Training-free Cross-domain Few-shot Segmentation via Robust Semantic Representation and Matching

Sujun Sun[1,2], Mingwu Ren[1,2], and Haofeng Zhang[1,2](✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, China
{egg, renmingwu, zhanghf}@njust.edu.cn

Abstract. Cross-domain Few-shot Segmentation (CD-FSS) aims to transfer knowledge learned from source domain to distinct target domains, segmenting unseen target classes with only a few annotated samples. Although existing methods have made significant progress, they still rely on training or fine-tuning processes, which incur high computational costs and risk overfitting. We observe that when powerful and general-purpose vision foundation models are incorporated into these methods, their performance shows only marginal improvement or even degrades due to overfitting. To address this, we eliminate trainable parameters and propose a training-free framework to avoid both training overhead and overfitting. Built upon the self-supervised vision encoder DINOv3, our framework addresses cross-domain challenges through three core modules. First, the Semantic-aware Feature Re-fusion (SAFR) module identifies and re-fuses features that emphasize semantic patterns, generating representations with enhanced semantic discriminability. Additionally, the Adaptive Support Enhancement (ASE) module narrows semantic gaps between support and query through robust query information aggregation. Finally, the Hybrid Prototype Matching (HPM) module integrates matching results from diverse prototypes to adapt to varying semantic complexity across domains. Extensive experiments on four target domain datasets demonstrate that our method achieves state-of-the-art performance in CD-FSS without any training. Our code is available at https://github.com/Sparkling-Water/RSRM.

Keywords: Cross-domain Few-shot Segmentation · Training-free Segmentation · Vision Foundation Models

1 Introduction

As a fundamental task in computer vision, semantic segmentation has witnessed significant progress in recent years [5,12,49]. However, most existing methods re-

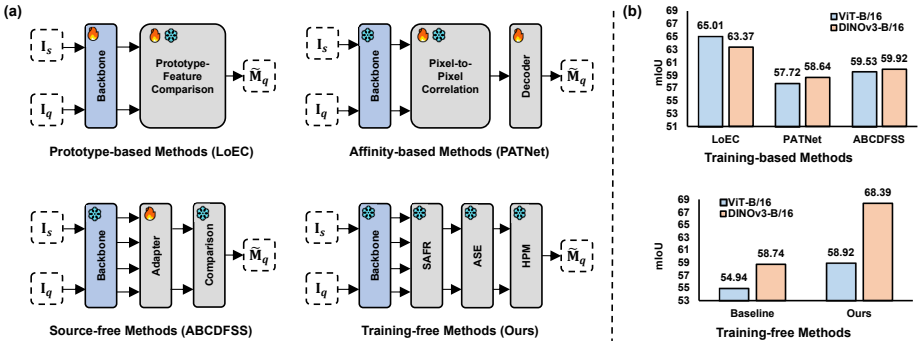

Fig. 1: (a) Existing CD-FSS methods across three paradigms all involve different trainable parameters, while our method requires no additional training. (b) When DINOv3 is incorporated into three existing CD-FSS paradigms, their performance shows only marginal improvement or even degrades. In contrast, our training-free method eliminates overfitting risk and achieves state-of-the-art performance.

main limited to training classes with abundant pixel-level annotations and struggle to extend to novel classes. To address this bottleneck, Few-shot Segmentation (FSS) has been introduced [11,31,37,42]. It leverages meta-learning on base classes with abundant annotated samples to learn generalizable class-agnostic knowledge, enabling the segmentation of novel classes using only a few annotated samples. While achieving remarkable success on in-domain benchmarks, existing FSS methods suffer severe performance degradation when domain gaps exist between training and testing data, limiting their practical applicability.

To bridge domain gaps, Cross-domain Few-shot Segmentation (CD-FSS) [21] has recently attracted increasing attention. Current CD-FSS methods can be broadly categorized into three paradigms: prototype-based methods [28, 29, 35, 39], affinity-based methods [21,38,40], and source-free fine-tuning methods [16]. Although these methods design various strategies to tackle cross-domain challenges, they still rely on source domain training or target domain fine-tuning, which incur high computational costs and risk overfitting. Furthermore, their backbone pretrained on ImageNet [33] limits their performance. Meanwhile, vision foundation models (VFMs) [3,20,30], pretrained on large-scale datasets, can provide powerful and universal visual representations and have demonstrated superior performance across various downstream tasks [13,18,47,51]. Recently, the self-supervised vision encoder DINOv3 [34] has further advanced numerous dense vision tasks [24, 48, 50]. A natural expectation is that leveraging these VFMs as feature extractors could substantially enhance the performance of CD-FSS.

To validate this, we design a simple baseline that directly uses the pretrained model to extract features and obtains predictions via similarity between support prototypes and query features. As shown in Fig. 1(b), experiments on the baseline reveal the excellent cross-domain potential of VFMs. However, integrating the same model into three CD-FSS paradigms yields only marginal improvement

or even degradation. This phenomenon highlights the key limitation of existing CD-FSS paradigms: the trainable parameters in these methods are prone to overfitting on the training data, and the strong initialization provided by powerful visual representations may further accelerate this process.

Based on these observations, we eliminate trainable parameters and adopt a training-free framework to avoid both training overhead and overfitting. Several FSS methods [26, 51, 54] perform training-free adaptation on top of SAM's [20] class-agnostic segmentation framework but suffer from the following limitations: 1) Due to the limited capability of SAM's encoder for cross-image semantic matching, these methods typically need to be combined with models that possess strong cross-image semantic matching ability (*e.g.*, DINOv2 [30]). 2) These methods are designed for single-domain, and their performance is limited in cross-domain scenarios. Moreover, encoder–decoder methods [37, 54] introduce more overfitting-prone parameters, generally resulting in inferior performance and efficiency compared to encoder-only methods [11, 28, 53] in CD-FSS. Therefore, we perform training-free adaptation on a single VFM encoder.

