few-shot segmentation," 2025. [Online]. Available: <https://arxiv.org/abs/2505.14100> 3
- [36] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, Y. Qiao, G. Peng, and H. Li, "Personalize segment anything model with one shot," in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki,

- M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 18 250–18 279. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/4f44e5c82dc7b0ba4472083e5dedb08e-Paper-Conference.pdf 3, 2
- [37] S. Lei, X. Zhang, J. He, F. Chen, B. Du, and C.-T. Lu, “Cross-domain few-shot semantic segmentation,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 73–90. 3, 9
- [38] J. Nie, Y. Xing, G. Zhang, P. Yan, A. Xiao, Y.-P. Tan, A. C. Kot, and S. Lu, “Cross-domain few-shot segmentation via iterative support-query correspondence mining,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3380–3390. 3, 6, 9, 2
- [39] J. Tong, Y. Zou, Y. Li, and R. Li, “Lightweight frequency masker for cross-domain few-shot semantic segmentation,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 96 728–96 749. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/af536fa4312b5b2e5c541156cd31e93b-Paper-Conference.pdf 3, 9
- [40] Y. Liu, Y. Zou, Y. Li, and R. Li, “The devil is in low-level features for cross-domain few-shot segmentation,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 4618–4627. 3
- [41] J. Tong, R. Ma, Y. Zou, G. Chen, Y. Li, and R. Li, “Adapter naturally serves as decoupler for cross-domain few-shot semantic segmentation,” in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 59 829–59 845. [Online]. Available: <https://proceedings.mlr.press/v267/tong25b.html> 3
- [42] Q. Xu, X. Liu, L. Zhu, G. Lin, C. Long, Z. Li, and R. Zhao, “Hybrid mamba for few-shot segmentation,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 73 858–73 883. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/86fe62e3b315d2578721562d9fd1a433-Paper-Conference.pdf 7
- [43] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, “Base and meta: A new perspective on few-shot segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 669–10 686, 2023. 6, 7, 8, 10, 1
- [44] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, “Images speak in images: A generalist painter for in-context visual learning,” *CoRR*, vol. abs/2212.02499, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2212.02499> 7, 8
- [45] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, “Seggpt: Towards segmenting everything in context,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. 7, 8
- [46] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, “Fss-1000: A 1000-class dataset for few-shot segmentation,” *CVPR*, 2020. 6, 1
- [47] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “Deepglobe 2018: A challenge to parse the earth through satellite images,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 172–17209. 6, 1
- [48] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172. 6, 1
- [49] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, “Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014. 6, 1
- [50] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald, “Automatic tuberculosis screening using chest radiographs,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014. 6, 1
- [51] J. Herzog, “Adapt before comparison: A new perspective on cross-domain few-shot segmentation,” *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23 605–23 615, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:268031829> 9
- [52] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, pp. 303–338, 06 2010. 1
- [53] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in *European Conference on Computer Vision (ECCV)*, 2014. 1
- [54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755. 1
- [55] A. Gupta, P. Dollár, and R. Girshick, “Lvis: A dataset for large vocabulary instance segmentation,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5351–5359. 1
- [56] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, “Detect what you can: Detecting and representing objects using holistic models and body parts,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1
- [57] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, A. Mousavi, Y. Song, A. Dubey, and D. Mahajan, “Paco: Parts and attributes of common objects,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7141–7151. 1
- [58] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, and J. Ye, “Fives: A fundus image dataset for artificial intelligence based vessel segmentation,” *Scientific Data*, vol. 9, no. 1, p. 475, 2022. [Online]. Available: <https://doi.org/10.1038/s41597-022-01564-3> 1
- [59] A. Hoover, V. Kouznetsova, and M. Goldbaum, “Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, 2000. 1
- [60] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “DINOv2: Learning Robust Visual Features without Supervision,” *arXiv e-prints*, p. arXiv:2304.07193, Apr. 2023. 2
- [61] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ArXiv*, vol. abs/2010.11929, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:225039882> 2
- [62] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, “Axiom-based grad-cam: Towards accurate visualization and explanation of cnns,” 2020. 3
- [63] R. L. Draelos and L. Carin, “Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks,” 2021. 3
- [64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. 3
- [65] L. Zhu, Q. Chen, L. Jin, Y. You, and Y. Lu, “Bagging regional classification activation maps for weakly supervised object localization,” in *European Conference on Computer Vision*. Springer, 2022, pp. 176–192. 3
- [66] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847. 3

