Hierarchical Spatial and Channel Aggregation for Cross-domain Few-shot Segmentation

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 learn generalizable segmentation capability from abundant annotated samples in the source domain, enabling accurate segmentation of novel classes in the target domain with only a few annotated samples. Existing CD-FSS methods mainly focus on mitigating feature distribution shifts caused by style gaps while ignoring significant differences in class semantic granularity and discriminative attributes across domains, leading to two key degradations in support-query matching: semantic over-alignment and attribute over-alignment. To this end, we propose the Dual Hierarchical Aggregation Network (DHANet), which comprises three key modules. First, the Hierarchical Spatial Aggregation (HSA) module performs multi-scale region aggregation of pixel features along the spatial dimension, generating hierarchical semantic-enhanced features to alleviate semantic over-alignment. Additionally, the HCA module conducts multi-scale attribute aggregation along the channel dimension, generating hierarchical attribute-enhanced features to mitigate attribute over-alignment. Finally, we propose the Online Probabilistic Semantic Bank (OPSB), which progressively constructs and updates class probability distributions from query predictions during inference, and samples multiple pseudo-prototypes as additional support information to mitigate insufficient support. Extensive experiments on four target-domain datasets demonstrate that our method achieves state-of-the-art performance.

Keywords: Few-shot Segmentation · Cross-domain Learning · Hierarchical Feature Aggregation

1 Introduction

Driven by robust network designs [11–13, 46] and large-scale densely annotated datasets [22, 51], semantic segmentation has achieved remarkable progress in recent years. However, when confronted with novel classes that have only a few annotated samples, the performance of these models often degrades sharply,

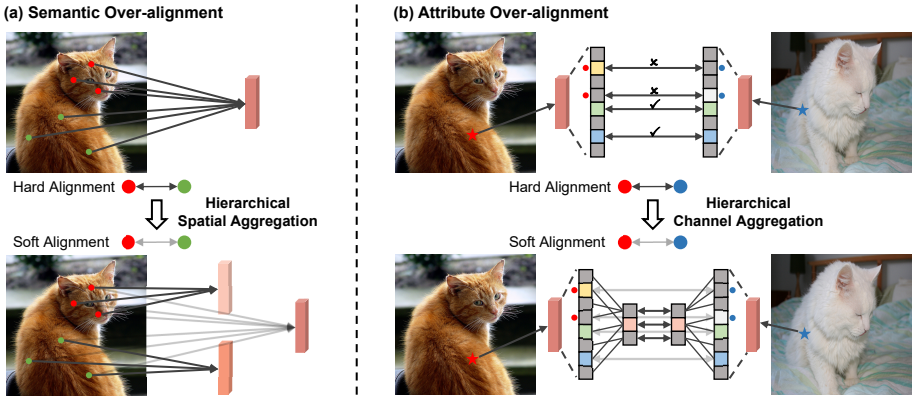

Fig. 1: Motivation of DHANet. (a) Semantic over-alignment prevents the model from distinguishing semantic categories at other segmentation granularities (*e.g.*, a cat's head and body). Hierarchical spatial aggregation aligns features at multiple granularities, adapting to target domains with varying granularities. (b) Attribute over-alignment causes the degradation of source-insensitive attributes. Hierarchical channel aggregation mitigates hard channel-wise alignment, preserving sensitivity to diverse attributes.

failing to meet the demand for rapid deployment in real-world scenarios. To alleviate the reliance on large-scale annotated data, Few-shot Segmentation (FSS) has been proposed [10, 27, 34, 44]. It leverages meta-learning to learn transferable category correspondence on base classes, enabling the segmentation of novel classes using only a few annotated samples. Conventional FSS methods inherently assume that base and novel classes share the same data domain. However, in realistic scenarios, substantial domain gaps typically exist between the source and target domains, leading to significant performance drops of existing FSS models. This practical challenge has motivated Cross-domain Few-shot Segmentation (CD-FSS) [19], which has attracted increasing attention.

Existing CD-FSS methods typically focus on mitigating cross-domain distribution shifts caused by style gaps. Prevalent techniques include performing style perturbations [24, 32], extracting domain-agnostic features [15, 19, 35], or fine-tuning with limited target domain annotated data [28, 31]. Although effective in reducing global feature distribution gaps between source and target domains, these strategies are ignoring substantial category differences across domains, which leads to support–query over-alignment during training, including semantic over-alignment and attribute over-alignment.