Specifically, we employ DINOv3 as the feature extractor and generalize it to cross-domain scenarios through three key modules. The core idea of our method is to generate discriminative semantic representations and perform robust matching across different domains. First, we observe that certain intermediate features in DINOv3 overemphasize local consistency and propagate to the final output via residual connections, limiting its semantic discriminability. To address this, we design a Semantic-aware Feature Re-fusion (SAFR) module, which adaptively identifies and re-fuses features that emphasize semantic patterns, enhancing the semantic representation quality. In addition, support and query images exhibit varying intra-class instance differences across domains, resulting in significant feature discrepancies. To improve semantic alignment between support and query, we design an Adaptive Support Enhancement (ASE) module. ASE adaptively identifies robust foreground and background regions in the query and aggregates robust query information into support features via masked cross-attention. Finally, we introduce a Hybrid Prototype Matching (HPM) module to handle the variations in intra-image semantic complexity across domains, generating final predictions by fusing the matching results of global, regional, and pixel-level prototypes. Our contributions are as follows:

- We identify that existing CD-FSS paradigms are prone to overfitting when combined with vision foundation models and propose to avoid both training overhead and overfitting in a training-free manner.
- We propose a training-free framework that incorporates three novel modules: SAFR enhances semantic discriminability of features, ASE improves semantic consistency between support and query, and HPM adapts to varying semantic complexity across domains.
- Extensive experiments on four standard target domain datasets demonstrate that our method significantly outperforms current state-of-the-art CD-FSS methods, even without any training.

2 Related Works

2.1 Few-shot Segmentation

The goal of FSS is to segment novel classes in query images using only a few annotated examples. Based on how support samples are utilized, existing FSS methods can be broadly categorized into prototype-based and affinity-based methods. Prototype-based methods adopt global [11, 52] or local [22, 46] prototypes to represent target classes in support images and then match these prototypes with query features through cosine similarity or convolution operations to generate segmentation masks. In contrast, affinity-based methods [31, 43, 44] argue that prototype computation ignores spatial details, thus focusing on constructing pixel-level correlations between support and query features and decoding them to predict query masks. Recently, some works [4, 37, 45] have leveraged the well-learned knowledge of vision foundation models (*e.g.*, DINOv2 [30], SAM [20] and SAM2 [32]) to simplify the learning of FSS, achieving significant performance improvements. However, these methods assume that base and novel classes originate from the same domain and fail to generalize well to unseen domains.

2.2 Cross-domain Few-shot Segmentation

Compared with FSS, CD-FSS is a more realistic setting, requiring the model trained on the source domain to generalize to novel classes in different target domains. Similar to the categorization in FSS, existing CD-FSS methods can be mainly divided into three paradigms: prototype-based methods, affinity-based methods, and source-free fine-tuning methods. Prototype-based methods [28, 29, 35, 36, 39] typically train the backbone on the source domain, introducing strategies such as domain perturbation or feature decomposition to prevent overfitting, and generate query masks via a prototype-feature matching module. Affinity-based methods [21, 38, 40] freeze the pretrained backbone, transform original features into domain-agnostic representations, compute dense affinities between support and query features, and then generate query masks using a decoder trained on the source domain. Source-free fine-tuning methods [16] also freeze the pretrained backbone but introduce a set of learnable adaptation layers for its hierarchical features, which are then fine-tuned with limited target domain samples. All these methods rely on training or fine-tuning processes, which incur substantial computational costs and risk overfitting. In this paper, we discard both training and fine-tuning processes, and avoid training overhead and overfitting in a training-free manner.

2.3 Training-free Few-shot Segmentation

Some FSS methods [26, 51, 54] and Few-shot Medical Image Segmentation (FSMIS) approaches [27, 55] have achieved promising performance by improving upon the class-agnostic segmentation framework of SAM [20] without introducing learnable parameters, thereby avoiding high training costs and the risk of

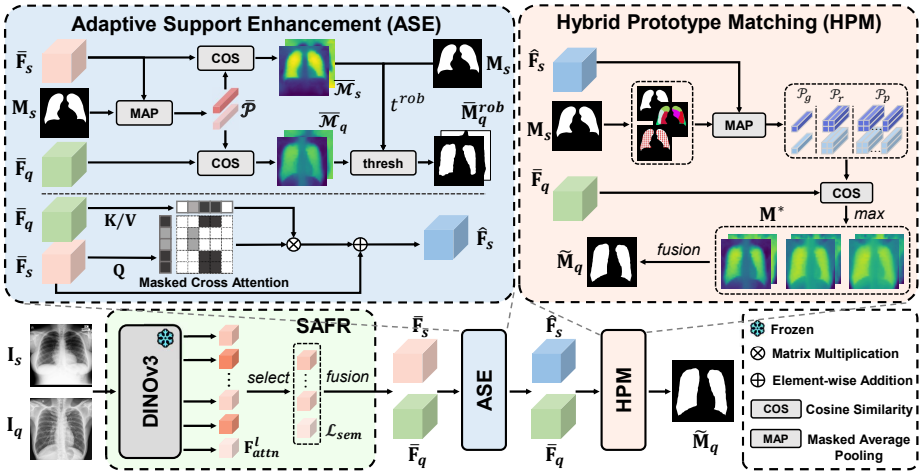

Fig. 2: Overview of our method. We first utilize DINOv3 to extract multi-layer features of both support and query images. The SAFR then locates and re-fuses representative semantic-aware features from these extracted features. Subsequently, the fused support features are enhanced by the ASE module. Finally, the HPM computes and integrates the matching results between diverse support prototypes and query features to generate the final query prediction.

overfitting. However, these methods typically require integration with additional vision foundation models, resulting in a more complex pipeline. Moreover, they are primarily designed for single-domain data (such as natural images or medical images), and their effectiveness in cross-domain scenarios remains unverified. In contrast, we aim to generalize a single pretrained vision encoder across different target domains.

3 Methodology

3.1 Problem Definition

In the traditional CD-FSS task, the model is trained on the source domain $\mathcal{D}_s = \{X_s, Y_s\}$ and evaluated on the target domain $\mathcal{D}_t = \{X_t, Y_t\}$, where X denotes the data distribution and Y represents the corresponding label space. The source and target domains have no overlap in both data distributions and label spaces, *i.e.*, $X_s \neq X_t$ and $Y_s \cap Y_t = \emptyset$.

Unlike traditional CD-FSS work, we directly generalize the model to different target domains during testing, without utilizing any source domain data for training. Each testing task is organized as an episode, consisting of a support set $\mathcal{S} = \{(\mathbf{I}_s^i, \mathbf{M}_s^i)\}_{i=1}^K$ and a query set $\mathcal{Q} = \{(\mathbf{I}_q, \mathbf{M}_q)\}$, where K denotes the number of support samples, $\mathbf{I}$ denotes an image, and $\mathbf{M}$ denotes its binary mask. In each episode, the model predicts the query mask based on the support set $\mathcal{S}$ and the query image $\mathbf{I}_q$.