APPENDIX

A. CyCTR Method

CyCTR [6] considers the cycle consistency between the support and query features: if pixel j in the support features is most similar to pixel max_q in the query features, and pixel max_q in the query features is most similar to pixel max_s in the support features, then the masks of j and max_s should be consistent, as shown in Algorithm S1. We also present our DDM algorithm in Algorithm S2 for better comparison.

Algorithm S1: CyCTR

```

1 Input:
2 cross attention map  $A \in \mathbb{R}^{h_q w_q \times h_s w_s}$ ,
3 flatten support mask  $M^s \in \mathbb{R}^{h_s w_s}$ ,
4 cross attention mask  $A^s \in \mathbb{R}^{h_q w_q \times h_s w_s}$  initialized by 0
   Output: cross attention map  $A \in \mathbb{R}^{h_q w_q \times h_s w_s}$ 
5 for  $j = 1$  to  $h_s w_s$  do
6    $max_q = \arg \max_i A[i, j];$ 
7    $max_s = \arg \max_k A[max_q, k];$ 
8   if  $M^s[j] \neq M^s[max_s]$  then
9      $A^s[:, j] = -\infty;$ 
10  end
11 end
12  $A = A + A^s;$ 

```

Algorithm S2: Directional Dropout Module

```

1 Input:
2 cross attention map  $A \in \mathbb{R}^{h_q w_q \times h_s w_s}$ ,
3 flatten support mask  $M^s \in \mathbb{R}^{h_s w_s}$ ,
4 flatten query mask  $M^q \in \mathbb{R}^{h_q w_q}$ 
5 Output: masked cross attention map  $A' \in \mathbb{R}^{h_q w_q \times h_s w_s}$ 
6 Initialize cross attention mask  $A^s \in \mathbb{R}^{h_q w_q \times h_s w_s}$  and
   mask ratio  $r$  with 0;
7 for  $i = 1$  to  $h_q w_q$  do
8    $j = \arg \max_{k \in \{1, 2, \dots, h_s w_s\}} A[i, k];$ 
9   if  $M^q[i] \neq M^s[j]$  then
10     $A^s[i, j] = -\infty;$ 
11     $r = r + \frac{1}{h_q w_q \times h_s w_s};$ 
12  end
13 end
14  $A' = A + A^s;$ 
15  $A' = A' \times \frac{1}{1-r};$ 

```

B. Datasets

We select a total of 9 existing and 2 new few shot datasets across three tasks for experimental validation:

- 1) PASCAL-5ⁱ [4]: *FSS*. The PASCAL-5ⁱ consists of PASCAL VOC 2012 dataset [52] and the SBD dataset [53]. The 20 classes in PASCAL-5ⁱ are divided into 4 folds, with each fold containing 5 classes. During training, we use 3 fold for training and reserve 1 fold for testing, following SG-ONE [4].