As shown in Fig. 1(a), the training objective forces all pixel features within the same source domain class to align with the class prototype, causing the model to only excel at distinguishing foreground and background at the same segmentation granularity as the source domain (*e.g.*, cat and other classes), while failing to discriminate semantic objects at other granularities (*e.g.*, a cat's head and body), *i.e.*, semantic over-alignment in the source domain. When the target

domain requires a finer or coarser segmentation granularity, the model's foreground–background discrimination capability degrades significantly. As shown in Fig. 1(b), different instances of the same class often exhibit attribute variations (*e.g.*, white and orange cats differ in the color attribute). These source-insensitive attributes may be activated by different channels in support and query features, and the incorrect matching caused by channel-wise alignment forces these channels to gradually degenerate, rendering them unable to capture such source-insensitive attributes, *i.e.*, attribute over-alignment in the source domain. When the target domain is sensitive to these attributes (*e.g.*, dermoscopic images [5] rely on color to distinguish lesions), the consistency of intra-class features across images weakens, substantially degrading segmentation performance.

Although some recent works [33, 36] have started to address semantic over-alignment, they typically rely on external models [45] or exhibit limited adaptability to target domains with varying segmentation granularities, while also ignoring the attribute over-alignment. To simultaneously alleviate semantic and attribute over-alignment, we propose the Dual Hierarchical Aggregation Network (DHANet). The core idea is to perform hierarchical aggregation along the spatial and channel dimensions, respectively, to capture semantic and attribute representations that are discriminative at different granularities. Such hierarchical representations not only convert hard alignment at a single granularity into soft alignment during training, mitigating over-alignment, but also provide multi-granularity information during testing, adapting to differences in class semantic granularity and discriminative attributes across different domains.

Specifically, our DHANet consists of three key modules. First, the Hierarchical Spatial Aggregation (HSA) module performs multi-scale region aggregation of pixel features along the spatial dimension via multiple groups of spatial slots and slot attention, producing hierarchical region prototypes. These prototypes are then used to enhance original features, generating hierarchical semantic-enhanced features. Second, the Hierarchical Channel Aggregation (HCA) module treats channels as aggregation primitives and performs aggregation along the channel dimension with multi-level channel slots, forming hierarchical attribute prototypes. These prototypes are then concatenated with original features to produce hierarchical attribute-enhanced representations. To prevent slot homogenization, we impose orthogonal constraints on the spatial and channel slots at each level. Finally, to further alleviate insufficient support information at test time, we propose an Online Probabilistic Semantic Bank (OPSB) module, which models class prototypes as probability distributions. It constructs and updates class probabilistic prototypes using high-confidence features extracted from the query predictions of each episode, and samples multiple pseudo-prototypes from the latest distributions as additional support information.

In summary, our contributions are as follows:

- We are the first to simultaneously focus on semantic over-alignment and attribute over-alignment in CD-FSS, and propose the HSA and HCA modules to aggregate discriminative semantic and attribute representations at different granularities along the spatial and channel dimensions, respectively.

- We propose an OPSB module to alleviate insufficient support information at test time, which constructs and updates class probabilistic prototypes using high-confidence query predictions and samples multiple pseudo-prototypes to provide additional support.
- Extensive experiments on four standard target-domain datasets demonstrate that our method achieves state-of-the-art performance on CD-FSS.

2 Related Works

2.1 Few-shot Segmentation

FSS aims to predict dense masks for novel classes in query images using only a few annotated support images. Early FSS methods [10, 23, 40, 47] typically summarize the support set into class-representative prototypes and perform segmentation by measuring the similarity between query features and these prototypes. To address the loss of spatial details inherent in prototype computation, matching-based methods [27,30,42,43] fully exploit support information by computing dense pixel-to-pixel correlations between support and query. Recently, several methods [1,20,39] introduce textual semantic features as additional support to alleviate the insufficiency of support information caused by large intra-class variance. Meanwhile, some methods [34,44,48] leverage the large-scale pre-training and generalization capabilities of vision foundation models [18, 29] to enhance FSS, achieving remarkable performance gains. However, these methods are not designed with cross-domain transfer in mind, resulting in limited generalization to novel domains.