3.2 Method Overview

The overall architecture of our proposed training-free framework is illustrated in Fig. 2. First, support and query images from the target domain are fed into DINOv3 to extract multi-layer raw features. The Semantic-aware Feature Re-fusion (SAFR) adaptively identifies and re-fuses features that emphasize semantic patterns, thereby generating support and query features with enhanced semantic discriminability. After computing support prototypes, the Adaptive Support Enhancement (ASE) module constructs the similarity map between these prototypes and query features, generating robust query foreground and background region masks based on adaptive thresholds. Subsequently, the support features aggregate information from corresponding query regions through masked cross-attention, thereby improving semantic consistency between support and query features. Finally, the Hybrid Prototype Matching (HPM) module extracts global, regional, and pixel-level prototypes from enhanced support features and fuses the matching results between query features and these prototypes to generate the final query prediction. Note that all features refer to 2D feature maps obtained by spatially reshaping the local patch tokens.

3.3 Semantic-aware Feature Re-fusion

Analysis of DINOv3 Features. Extracting features with strong semantic discriminability is a key prerequisite for CD-FSS. To this end, given an input image $\mathbf{I}$, we first perform principal component analysis (PCA) on the last-layer feature $\mathbf{F}_{last} \in \mathbb{R}^{H\times W\times C}$ extracted by DINOv3 and visualize the top three components, where H, W, and C denote the height, width, and channel depth of the feature, respectively. As shown in Fig. 3(a), a notable observation is that these features exhibit strong local consistency, while their semantic discriminability is relatively weak. For further investigation, we extract and visualize the top four channels with the highest variance, finding that these channels typically encode patterns related to pixel positions rather than semantically meaningful concepts.

To trace the origin of these position-sensitive channels, we extract the raw outputs $\{\mathbf{F}^l_{attn}\}^L_{l=1}$ of attention modules and the fused outputs $\{\mathbf{F}^l_{sum}\}^L_{l=1}$ after residual connections in each block, where L is the number of layers. Additionally, $\mathbf{F}^l_{sum} = \mathrm{FFN}(\mathrm{LN}(\mathbf{F}^{l-1}_{sum}+\mathbf{F}^l_{attn}))$ and $\mathbf{F}_{last} = \mathbf{F}^L_{sum}$, where LN denotes layer normalization, and FFN stands for a feed-forward network. Feature visualizations show that DINOv3 layers alternate between focusing on semantic patterns and positional variations. These position-aware features propagate via residual connections to $\mathbf{F}_{last}$, limiting its semantic discriminability. See the supplementary material for details.

For quantitative validation, we select the top $N_c = \mathrm{round}(C/\gamma)$ highest-variance channels from F_{last} as the position-sensitive channel set $\mathcal{C}$, where γ is a hyperparameter controlling the size of $\mathcal{C}$. Then, we compute the variance proportion r of $\mathcal{C}$ in each feature $\mathbf{F}$:

$$r = \frac{\sum_{c\in\mathcal{C}} \sigma^2(\mathbf{F}_c)}{\sum_{c=1}^{C} \sigma^2(\mathbf{F}_c)}, \tag{1}$$

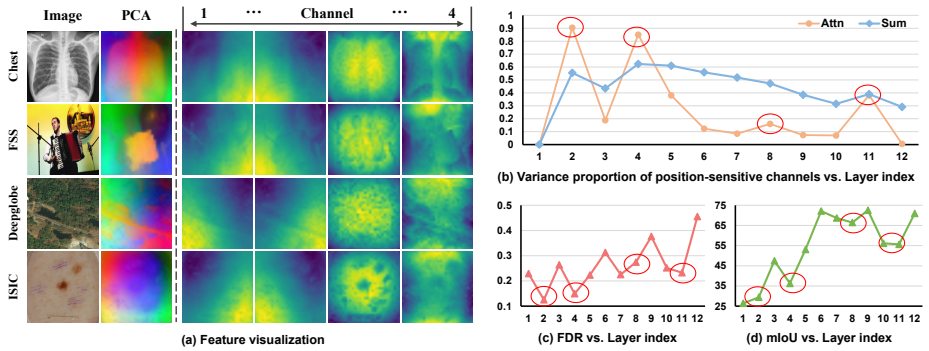

Fig. 3: (a) We perform PCA on the last-layer feature $\mathbf{F}_{last}$ of DINOv3 and visualize the first three components, indicating that these features are overly focused on local consistency. Further visualization of the top four channels with the highest variance in $\mathbf{F}_{last}$ reveals that these channels typically encode patterns related to pixel positions rather than semantically meaningful concepts. (b)-(d) Statistical analysis of features across layers in DINOv3 on the Chest X-ray dataset.

where $\sigma(\mathbf{F}_c)$ denotes the standard deviation of the c-th channel in feature $\mathbf{F}$. As shown in Fig. 3(b), the position-aware feature $\mathbf{F}^l_{attn}$ with higher r^l_{attn} increases r^l_{sum} of the fused feature $\mathbf{F}^l_{sum}$, and this persists until the last layer.

Furthermore, for each $\mathbf{F}^l_{attn}$, we compute its Fisher discriminant ratio (FDR) [19] f^l_{attn} and evaluate its CD-FSS performance. The FDR is defined as the trace ratio between the between-class and within-class scatter matrices, reflecting the inter-class discriminability of features. Fig. 3 shows that layers with higher variance proportions generally exhibit lower FDR, indicating weaker foreground-background discriminability, and these layers also demonstrate relatively inferior segmentation performance.

Feature Re-fusion. Position-aware features weaken the semantic discriminability of $\mathbf{F}_{last}$. As shown in Tab. 3, directly modifying or discarding these features alters the feature distribution in subsequent layers, resulting in a further drop in segmentation performance. Therefore, instead of using the standard final output $\mathbf{F}_{last}$, we re-fuse semantically-aware features to construct a representation with enhanced semantic discriminability.