- 2) COCO-20ⁱ [4]: *FSS*. COCO-20ⁱ is originated from COCO [54], the data are split into 4 folds, with each fold containing 20 classes.
- 3) LVIS-92ⁱ [32]: *FSS*. LVIS-92ⁱ is based on the LVIS dataset [55], a more challenging benchmark for evaluating the generalization of a model across datasets. After removing the classes with less than two images, the dataset retain 920 classes for training and testing. These classes were then divided into 10 equal folds.
- 4) PACSCAL-Part [32]: *OPFSS*. The dataset is built based on PASCAL [52] and its body part annotations [56]. The dataset consists of four superclasses: animals, indoor, person, and vehicles.
- 5) PACO-Part [32]: *OPFSS*. The dataset is built based on PACO dataset [57] with 303 object part categories. The dataset consists four folds, each with about 76 object parts.
- 6) FSS-1000 [46]: *CDFSS (natural $\rightarrow$ natural)*. The dataset comprising 1000 natural image categories, with each category contains 10 samples.
- 7) Deepglobe [47]: *CDFSS (natural $\rightarrow$ remote sensing)*. The dataset consists of satellite images with 7 foreground classes: urban, agriculture, rangeland, forest, water, barren, and unknown.
- 8) ISIC2018 [48]: *CDFSS (natural $\rightarrow$ medical)*. The dataset comprises 2,596 images of skin cancer screening, with each image depicting exactly one primary lesion.
- 9) Chest X-ray [49], [50]: *CDFSS (natural $\rightarrow$ medical)*. Dataset consists X-ray Tuberculosis images. It includes 566 images, which are collected from 58 cases with a manifestation of Tuberculosis and 80 normal cases.
- 10) Fives: *CDFSS (natural $\rightarrow$ medical)*. The FIVES dataset [58] consists of 800 color fundus photographs with 2048×2048 multi-disease pixel-wise retinal vessel annotation, including 200 healthy, 200 age related macular degeneration, 200 diabetic retinopathy, and 200 glaucoma. To the best of our knowledge, this is the largest retinal vessel segmentation dataset.
- 11) Stare: *CDFSS (natural $\rightarrow$ medical) few shot semantic segmentation*. The STARE dataset [59] contains 20 fundus images with a resolution of 700×605 pixels. Images cover various lesion types, such as macular degeneration, hypertensive retinopathy, and diabetic retinopathy, among others.

C. Implementation Details

The implementation details for each dataset follow their respective baselines, as specified below:

- 1) PASCAL-5ⁱ and COCO-20ⁱ: Our model adopts BAM [43] as its baseline. Since the first stage of BAM fine-tunes the encoder model, we incorporate the KL divergence loss into the classification head to maintain the consistency of class mapping, which ensures that CAM can correctly map to foreground regions. During the inference phase, the Directional Dropout Module is not employed due to the unknown query mask. For this study, ResNet50 and ResNet101 [1] are used as

TABLE S1
PERFORMANCE OF USING DIFFERENT THE NUMBER OF CROSS DECODER BLOCKS AND DECODER BLOCKS, WHERE CAB IS THE ABBREVIATION OF CROSS ATTENTION BLOCK. (THE CROSS ATTENTION BLOCK NUMBER MUST BE LESS THAN DECODER NUMBER)

mIoU \ Decoder	2	3	4	5	6
CAB					
0	70.87	71.49	71.59	71.61	71.53
1	71.03	71.57	71.61	71.83	71.95
2	71.55	71.69	71.89	72.04	72.18
3	-	71.82	71.83	72.23	72.31
4	-	-	72.28	72.67	72.49
5	-	-	-	72.45	72.91
6	-	-	-	-	72.53

backbones. For the PASCAL-5ⁱ dataset, the image size, batch size, and learning rate are set to 473×473, 8, and 0.05, respectively. For COCO-20ⁱ, these parameters are configured as 641×641, 8, and 0.05, respectively.

- 2) LVIS-92ⁱ: We follow the setting of Matcher [32], which uses the model trained on COCO-20ⁱ for testing directly.
- 3) PASCAL-part and PACO-Part: Following the experimental settings of PerSAM [36] and Matcher [32], we use DINOv2 [60] (with ViT-L/14 [61]) as the feature extraction backbone. For the PASCAL-Part dataset, the model is first trained on the PASCAL-5ⁱ and then transferred to PASCAL-Part; For the PACO-Part dataset, the model is first trained on the COCO-20ⁱ and then transferred to PACO-Part. For both datasets, the image size is set to 518×518, batch size to 4.
- 4) Cross domain few shot datasets: Our model takes IFA [38] as its baseline. The initial model is first trained on PASCAL-5ⁱ and then fine-tuned on a cross-domain dataset following the second stage of IFA. During this fine-tuning stage, the augmented support image-mask pair is utilized as the query image-mask pair for domain adaptation. Finally, the model is tested on the specific dataset. We adopt ResNet50 [1] as the backbone of our model. Following the baseline IFA [38], the size of both the support and query images is set to 400×400. During the fine-tuning step, the SGD optimizer is used and the learning rate is set to 0.01. The weight decay is set to 1e-2 for ISIC and 1e-3 for other datasets. Since the pixel-wise retinal vessel segmentation task has the problem of unbalanced foreground and background, the AdamW optimizer with a learning rate of 5e-4 and the weighted BCE loss are employed.