2.2 Cross-domain Few-shot Segmentation

The limitations of FSS methods in cross-domain scenarios have motivated increasing attention to CD-FSS, which requires models trained on the source domain to generalize to novel classes across diverse domains. Some methods [24,32] apply style perturbations to features during source training to simulate cross-domain style variations, thereby narrowing the feature distribution gaps between source and diverse target domains. Others [15, 19, 35] aim to decompose or transform features into domain-agnostic spaces, generating domain-agnostic representations with cross-domain generalizability. Furthermore, most methods [16,28,31,37] leverage limited target annotated samples for fine-tuning to better adapt to the target domain. However, these methods ignore the substantial differences in semantic granularity and discriminative attributes across domains, leading to semantic over-alignment and attribute over-alignment. SDRC [36] decomposes features into representations focusing on distinct semantic regions to alleviate semantic over-alignment, but this single-granularity decoupling exhibits insufficient adaptability to different domains. Similarly, HSL [33] learns hierarchical semantic features with the assistance of multi-scale superpixel segmentation priors, yet its reliance on superpixel segmentation models limits its applicability. In contrast, we adaptively construct hierarchical semantic and attribute structures to simultaneously address both semantic and attribute over-alignment.

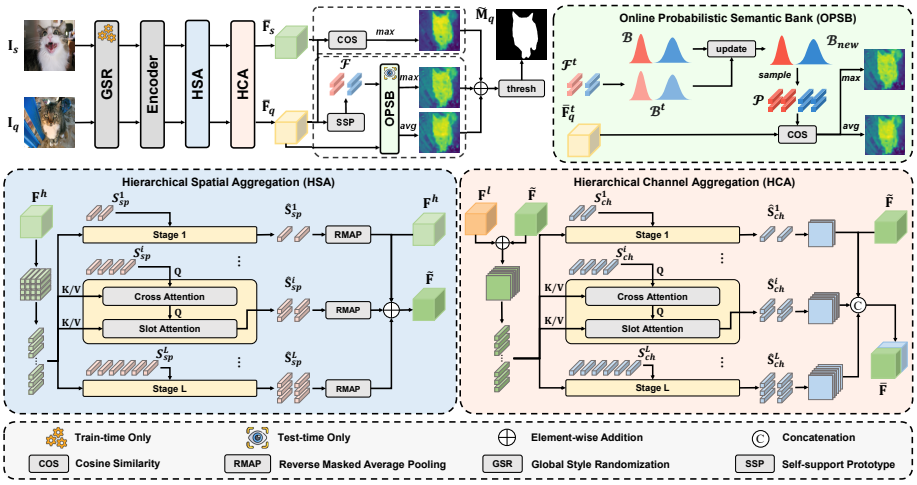

Fig. 2: Overview of our method. The support and query images are first fed into GSR for style perturbation, and features are extracted by the image encoder. Subsequently, the HSA and HCA modules sequentially perform hierarchical semantic and attribute enhancement on the features. The enhanced features are then fed into the main branch and the auxiliary branch to compute the query foreground confidence maps, respectively. Finally, the foreground confidence maps from two branches are fused and thresholded to obtain the query prediction.

3 Methodology

3.1 Problem Definition

CD-FSS aims to transfer a model trained on the source domain $\mathcal{D}_s = \{X_s, Y_s\}$ to the target domain $\mathcal{D}_t = \{X_t, Y_t\}$, where the source and target domains are disjoint in both the data distribution X and class labels Y, *i.e.*, $X_s \neq X_t$ and $Y_s \cap Y_t = \emptyset$. CD-FSS models are typically trained using the episode-based meta-learning paradigm, where the training and testing data consist of multiple episodes randomly sampled from $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. Each episode contains 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, $\mathbf{I}_s$ and $\mathbf{I}_q$ represent the support and query images, and $\mathbf{M}_s$ and $\mathbf{M}_q$ are their corresponding binary masks. The model extracts support information from the support set $\mathcal{S}$ to predict the segmentation mask for the query image $\mathbf{I}_q$.

3.2 Overview

To alleviate both semantic over-alignment and attribute over-alignment during training, as well as insufficient support information during testing, we propose the Dual Hierarchical Aggregation Network (DHANet), whose overall architecture is illustrated in Fig. 2.