Specifically, given an input image $\mathbf{I}$, we first extract the raw outputs $\{\mathbf{F}^l_{attn}\}^L_{l=1}$ and the final output $\mathbf{F}_{last}$, computing the position-sensitive channel set $\mathcal{C}$ from $\mathbf{F}_{last}$. We then compute the variance proportions $\{r^l_{attn}\}^L_{l=1}$ and FDRs $\{f^l_{attn}\}^L_{l=1}$ for each layer. Subsequently, we identify all layers where r_{attn} exceeds the average as position-aware layers and add them to the position-aware layer set $\mathcal{L}_{pos}$:

$$\mathcal{L}_{pos} = \{l^i_{pos}\}^{|\mathcal{L}_{pos}|}_{i=1} = \left\{ l \;\middle|\; r^l_{attn} > \frac{1}{L}\sum_{l'=1}^{L} r^{l'}_{attn} \right\}. \tag{2}$$

We also add $L+1$ to $\mathcal{L}_{pos}$ and sort it in ascending order. Next, we sequentially select the layer with the minimum r_{attn} between adjacent position-aware layers to form the representative semantic-aware layer set $\mathcal{L}_{sem}$:

$$\mathcal{L}_{sem} = \bigcup_{i=0}^{|\mathcal{L}_{pos}|-1} \left\{ \arg \min_{l \in (l_{pos}^{i}, l_{pos}^{i+1})} r_{attn}^{l} \right\}. \tag{3}$$

Finally, all representative semantic-aware features are weighted to obtain the fused feature $\bar{\mathbf{F}}$:

$$\bar{\mathbf{F}} = \sum_{l \in \mathcal{L}_{sem}} \omega^{l} \mathbf{F}_{attn}^{l}, \quad \omega^{l} = \frac{f_{attn}^{l} / r_{attn}^{l}}{\sum_{l' \in \mathcal{L}_{sem}} f_{attn}^{l'} / r_{attn}^{l'}}, \tag{4}$$

where layers with lower r_{attn} and higher f_{attn} are assigned higher weights. We utilize support features to compute $\mathcal{L}_{pos}$, $\mathcal{L}_{sem}$, and ω^{l}, applying them to both support and query features.

3.4 Adaptive Support Enhancement

Due to variations in foreground instances or background categories, the semantic consistency between some support and query features is weak, leading to insufficient cross-image semantic discrimination. Some FSS methods [11, 31, 44] mitigate this by introducing query-support feature interaction or fusion modules. However, these methods typically involve trainable parameters or fixed thresholds, which fail to generalize across domains. To address this, we aim to design a support enhancement module that is training-free and adaptable to different domains, thereby reducing the semantic gap between support and query. The core idea is to aggregate query information into the support features via cross-attention without projection layers, while restricting the aggregation to robust query foreground/background regions to prevent feature confusion.

Specifically, after obtaining the fused features $\bar{\mathbf{F}}_s$ and $\bar{\mathbf{F}}_q$, we compute the support prototypes $\bar{\mathcal{P}} = \{\bar{\mathbf{p}}_f, \bar{\mathbf{p}}_b\}$ via masked average pooling (MAP), further computing their cosine similarity with $\bar{\mathbf{F}}_q$ to obtain the query similarity map $\bar{\mathcal{M}}_q = \{\bar{\mathbf{M}}_{q,f}, \bar{\mathbf{M}}_{q,b}\} = \text{softmax}(\text{cosine}(\bar{\mathcal{P}}, \bar{\mathbf{F}}_q))$. The support similarity map $\bar{\mathcal{M}}_s = \{\bar{\mathbf{M}}_{s,f}, \bar{\mathbf{M}}_{s,b}\}$ is computed in the same way.

To locate robust query foreground regions, we first sample a set of thresholds $\mathcal{T}_f = \{t_f^i\}_{i=1}^{N_t}$ uniformly from the interval (0, 1), where N_t is the number of thresholds. Each t_f^i is utilized to binarize the support foreground similarity map $\bar{\mathbf{M}}_{s,f}$ to obtain the corresponding support prediction mask $\bar{\mathbf{M}}_s^i = \mathbb{I}(\bar{\mathbf{M}}_{s,f} > t_f^i)$, where $\mathbb{I}$ is the indicator function. Subsequently, we compute the mIoU between $\bar{\mathbf{M}}_s^i$ and the ground-truth mask $\mathbf{M}_s$, selecting the threshold with the highest mIoU as the optimal threshold t_f^*. Since the values of support-support similarity are typically higher than support-query similarity, we compute a more robust threshold t_f^{rob} instead of directly using t_f^*:

$$t_f^{rob} = \frac{1}{|\mathcal{T}_f^{rob}|} \sum_{j \in \mathcal{T}_f^{rob}} t_f^j, \quad \mathcal{T}_f^{rob} = \{t | t \in \mathcal{T}_f, t > \alpha t_f^*\}, \tag{5}$$

where $\alpha = 0.75$ is a hyperparameter. We then use t_f^{rob} to binarize $\bar{\mathbf{M}}_{q,f}$, obtaining the robust query foreground mask $\bar{\mathbf{M}}_{q,f}^{rob}$. The robust query background mask $\bar{\mathbf{M}}_{q,b}^{rob}$ is obtained in the same way.

Next, we reshape $\bar{\mathbf{F}}_s$ into $\mathbf{Q} \in \mathbb{R}^{HW \times C}$ and $\bar{\mathbf{F}}_q$ into $\mathbf{K} \in \mathbb{R}^{HW \times C}$ and $\mathbf{V} \in \mathbb{R}^{HW \times C}$, aggregating query information into support features through training-free cross-attention. To prevent information confusion, we restrict support foreground/background regions to interact only with robust query foreground/background regions, which is achieved by adding a mask $\mathbf{M}_{sq} \in \mathbb{R}^{HW \times HW}$ to the attention matrix. $\mathbf{M}_{sq}$ is defined as:

$$\mathbf{M}_{sq}(\mathbf{u}_s, \mathbf{u}_q) = \begin{cases} 0, & \text{if } \left(\bar{\mathbf{M}}_{q,f}^{rob}(\mathbf{u}_q)\mathbf{M}_s(\mathbf{u}_s) = 1\right) \\ & \vee \left(\bar{\mathbf{M}}_{q,b}^{rob}(\mathbf{u}_q)\mathbf{M}_{s,b}(\mathbf{u}_s) = 1\right), \\ -\infty, & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{u}_s = (x_s, y_s)$ and $\mathbf{u}_q = (x_q, y_q)$ denote spatial coordinates, and $\mathbf{M}_{s,b} = 1 - \mathbf{M}_s$ is the support background mask. The query information aggregation is formulated as:

$$\mathbf{A}_{sq} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}} + \mathbf{M}_{sq})\mathbf{V}. \tag{7}$$

Finally, we reshape $\mathbf{A}_{sq}$ into $\mathbf{F}_{sq} \in \mathbb{R}^{H \times W \times C}$ and fuse it with $\bar{\mathbf{F}}_s$ to obtain the enhanced feature $\hat{\mathbf{F}}_s = (\bar{\mathbf{F}}_s + \mathbf{F}_{sq})/2$.

3.5 Hybrid Prototype Matching

Images from different target domains exhibit varying semantic complexity, making it difficult for a single form of support-query matching to adapt universally. We propose to fuse matching results from global, regional, and pixel-level prototypes to generate the final query prediction.