D. More Ablation Study

1) *Decoder Number*: The decoder number is a key factor affecting the results. Table S1 and Fig. S1 demonstrate that when the number of decoders is 6 and the number of CABs is 5, the mIoU reaches 72.91%. If the number of CABs is less than 5, the support-query feature interaction is insufficient (the mIoU is only 72.18% when the number of CABs is 2). If the number of decoders exceeds 6, the performance enters a saturation stage (the mIoU drops to 72.53% when the number of decoders is 7) with increased computational cost. This

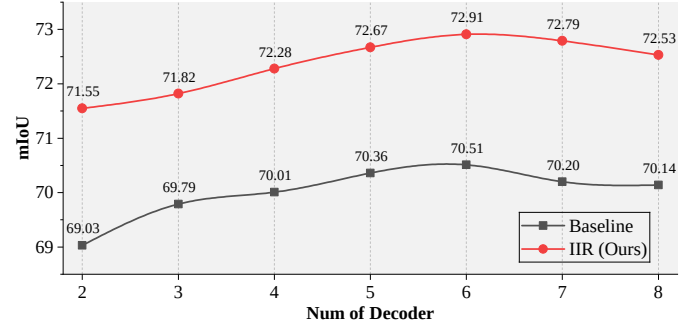


Fig. S1. Saturation experiments on the number of decoders.

verifies that the architecture of IIR achieves an optimal balance between feature interaction and computational efficiency.

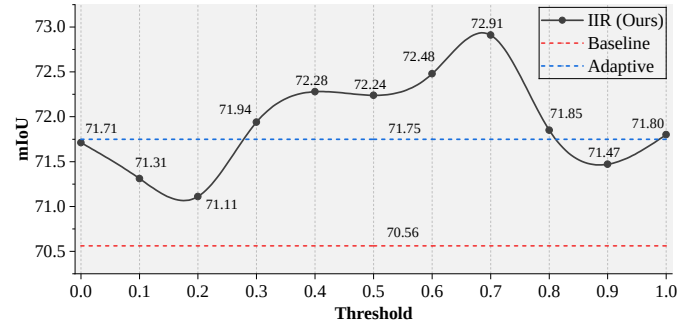


Fig. S2. Performance of using different CAM threshold.

2) *CAM Methods & Thresholds*: Different CAM methods and thresholds focus on different aspects of images, and thus affect the prototype generation regions, thereby influencing the accuracy of prior maps and final prediction. Table S2 demonstrates that XGradCAM achieves the highest mIoU of 72.91% among all CAM methods, outperforming methods such as GradCAM, as it can more accurately localize class-relevant regions. Fig. S2 further confirms that the optimal performance is achieved when the CAM threshold is set to 0.7: a threshold below 0.3 leads to scattered foreground in the CAM, confusing foreground and background, while a threshold above 0.7 results in the loss of local features (e.g., target edge details). This component selection ensures that the dual prototypes generated by PMGM possess both global semantics and local specificity. The adaptive threshold is defined within the range [0, 1] with an interval of 0.01. Specifically, it refers to the value that maximizes the mIoU between the corresponding support pseudo mask and the support ground truth mask. We found that the adaptive threshold achieves a mIoU of 71.75%, with the optimal thresholds typically falling within the range of [0.2, 0.5]. This limited performance is due to the fact that thresholds within this range leads to decentralization of the CAM foreground, causing confusion between foreground and background features.

E. More Visualization

We provide supplementary visualizations of the PASCAL-5ⁱ, LVIS-92ⁱ, and paco-part datasets.