Given support and query images, we first apply Global Style Randomization (GSR) to the input images using random convolution, following [24,33]. The perturbed images are then fed into a weight-shared encoder to extract the original support and query features. Next, the features are sequentially passed through Hierarchical Spatial Aggregation (HSA) and Hierarchical Channel Aggregation (HCA) modules to obtain features with hierarchical enhancement in both semantics and attributes. In the main branch, the initial foreground confidence map is obtained through dense support-query feature matching. In the auxiliary branch, the Online Probabilistic Semantic Bank (OPSB) module updates class probability distributions and samples pseudo-prototypes from them to provide additional support information for computing an auxiliary foreground confidence map. Finally, the confidence maps from two branches are combined, and the final query prediction is obtained by adaptive thresholding in [33]. Note that GSR is used only during training, and the auxiliary branch is used only during testing.

3.3 Hierarchical Spatial Aggregation

The core idea of HSA is to generate hierarchical semantic-enhanced features that are inter-class discriminative and intra-class consistent at different semantic granularities. During source domain training, these hierarchical representations enable support and query to align at multiple semantic granularities, thereby mitigating over-alignment at a single granularity. During target domain testing, these representations enhance regional consistency at different scales, preventing interference from noise. Motivated by slot attention [25, 26], which aggregates features from specific image regions via a zero-sum game mechanism for unsupervised region partitioning, we introduce a hierarchical spatial aggregation process. Specifically, we employ varying numbers of slots at each level to aggregate semantic features at diverse granularities.

Specifically, HSA is designed as a hierarchical architecture with L_s stages. At the i-th stage, we allocate a set of learnable spatial slots $\mathbf{S}_{sp}^{i} \in \mathbb{R}^{N_{sp}^{i} \times C}$, where N_{sp}^{i} denotes the number of spatial slots at this stage and C is the channel depth. Given the deep high-level feature $\mathbf{F}^{h} \in \mathbb{R}^{H \times W \times C}$ extracted by the image encoder, where H and W denote the feature height and width, respectively, we flatten the spatial dimensions to obtain $\mathbf{F}^{h} \in \mathbb{R}^{HW \times C}$. First, we aggregate initial information for the slots via cross-attention, employing $\mathbf{S}_{sp}^{i}$ as queries and $\mathbf{F}^{h}$ as keys and values:

$$\tilde{\mathbf{S}}_{sp}^{i} = \mathrm{CrossAttn}(\mathbf{S}_{sp}^{i}, \mathbf{F}^{h}), \tag{1}$$

where $\tilde{\mathbf{S}}_{sp}^{i}$ denotes the enhanced spatial slots. Then, $\tilde{\mathbf{S}}_{sp}^{i}$ is fed into slot attention for iterative updating:

$$\hat{\mathbf{S}}_{sp}^{i}, \hat{\mathbf{A}}_{sp}^{i} = \mathrm{SlotAttn}(\tilde{\mathbf{S}}_{sp}^{i}, \mathbf{F}^{h}), \tag{2}$$

where $\hat{\mathbf{S}}_{sp}^{i} \in \mathbb{R}^{N_{sp}^{i} \times C}$ represents region prototypes aggregated from different spatial regions of $\mathbf{F}^{h}$, and $\hat{\mathbf{A}}_{sp}^{i} \in \mathbb{R}^{N_{sp}^{i} \times HW}$ denotes the spatial attention maps

between $\hat{\mathbf{S}}_{sp}^{i}$ and $\mathbf{F}^{h}$. To obtain the spatial region corresponding to each region prototype, we determine the index of the maximum value of $\hat{\mathbf{A}}_{sp}^{i}$ along the slot dimension N_{sp}^{i}, yielding the spatial assignment masks $\mathbf{M}_{sp}^{i} \in \mathbb{R}^{N_{sp}^{i} \times H \times W}$ for each region prototype.