Specifically, we first compute three types of support foreground prototype sets. The global foreground prototype $\mathbf{p}_{g,f}$ is obtained by applying MAP to $\hat{\mathbf{F}}_s$ and $\mathbf{M}_s$, forming the global prototype set $\mathcal{P}_{g,f} = \{\mathbf{p}_{g,f}\}$. Simultaneously, we treat each foreground feature in $\hat{\mathbf{F}}_s$ as a pixel-level foreground prototype, constructing the pixel-level prototype set $\mathcal{P}_{p,f} = \{\mathbf{p}_{p,f}^i\}_{i=1}^{Ns,f}$, where $N_{s,f}$ denotes the number of foreground pixels in $\hat{\mathbf{F}}_s$. Additionally, we cluster the support foreground features into N_r clusters using k-means++ [1] and regard their centers as regional prototypes, forming the regional prototype set $\mathcal{P}_{r,f} = \{\mathbf{p}_{r,f}^i\}_{i=1}^{N_r}$. Corresponding background prototype sets $\mathcal{P}_{g,b}$, $\mathcal{P}_{p,b}$, and $\mathcal{P}_{r,b}$ are obtained in a similar manner. The only difference is that, since the background typically contains more semantic categories, the support background features are clustered into $2N_r$ clusters to generate $\mathcal{P}_{r,b}$.

Next, we compute the similarity map for each prototype set. Specifically, given a prototype set $\mathcal{P}_{m,n} = \{\mathbf{p}_{m,n}^i\}_{i=1}^{|\mathcal{P}_{m,n}|}$, where $m \in \{g, r, p\}$ and $n \in \{f, b\}$, we compute the cosine similarity between each prototype $\mathbf{p}_{m,n}^i$ and $\bar{\mathbf{F}}_q$, obtaining

a similarity map $\mathbf{M}^{i}_{m,n} \in \mathbb{R}^{H\times W}$. We then aggregate by taking the pixel-wise maximum to capture the strongest association between each pixel and any prototype, generating the similarity map $\mathbf{M}^{*}_{m,n}$:

$$\mathbf{M}^{*}_{m,n}(x,y) = \max_{i\in\{1,\dots,|\mathcal{P}_{m,n}|\}} \mathbf{M}^{i}_{m,n}(x,y). \tag{8}$$

Subsequently, we compute the pseudo query mask for each prototype type as $\tilde{\mathbf{M}}_m = \mathbb{I}(\mathbf{M}^{*}_{m,f} > \mathbf{M}^{*}_{m,b})$, which is further utilized to compute the average foreground similarity $\mu_{m,f}$ and average background similarity $\mu_{m,b}$ from $\mathbf{M}^{*}_{m,f}$. The difference between these averages, $\omega_m = \mu_{m,f} - \mu_{m,b}$, is employed as both the confidence measure and fusion weight for prototype type m. After normalizing the weights across types with softmax, we compute the fused foreground and background similarity maps as:

$$\tilde{\mathbf{M}}_f = \sum_m \omega_m \mathbf{M}^{*}_{m,f}, \quad \tilde{\mathbf{M}}_b = \sum_m \omega_m \mathbf{M}^{*}_{m,b}. \tag{9}$$

Finally, the query prediction $\tilde{\mathbf{M}}_q$ is obtained by comparing the fused maps, *i.e.*, $\tilde{\mathbf{M}}_q = \mathbb{I}(\tilde{\mathbf{M}}_f > \tilde{\mathbf{M}}_b)$.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. Following the setup of PATNet [21], we evaluate our method on four target domain datasets: Deepglobe [7], ISIC [6,41], Chest X-ray [2,17], and FSS-1000 [23]. Unlike existing CD-FSS methods, our approach does not leverage any additional source domain datasets (*e.g.*, PASCAL [9] and COCO [25]) for training. Please refer to our supplementary material for more details.

Implementation Details. We employ DINOv3 ViT-B/16 [34] as our backbone to extract high-quality visual representations. Following LoEC [28], we resize all images to 480×480. The hyperparameters γ, α, N_t, and N_r are set to 200, 0.75, 20, and 9, respectively. Consistent with previous methods, we use mean intersection over union (mIoU) as the evaluation metric and report the average results over 5 random seeds. Each run consists of 1200 episodes sampled from Deepglobe, ISIC, and Chest X-ray, respectively, and 2400 episodes from FSS-1000. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

4.2 Comparison with State-of-the-art Methods

In Tab. 1, we compare our method with existing methods, including those using ResNet-50 [14] and ViT-B/16 [8] with ImageNet [33] pretrained weights as backbones, as well as methods employing vision foundation models (VFMs) such as DINOv2 [30] and SAM [20] as backbones. For fair comparison, we also replace the backbone of several CD-FSS and training-free methods with DINOv3-B/16.

Table 1: Mean-IoU of 1-shot and 5-shot results compared with previous methods. The best and second-best methods are highlighted in **bold** and underlined, respectively. † marks methods implemented or reproduced by us.