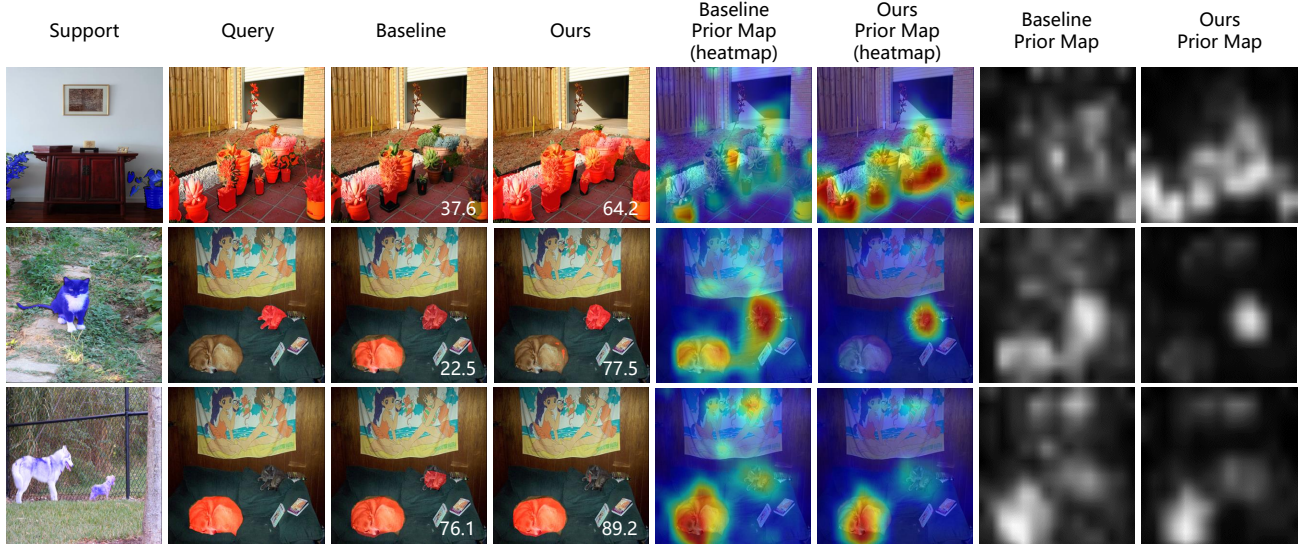


Fig. S3. More visualization on PASCAL-5ⁱ dataset.

TABLE S2
PERFORMANCE OF USING DIFFERENT CAM METHOD

CAM Method	mIoU
XGradCAM(2020) [62]	72.91
HiResCAM(2021) [63]	71.91
GradCAM(2017) [64]	72.23
BagCAM(2023) [65]	71.54
GradCAM++(2018) [66]	71.67

Additional demonstrations for the PASCAL-5ⁱ dataset under the ResNet50 1-shot configuration are presented in Fig. S3. We have additionally conducted corresponding visualizations on the LVIS-92ⁱ dataset, which are displayed in Fig. S4. These results illustrate that our approach achieves superior segmentation performance for multiple targets, targets of varying sizes, and irregularly shaped targets. Fig. S5 showcases several representative segmentation results to depict typical scenarios in the few shot object part segmentation task. In this task, the model is required to perform segmentation on cropped images, as illustrated in the second column of the figure.

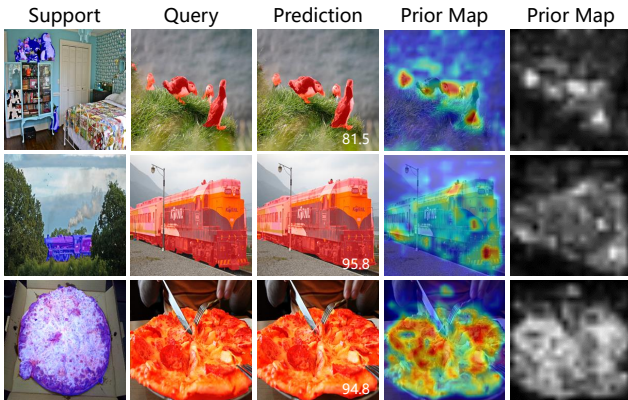


Fig. S4. Visualization on LVIS-92ⁱ dataset for multiple targets, size-varying targets, and irregularly shaped targets.