To apply each region prototype to its associated pixels, we employ the reverse process of masked average pooling (RMAP) to obtain a region feature map $\mathbf{F}^{i}$. This process is formulated as:

$$\mathbf{F}^{i}(x, y) = \sum\nolimits_{j}^{N_{sp}^{i}} \hat{\mathbf{S}}_{sp}^{ij} \mathbf{M}_{sp}^{ij}(x, y), \tag{3}$$

where (x, y) denotes the spatial coordinate, $\hat{\mathbf{S}}_{sp}^{ij} \in \mathbb{R}^{C}$ is the j-th region prototype at stage i, and $\mathbf{M}_{sp}^{ij}$ is its corresponding spatial assignment mask.

Finally, we enhance $\mathbf{F}^{h}$ with region feature maps from all stages to obtain the hierarchical semantic-enhanced feature $\tilde{\mathbf{F}}$:

$$\tilde{\mathbf{F}} = \mathbf{F}^{h} + \frac{1}{L_s} \sum\nolimits_{i}^{L_s} \mathbf{F}^{i}. \tag{4}$$

3.4 Hierarchical Channel Aggregation

Intra-class instances may exhibit attribute variations, meaning the class is insensitive to certain attributes. As a result, the same semantic region may be activated by different channels across support and query features, causing incorrect matching under channel-wise alignment. To address this, we propose an HCA module, which performs hierarchical aggregation along the channel dimension to group channels that activate the same region into hierarchical attribute prototypes. These prototypes are then used for support–query alignment, mitigating the adverse effect of hard channel-wise matching.

HCA follows a process similar to HSA. It is also designed as a hierarchical architecture with L_c stages, while the aggregation is performed along the channel dimension rather than the spatial dimension. Specifically, at the i-th stage, we allocate a set of learnable channel slots $\mathbf{S}_{ch}^{i} \in \mathbb{R}^{N_{ch}^{i} \times HW}$, where N_{ch}^{i} is the number of channel slots at this stage. Given the feature $\tilde{\mathbf{F}}$, we first combine it with the shallow feature $\mathbf{F}^{l} \in \mathbb{R}^{H \times W \times C}$ to produce a more detail-rich feature $\hat{\mathbf{F}} = \tilde{\mathbf{F}} + \mathbf{F}^{l}$. We then flatten its spatial dimensions and transpose it to obtain $\hat{\mathbf{F}} \in \mathbb{R}^{C \times HW}$. Following the same steps as Eq. (1) and Eq. (2), we obtain attribute prototypes $\hat{\mathbf{S}}_{ch}^{i} \in \mathbb{R}^{N_{ch}^{i} \times HW}$ aggregated from the channels of $\hat{\mathbf{F}}$:

$$\tilde{\mathbf{S}}_{ch}^{i} = \mathrm{CrossAttn}(\mathbf{S}_{ch}^{i}, \hat{\mathbf{F}}), \quad \hat{\mathbf{S}}_{ch}^{i}, \hat{\mathbf{A}}_{ch}^{i} = \mathrm{SlotAttn}(\tilde{\mathbf{S}}_{ch}^{i}, \hat{\mathbf{F}}). \tag{5}$$

To avoid undermining the attribute diversity of features, we do not apply attribute prototypes to their corresponding channels using the region-filling strategy in HSA. In contrast, we concatenate $\hat{\mathbf{S}}_{ch}^{i}$ from all stages with the feature $\tilde{\mathbf{F}}$ along the channel dimension to obtain the hierarchical attribute-enhanced feature $\bar{\mathbf{F}}$:

$$\bar{\mathbf{F}} = \left[\tilde{\mathbf{F}}; \hat{\mathbf{S}}_{ch}^{1}; \hat{\mathbf{S}}_{ch}^{2}; \ldots; \hat{\mathbf{S}}_{ch}^{L_c}\right] \in \mathbb{R}^{C' \times HW}, \quad C' = C + \sum\nolimits_{i=1}^{L_c} N_{ch}^{i}, \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation along the channel dimension. Finally, we restore the spatial dimensions of $\bar{\mathbf{F}}$ and transpose it to obtain the final hierarchical attribute-enhanced feature $\bar{\mathbf{F}} \in \mathbb{R}^{H\times W\times C'}$.

3.5 Online Probabilistic Semantic Bank

The limited labeled support data available at test time offers restricted class information. Some works [4,19] perform continuous fine-tuning using query predictions from each episode during testing to exploit additional class information in query images. However, such continuous fine-tuning entails the risk of knowledge forgetting. Instead, we avoid per-episode fine-tuning and dynamically accumulate reliable class priors to construct a semantic bank. An intuitive method is to progressively maintain and update a single class prototype using query predictions. However, for target domain classes with large intra-class variations, query features may still exhibit a considerable distance from the accurate prototype. We overcome this limitation by establishing a probabilistic semantic bank.