Methods	Task	Mark	Backbone	Deepglobe		ISIC		Chest X-ray		FSS-1000		Average	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Training-based Methods													
SSP [11]	FSS	ECCV-22	ResNet-50	40.00	48.68	35.49	45.86	74.44	74.26	78.91	80.59	57.21	62.35
FPTrans [53]	FSS	NIPS-22	ViT-B	38.36	49.30	48.65	60.37	80.92	82.91	80.74	83.65	62.17	69.06
HDMNet [31]	FSS	CVPR-23	ResNet-50	25.40	39.10	33.00	35.00	30.60	31.30	75.10	78.60	41.00	46.00
FSSAM† [45]	FSS	ICML-25	DINOv2-B+SAM2	40.25	46.58	47.75	59.68	77.11	86.56	80.36	87.11	61.37	69.98
PATNet [21]	CD-FSS	ECCV-22	ResNet-50	37.89	42.97	41.16	53.58	66.61	70.20	78.59	81.23	56.06	61.99
PATNet† [21]	CD-FSS	ECCV-22	DINOv3-B	33.78	40.45	<u>54.37</u>	62.36	62.69	64.42	83.71	84.65	58.64	62.97
ABCDFSS [16]	CD-FSS	CVPR-24	ResNet-50	42.60	49.00	45.70	53.30	79.80	81.40	74.60	76.20	60.67	64.97
ABCDFSS† [16]	CD-FSS	CVPR-24	DINOv3-B	32.57	39.26	49.56	56.32	80.73	82.91	76.80	79.39	59.92	64.47
DRA [35]	CD-FSS	CVPR-24	ResNet-50	41.29	50.12	40.77	48.87	82.35	82.31	79.05	80.40	60.86	65.42
APSeg [15]	CD-FSS	CVPR-24	SAM-B	35.94	39.98	45.43	53.98	<u>84.10</u>	84.50	79.71	81.90	61.30	65.09
APM [40]	CD-FSS	NIPS-24	ResNet-50	40.86	44.92	41.71	51.16	78.25	82.81	79.29	81.83	60.03	65.18
LoEC [28]	CD-FSS	CVPR-25	ViT-B	42.12	51.48	52.91	<u>62.43</u>	83.94	84.12	81.05	83.69	<u>65.01</u>	<u>70.43</u>
LoEC† [28]	CD-FSS	CVPR-25	DINOv3-B	44.34	53.25	50.98	61.38	73.83	74.58	84.33	86.16	63.37	68.84
SDRC [39]	CD-FSS	ICML-25	ViT-B	43.15	46.83	46.57	55.02	82.86	84.79	80.31	82.55	63.22	67.30
DFN [38]	CD-FSS	ICML-25	ViT-B	39.45	47.67	50.36	58.53	83.18	**87.14**	82.97	85.72	63.99	69.77
ISA [10]	CD-FSS	ICCV-25	ResNet-50	44.30	52.70	37.20	56.10	83.40	86.30	78.80	86.00	60.90	70.30
Training-free Methods													
PerSAM [54]	FSS	ICLR-24	SAM-B	36.08	40.65	23.27	25.33	29.95	30.05	60.92	66.53	37.56	40.64
Matcher† [26]	FSS	ICLR-24	DINOv2-B+SAM-B	47.01	50.98	38.29	39.22	76.42	79.04	85.27	**88.52**	61.75	64.44
Matcher† [26]	FSS	ICLR-24	DINOv3-B+SAM-B	46.85	53.73	38.68	46.85	75.86	81.07	83.39	87.44	61.20	67.27
GF-SAM† [51]	FSS	NIPS-24	DINOv2-B+SAM-B	<u>49.05</u>	55.13	44.47	50.25	38.70	37.93	<u>85.75</u>	86.81	54.49	57.53
GF-SAM† [51]	FSS	NIPS-24	DINOv3-B+SAM-B	48.75	<u>57.86</u>	49.13	56.91	43.37	42.63	**86.51**	<u>87.57</u>	56.94	61.24
MAUP† [55]	FSMIS	MICCAI-25	DINOv2-B+SAM-B	41.50	45.06	33.88	36.35	73.96	75.63	73.29	75.97	55.66	58.25
Ours	CD-FSS	Ours	DINOv3-B	**49.71**	**58.88**	**55.73**	**62.55**	**85.44**	<u>87.01</u>	82.67	83.89	**68.39**	**73.08**

The results demonstrate that our method achieves significant improvements under both 1-shot and 5-shot settings, surpassing all existing training-based and training-free methods without any training. Specifically, our method surpasses the state-of-the-art training-based method LoEC [28] by 3.38% and 2.65%, and exceeds the state-of-the-art training-free method Matcher [26] by 7.19% and 5.81% under 1-shot and 5-shot settings, respectively.

Notably, introducing DINOv3 into training-based methods yields marginal gains or even performance degradation, and other VFM-based methods [15, 45] also underperform compared to some VFM-free methods [28, 53]. This indicates that training-based methods still suffer from the risk of overfitting. Existing training-free methods also exhibit suboptimal performance on CD-FSS, which we attribute to two primary factors: 1) The modules or parameters in these methods are carefully designed for single-domain settings, limiting their cross-domain generalization. 2) While these methods benefit from SAM's strong semantic discrimination on natural images (*e.g.*, FSS-1000), this ability diminishes on domains with large gaps, failing to compensate for its inherently limited cross-image matching capability.

4.3 Ablation Studies

Effects of Each Component. We conduct ablation studies on each proposed component, with results shown in Tab. 2. SAFR generates features with enhanced semantic discriminability by re-fusing representative semantic-aware features, improving the average mIoU by 3.79% and 4.10% under 1-shot and 5-shot

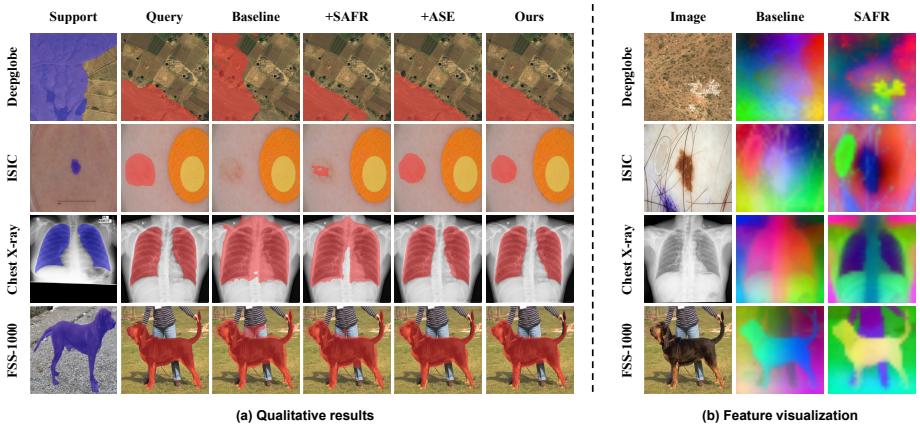

Fig. 4: (a) Qualitative results with progressive module integration in 1-way 1-shot setting. (b) Visualizations of the first three PCA components demonstrate that SAFR generates representations with enhanced semantic discriminability.

Table 2: Ablation studies for each proposed component.

SAFR	ASE	HPM	1-shot	5-shot
×	×	×	58.74	64.54
✓	×	×	62.53	68.64
✓	✓	×	66.16	70.86
✓	×	✓	65.99	71.61
✓	✓	✓	**68.39**	**73.08**

Table 3: Ablation studies for different semantic discriminability enhancement strategies.

Strategy	Deepglobe	ISIC	Chest	FSS	Average
Baseline	42.32	51.15	67.18	74.32	58.74
Reduce weight	40.75	49.64	70.20	72.53	58.28
Reduce variance	41.40	50.54	65.06	73.74	57.69
Remove	48.34	53.26	71.42	75.58	62.15
SAFR	48.78	53.69	72.03	75.60	**62.53**

settings, respectively. Building on this, ASE aggregates robust query information into the support features, enhancing semantic consistency between support and query features, and further improves the average mIoU by 3.63% and 2.22%. HPM integrates matching results from diverse prototypes, resulting in additional mIoU gains of 3.46% and 2.97% for 1-shot and 5-shot settings, respectively. Finally, the combination of all three modules leads to further performance improvements. We also present qualitative results with progressive module integration in Fig. 4(a). These results clearly demonstrate the effectiveness of each component.