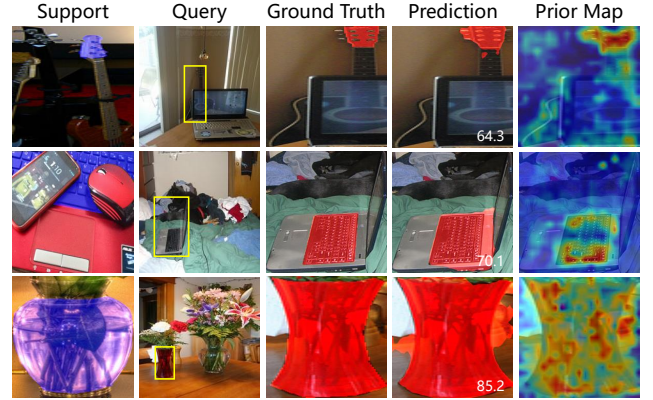


Fig. S5. Visualization on PACO-Part dataset. The yellow box in the query image (second column) serves as model input (third column).

F. Error Cases Analysis

We also investigate several error cases in the datasets. These cases can be categorized as follows:

- Small foreground hidden behind large background objects: when the ground truth mask is obscured by a large object in the background, the model tends to misclassify that large object as the foreground. (e.g. the first row in Fig. S6 and S7)
- Camouflaged target or small target segmentation. We notice that some foreground objects have merged with the background or are difficult to discern in query or support images, leading the model to fail in extracting useful features from them (e.g. the second row in Fig. S6 and S7).
- Objects with similar features are classified into different classes. We observe that certain minority classes are difficult to distinguish, particularly in the COCO-20ⁱ dataset which contains more classes to be segmented (e.g. the third row in Fig. S7).
- Errors in mask labeling: we find that some annotations are

incomplete (e.g. the third row in Fig. S6 and the last row in Fig. S7) and certain class is annotated with foreground contours (e.g. the ‘bicycle’ class in PASCAL-5ⁱ, as shown in the last row in Fig. S6).

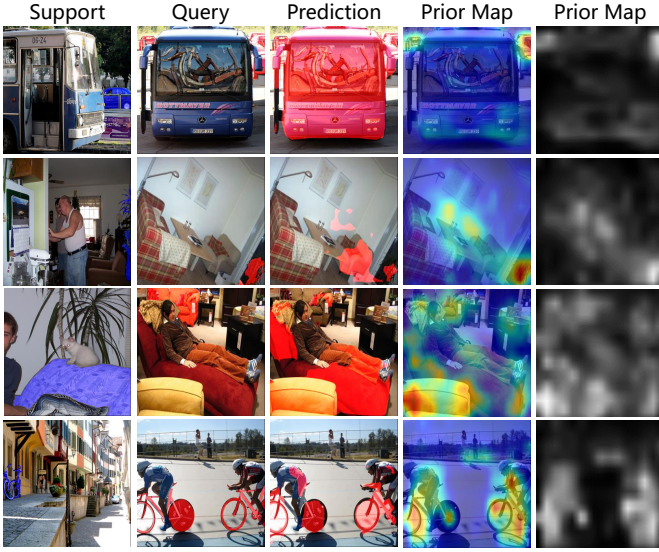


Fig. S6. Visualization of error cases on PASCAL-5ⁱ dataset.

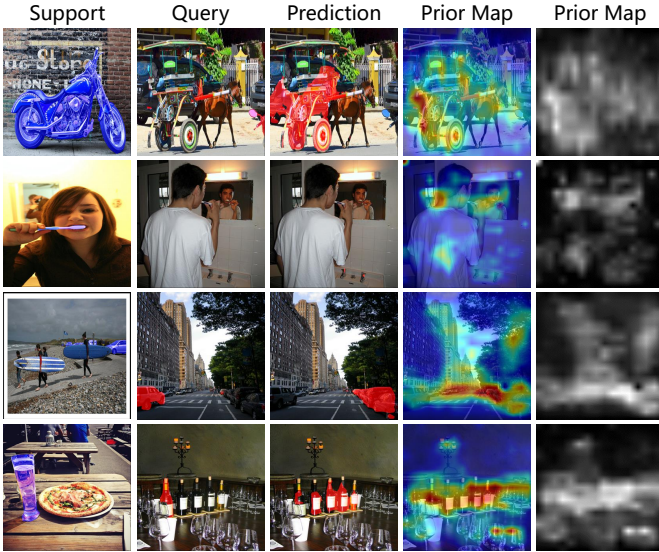


Fig. S7. Visualization of error cases on COCO-20ⁱ dataset.