Specifically, we model class prototypes as Gaussian distributions. Given a set of high-confidence features $\mathcal{F} = \{\mathbf{f}^i\}_{i=1}^N$, we estimate distribution parameters as:

$$\boldsymbol{\mu} = \frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{f}^i, \quad \boldsymbol{\sigma}^2 = \frac{1}{N-1}\sum\nolimits_{i=1}^{N}(\mathbf{f}^i - \boldsymbol{\mu})^2, \tag{7}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ denote the mean and variance, respectively, and N represents the number of samples.

During testing, we maintain an online semantic bank for each class:

$$\mathcal{B} = \{\{\boldsymbol{\mu}_{fg}, \boldsymbol{\sigma}^2_{fg}, N_{fg}\}, \{\boldsymbol{\mu}_{bg}, \boldsymbol{\sigma}^2_{bg}, N_{bg}\}\}. \tag{8}$$

For the t-th episode, we first obtain the support feature $\bar{\mathbf{F}}^t_s$ and query feature $\bar{\mathbf{F}}^t_q$. High-confidence query foreground feature set $\mathcal{F}^t_{fg}$ and background feature set $\mathcal{F}^t_{bg}$ are then extracted via SSP [10]. Subsequently, we estimate the episode-specific foreground distribution $\{\boldsymbol{\mu}^t_{fg}, (\boldsymbol{\sigma}^t_{fg})^2, N^t_{fg}\}$ using Eq. (7) and fuse it with the existing statistics in the semantic bank:

$$\begin{aligned} N_{new} &= N_{fg} + N^t_{fg}, \quad \boldsymbol{\mu}_{new} = \frac{N_{fg}\boldsymbol{\mu}_{fg} + N^t_{fg}\boldsymbol{\mu}^t_{fg}}{N_{new}}, \\ \boldsymbol{\sigma}^2_{new} &= \frac{N_{fg}(\boldsymbol{\sigma}^2_{fg} + (\boldsymbol{\mu}_{fg} - \boldsymbol{\mu}_{new})^2) + N^t_{fg}((\boldsymbol{\sigma}^t_{fg})^2 + (\boldsymbol{\mu}^t_{fg} - \boldsymbol{\mu}_{new})^2)}{N_{new}}. \end{aligned} \tag{9}$$

Finally, the semantic bank is updated with the fused statistics:

$$\{\boldsymbol{\mu}_{fg}, \boldsymbol{\sigma}^2_{fg}, N_{fg}\} \leftarrow \{\boldsymbol{\mu}_{new}, \boldsymbol{\sigma}^2_{new}, N_{new}\}. \tag{10}$$

The background distribution is updated in the same manner.

3.6 Segmentation

During testing, given the query feature $\bar{\mathbf{F}}_q$, support feature $\bar{\mathbf{F}}_s$, and support mask $\mathbf{M}_s$, we compute the query foreground confidence map via two branches.

In the main branch, we compute the cosine similarity between $\bar{\mathbf{F}}_q$ and all foreground support features, and take the maximum along the support dimension to obtain the foreground similarity map $\mathbf{M}_{fg}$:

$$\mathbf{M}_{fg}(\mathbf{u}_q) = \max_{\mathbf{u}_s} \cos(\bar{\mathbf{F}}_q(\mathbf{u}_q), \mathbf{M}_s(\mathbf{u}_s)\bar{\mathbf{F}}_s(\mathbf{u}_s)), \tag{11}$$

where $\mathbf{u}_s = (x_s, y_s)$ and $\mathbf{u}_q = (x_q, y_q)$ denote spatial coordinates. The background similarity map $\mathbf{M}_{bg}$ is obtained in the same manner, and the query foreground confidence map is further computed as $\mathbf{M}_{conf} = \mathbf{M}_{fg} - \mathbf{M}_{bg}$.