Semantic Discriminability Enhancement Strategies. We compare the performance of different semantic discriminability enhancement strategies under the 1-shot setting, as shown in Tab. 3. We first directly modify position-aware features, such as reducing their weight in residual connections or decreasing the variance proportion of position-sensitive channels through variance regularization. Since these strategies alter the feature distribution of subsequent layers, they actually lead to further degradation in performance. In contrast, explicitly removing the position-sensitive channels from the standard final output $\mathbf{F}_{last}$ results in a significant performance improvement, indicating that these channels suppress semantic discriminability. Our SAFR module avoids using $\mathbf{F}_{last}$ affected by position-sensitive channels, achieving better performance.

Table 4: Ablation studies for the effects of threshold types.

	Deepglobe	ISIC	Chest	FSS	Average
w/o mask	49.52	54.57	73.16	77.80	63.76
Fixed threshold	49.64	56.03	74.34	78.11	64.53
Best threshold	48.67	55.47	80.03	79.62	65.95
Robust threshold	49.48	55.98	80.14	79.03	**66.16**

Table 5: Ablation studies for the effects of prototype types.

Prototype	Deepglobe	ISIC	Chest	FSS	Average
Global	49.48	55.98	80.14	79.03	66.16
Regional	47.98	54.06	85.48	82.80	67.58
Pixel	47.65	51.03	84.55	83.64	66.72
HPM	49.71	55.73	85.44	82.67	**68.39**

Table 6: Results of combination with VFM-based methods.

	Deepglobe	ISIC	Chest	FSS	Average
GF-SAM [51]	49.05	44.47	38.70	85.75	54.49
GF-SAM + Ours	47.26	48.24	61.46	85.98	**60.74**
FSSAM [45]	40.25	47.75	77.11	80.36	61.37
FSSAM + Ours	47.67	43.59	84.36	85.23	**65.21**
Ours	49.71	55.73	85.44	82.67	**68.39**

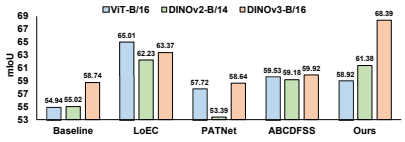

Fig. 5: Performance of CD-FSS methods with different pretrained models.

Effects of Threshold Types. To validate the effectiveness of the robust threshold in ASE, we compare the performance of different threshold types under 1-shot setting, as summarized in Tab. 4. First, the attention matrix without adding a mask may lead to information confusion, resulting in limited performance improvement. Simply using a fixed threshold of 0.5 to localize coarse foreground and background regions in the query image achieves further gains. Employing the proposed robust threshold better adapts to different target domains, achieving the best performance.

Effects of Prototype Types. As shown in Tab. 5, we conduct ablation studies under the 1-shot setting to validate the necessity of diverse prototypes. Specifically, we first extract and match only a single type of prototype, and the results indicate that different domains favor different prototype types. By integrating the matching results of diverse prototypes, our method achieves a more balanced and improved performance across different target domains.

More ablation studies can be found in the supplementary material.

4.4 More Analysis

Combination with VFM-based Methods. Existing VFM-based methods [45, 51] typically leverage features from models like DINOv2 [30] to compute support-query similarity as a prior for SAM [20, 32], reformulating cross-image matching segmentation into intra-image semantic partitioning at which SAM excels. The similarity maps computed by our method can also naturally serve as priors to combine with these methods. As shown in Tab. 6, our method significantly improves the performance of existing VFM-based methods on CD-FSS. However, the combined performance is inferior to that using our original priors alone, revealing the limitations of SAM in cross-domain scenarios.

Effects of pretrained models on CD-FSS. We incorporate DINOv2 [30] and DINOv3 [34] into three different CD-FSS paradigms to explore the impact

Table 7: Effects of different modules on cross-image semantic discriminability.

	Deepglobe	ISIC	Chest	FSS
Baseline	-0.0082	0.2471	0.2381	0.2724
Baseline+SAFR	0.0848	0.4112	0.2750	0.4130
Baseline+SAFR+ASE	0.4474	1.1880	0.5941	1.4488

Table 8: Model efficiency and performance under 1-shot setting.

Method	T. Params (M)	FLOPs (G)	FPS	mean-IoU
LoEC [28]	72.82	150.1	52.91	65.01
Matcher [26]	0	727.6	0.47	61.75
Ours	0	161.9	47.62	68.39

of stronger pretrained models on CD-FSS, as shown in Fig. 5. The baseline results approximately reflect the potential capability of models. For LoEC [28], which treats the backbone as trainable parameters, performance decreases after introducing VFMs because the rich prior accelerates overfitting. Methods freezing the pretrained model (PATNet [21] and ABCDFSS [16]) degrade with DINOv2 but improve slightly with DINOv3. We argue that VFMs still accelerate overfitting of trainable parameters in these methods while providing better feature representations during inference. When VFM priors are sufficiently strong, the benefits during inference outweigh the overfitting effects in training, leading to marginal improvements. In contrast, our training-free method completely avoids overfitting risks and fully exploits the well-learned knowledge in VFMs.

Semantic Discriminability. Semantic discriminability of features is crucial for effective foreground-background segmentation. As shown in Fig. 4(b), we perform PCA on the standard final output $\mathbf{F}_{last}$ and the fused feature $\bar{\mathbf{F}}$ obtained by SAFR, visualizing the first three components. The results demonstrate that our SAFR generates representations with enhanced semantic discriminability. Furthermore, cross-image semantic discriminability has a more direct impact on CD-FSS performance. Therefore, we compute the Fisher discriminant ratio between query foreground and support background, as well as between query foreground and support foreground, and use the difference between the two to measure cross-image semantic discriminability. As shown in Tab. 7, both SAFR and ASE effectively enhance the cross-image semantic discriminability.

Model Efficiency. We compare the model efficiency with the training-based state-of-the-art LoEC [28] and the training-free state-of-the-art Matcher [26], including the number of trainable parameters (T. Params), computational complexity (FLOPs), and inference speed (FPS), as shown in Tab. 8. Compared to LoEC, our model requires no additional training cost and achieves significantly better performance with comparable inference speed. Although Matcher is training-free, its complex multi-model architecture and pipeline result in prohibitive inference speed and inferior performance.