In the auxiliary branch, we obtain the foreground distribution $\mathcal{N}(\boldsymbol{\mu}_{fg}, \boldsymbol{\sigma}^2_{fg})$ from the semantic bank and sample P pseudo foreground prototypes $\{\mathbf{p}^i_{fg}\}^P_{i=1}$. We then compute the cosine similarity between $\bar{\mathbf{F}}_q$ and $\{\mathbf{p}^i_{fg}\}^P_{i=1}$. By maximizing and averaging along the prototype dimension, we obtain the maximum foreground similarity map $\mathbf{M}^{max}_{fg}$ and the average foreground similarity map $\mathbf{M}^{avg}_{fg}$:

$$\mathbf{M}^{max}_{fg} = \max_i \cos(\bar{\mathbf{F}}_q, \mathbf{p}^i_{fg}), \quad \mathbf{M}^{avg}_{fg} = \frac{1}{P}\sum\nolimits_i^P \cos(\bar{\mathbf{F}}_q, \mathbf{p}^i_{fg}). \tag{12}$$

The background similarity maps $\mathbf{M}^{max}_{bg}$ and $\mathbf{M}^{avg}_{bg}$ are computed in the same way, from which we further derive the max foreground confidence map $\mathbf{M}^{max}_{conf}$ and average foreground confidence map $\mathbf{M}^{avg}_{conf}$.

We combine the results from both branches to obtain the final foreground confidence map $\tilde{\mathbf{M}}_{conf} = \mathbf{M}_{conf} + \mathbf{M}^{max}_{conf} + \mathbf{M}^{avg}_{conf}$. Finally, we binarize $\tilde{\mathbf{M}}_{conf}$ using the adaptive threshold from [33] to produce the query prediction $\tilde{\mathbf{M}}_q$.

3.7 Loss Function

We train our model on the main branch using the Binary Cross Entropy (BCE) loss:

$$\mathcal{L}_{main} = \mathrm{BCE}(\mathbf{M}_{fg}, \mathbf{M}_{bg}, \mathbf{M}_q). \tag{13}$$

Following [33], we additionally adopt the same self-segmentation loss $\mathcal{L}_{ssp}$ as in SSP [10] to better supervise training.

To prevent the spatial slots $\hat{\mathbf{S}}^i_{sp}$ and channel slots $\hat{\mathbf{S}}^i_{ch}$ from collapsing into similar representations, we normalize all slots and impose an orthogonality constraint at each stage:

$$\mathcal{L}_{orth} = \frac{1}{L_s}\sum\nolimits_{i=1}^{L_s} \|\hat{\mathbf{S}}^i_{sp}(\hat{\mathbf{S}}^i_{sp})^T - \mathbf{I}\|^2_F + \frac{1}{L_c}\sum\nolimits_{i=1}^{L_c} \|\hat{\mathbf{S}}^i_{ch}(\hat{\mathbf{S}}^i_{ch})^T - \mathbf{I}\|^2_F. \tag{14}$$

Finally, the total loss function is $\mathcal{L} = \mathcal{L}_{main} + \mathcal{L}_{ssp} + \lambda\mathcal{L}_{orth}$, where $\lambda = 0.1$ is the weight parameter, and $\mathcal{L}_{ssp}$ is the same as that proposed in [10].

Table 1: Mean-IoU of 1-shot and 5-shot results compared with previous FSS and CD-FSS methods on the CD-FSS benchmark. The best and second-best methods are highlighted in **bold** and underlined, respectively.