5 Conclusion

In this paper, we propose a training-free framework for CD-FSS to avoid the training overhead and overfitting risks, which addresses CD-FSS challenges by enhancing feature semantic discriminability and performing robust matching. Specifically, we propose an SAFR module to identify and re-fuse semantic-aware features, generating representations with enhanced semantic discriminability. To

improve semantic consistency between support and query features, we propose an ASE module that adaptively locates robust query regions and aggregates query information from corresponding regions for support features. We also propose an HPM module that adapts to semantic complexity variations across different domains by integrating matching results from diverse prototypes. Extensive experiments demonstrate that our method achieves state-of-the-art performance on four datasets with different domain gaps without any training.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under the Grants No. 62371235 and No. U25A20444, in part by the Key Research and Development Plan of Jiangsu Province under Grant No. BE2023008-2.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Tech. rep., Stanford (2006)
2. Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE transactions on medical imaging **33**(2), 577–590 (2013)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chang, S., Zhang, L., Lu, H.: High-performance few-shot segmentation with foundation models: An empirical study. arXiv preprint arXiv:2409.06305 (2024)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
7. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 172–181 (2018)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)

10. Fan, Q., Liu, K., Liu, N., Cholakkal, H., Anwer, R.M., Li, W., Gao, Y.: Adapting in-domain few-shot segmentation to new domains without source domain retraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21429–21439 (2025)
11. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: European conference on computer vision. pp. 701–719. Springer (2022)
12. Fu, Y., Lou, M., Yu, Y.: Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19077–19087 (2025)
13. Gu, H., Pei, G., Sun, Z., Ren, M., Shu, X., Yao, Y., Shen, F.: Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42755–42764 (2026)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. He, W., Zhang, Y., Zhuo, W., Shen, L., Yang, J., Deng, S., Sun, L.: Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23762–23772 (2024)
16. Herzog, J.: Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23605–23615 (2024)
17. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging **33**(2), 233–245 (2013)
18. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24905–24916 (2025)
19. Kan, M., Shan, S., Chen, X.: Multi-view deep network for cross-view classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4847–4855 (2016)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
21. Lei, S., Zhang, X., He, J., Chen, F., Du, B., Lu, C.T.: Cross-domain few-shot semantic segmentation. In: European conference on computer vision. pp. 73–90. Springer (2022)
22. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8334–8343 (2021)
23. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2869–2878 (2020)
24. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: Meddinov3: How to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379 (2025)

25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
26. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023)
27. Liu, Y., Xiao, H., Chai, J., Zhang, Y., Wang, R., Meng, Z., Luo, Z.: Synpo: Boosting training-free few-shot medical segmentation via high-quality negative prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–603. Springer (2025)
28. Liu, Y., Zou, Y., Li, Y., Li, R.: The devil is in low-level features for cross-domain few-shot segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4618–4627 (2025)
29. Nie, J., Xing, Y., Zhang, G., Yan, P., Xiao, A., Tan, Y.P., Kot, A.C., Lu, S.: Cross-domain few-shot segmentation via iterative support-query correspondence mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3380–3390 (2024)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
31. Peng, B., Tian, Z., Wu, X., Wang, C., Liu, S., Su, J., Jia, J.: Hierarchical dense correlation distillation for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23641–23651 (2023)
32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
34. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
35. Su, J., Fan, Q., Pei, W., Lu, G., Chen, F.: Domain-rectifying adapter for cross-domain few-shot segmentation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 24036–24045 (2024)
36. Sun, S., Gu, H., Xie, C., Ren, Y., Ren, M., Zhang, H.: Bridging granularity gaps: Hierarchical semantic learning for cross-domain few-shot segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9215–9223 (2026)
37. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Ding, E., Wang, J., Li, Z.: Vrp-sam: Sam with visual reference prompt. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23565–23574 (2024)
38. Tong, J., Ma, R., Zou, Y., Chen, G., Li, Y., Li, R.: Adapter naturally serves as decoupler for cross-domain few-shot semantic segmentation. arXiv preprint arXiv:2506.07376 (2025)
39. Tong, J., Zou, Y., Chen, G., Li, Y., Li, R.: Self-disentanglement and re-composition for cross-domain few-shot segmentation. arXiv preprint arXiv:2506.02677 (2025)
40. Tong, J., Zou, Y., Li, Y., Li, R.: Lightweight frequency masker for cross-domain few-shot semantic segmentation. Advances in Neural Information Processing Systems **37**, 96728–96749 (2024)

41. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
42. Wang, J., Zhang, B., Pang, J., Chen, H., Liu, W.: Rethinking prior information generation with clip for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3941–3951 (2024)
43. Xu, Q., Lin, G., Loy, C.C., Long, C., Li, Z., Zhao, R.: Eliminating feature ambiguity for few-shot segmentation. In: European Conference on Computer Vision. pp. 416–433. Springer (2024)
44. Xu, Q., Zhao, W., Lin, G., Long, C.: Self-calibrated cross attention network for few-shot segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 655–665 (2023)
45. Xu, Q., Zhu, L., Liu, X., Lin, G., Long, C., Li, Z., Zhao, R.: Unlocking the power of sam 2 for few-shot segmentation. arXiv preprint arXiv:2505.14100 (2025)
46. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 763–778. Springer (2020)
47. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
48. Yang, S., Wang, H., Xing, Z., Chen, S., Zhu, L.: Segdino: An efficient design for medical and natural image segmentation with dino-v3. arXiv preprint arXiv:2509.00833 (2025)
49. Yu, H., Cho, Y., Kang, B., Moon, S., Kong, K., Kang, S.J.: Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In: European Conference on Computer Vision. pp. 92–110. Springer (2024)
50. Yuan, J., Ye, J., Chen, W., Gao, C.: Ad-dinov3: Enhancing dinov3 for zero-shot anomaly detection with anomaly-aware calibration. arXiv preprint arXiv:2509.14084 (2025)
51. Zhang, A., Gao, G., Jiao, J., Liu, C., Wei, Y.: Bridge the points: Graph-based few-shot segment anything semantically. Advances in Neural Information Processing Systems **37**, 33232–33261 (2024)
52. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5217–5226 (2019)
53. Zhang, J.W., Sun, Y., Yang, Y., Chen, W.: Feature-proxy transformer for few-shot segmentation. Advances in neural information processing systems **35**, 6575–6588 (2022)
54. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
55. Zhu, Y., Zhang, H.: Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 326–336. Springer (2025)