Methods	Publication	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
Few-shot Semantic Segmentation Methods												
RePRI [2]	CVPR-21	Res-50	25.03	27.41	23.27	26.23	65.08	65.48	70.96	74.23	46.09	48.34
HSNet [27]	ICCV-21	Res-50	29.65	35.08	31.20	35.10	51.88	54.36	77.53	80.99	47.57	51.38
SSP [10]	ECCV-22	Res-50	40.00	48.68	35.49	45.86	74.44	74.26	78.91	80.59	57.21	62.35
FPTrans [49]	NIPS-22	ViT-base	38.36	49.30	48.65	60.37	80.92	82.91	80.74	83.65	62.17	69.06
HDMNet [30]	CVPR-23	Res-50	25.40	39.10	33.00	35.00	30.60	31.30	75.10	78.60	41.00	46.00
PerSAM [50]	ICLR-24	ViT-base	36.08	40.65	23.27	25.33	29.95	30.05	60.92	66.53	37.56	40.64
VRP-SAM [34]	CVPR-24	ViT-base	40.43	44.75	28.21	31.96	30.54	29.99	80.78	83.18	44.99	47.47
Cross-domain Few-shot Semantic Segmentation Methods												
PATNet [19]	ECCV-22	Res-50	37.89	42.97	41.16	53.58	66.61	70.20	78.59	81.23	56.06	61.99
ABCDFSS [16]	CVPR-24	Res-50	42.60	49.00	45.70	53.30	79.80	81.40	74.60	76.20	60.67	64.97
DRA [32]	CVPR-24	Res-50	41.29	50.12	40.77	48.87	82.35	82.31	79.05	80.40	60.86	65.42
APSeg [15]	CVPR-24	ViT-base	35.94	39.98	45.43	53.98	84.10	84.50	79.71	81.90	61.30	65.09
DMTNet [4]	IJCAI-24	Res-50	40.14	51.17	43.55	52.30	73.74	77.30	81.52	83.28	59.74	66.01
APM [37]	NIPS-24	Res-50	40.86	44.92	41.71	51.16	78.25	82.81	79.29	81.83	60.03	65.18
LoEC [24]	CVPR-25	ViT-base	42.12	51.48	52.91	62.43	83.94	84.12	81.05	83.69	65.01	70.43
SDRC [36]	ICML-25	ViT-base	43.15	46.83	46.57	55.02	82.86	84.79	80.31	82.55	63.22	67.30
DFN [35]	ICML-25	ViT-base	39.45	47.67	50.36	58.53	83.18	**87.14**	82.97	85.72	63.99	69.77
ISA [9]	ICCV-25	Res-50	44.30	52.70	37.20	56.10	83.40	86.30	78.80	86.00	60.90	70.30
HSL [33]	AAAI-26	ViT-base	45.77	54.56	59.36	64.62	85.95	86.25	81.89	83.84	68.24	72.32
DHANet(Ours)	Ours	ViT-base	**49.79**	**56.13**	**64.25**	**67.38**	**86.06**	86.87	**83.61**	**86.67**	**70.93**	**74.26**

4 Experiments

4.1 Experiment Setup

Datasets and Metric. Following the experimental setup of PATNet [19], we use the PASCAL VOC 2012 [8] dataset augmented with SBD [14] as the source domain for training. Subsequently, the trained model is evaluated on four target datasets: Deepglobe [6], ISIC [5, 38], Chest X-ray [3, 17], and FSS-1000 [21]. Deepglobe [6] is a satellite remote sensing dataset comprising dense annotations for 7 categories: areas of urban, agriculture, rangeland, forest, water, barren, and unknown. ISIC [5,38] is a dermoscopic skin lesion dataset covering 3 major types of skin lesions. Chest X-ray [3,17] is a tuberculosis diagnosis X-ray dataset, where the grayscale images exhibit substantial style gaps compared to the natural images. FSS-1000 [21] contains 1,000 daily object categories with 10 annotated images per class.

We adopt mean Intersection over Union (mIoU) as the evaluation metric and report the average results over 5 runs with different random seeds.

Implementation Details. Consistent with previous works [24, 33], we adopt ViT-B/16 [7] as our backbone and resize all images to 480×480. During training, we follow FPTrans [49] to augment all images. We use SGD to optimize our model for 5 epochs, with a momentum of 0.9, a weight decay of 5e-4, a batch size of 6, and a constant learning rate of 1e-3. Following most works, we also perform fine-tuning using the support set from the first episode of each class in the target domain. During fine-tuning, we sample 12 episodes from the support set augmented using the strategy in [16] and train for one epoch. We set the

Table 2: Ablation studies for each component of our method.

HSA	HCA	OPSB	Deepglobe	ISIC	Chest	FSS-1000	Average
×	×	×	42.52	55.75	84.60	83.23	66.53
✓	×	×	45.24	59.23	86.07	83.16	68.43
×	✓	×	44.48	58.19	85.77	83.17	67.90
✓	✓	×	46.68	60.43	86.52	83.25	69.22
✓	✓	✓	49.79	64.25	86.06	83.61	**70.93**

Table 3: Effects of components in HSA and HCA.

	mean-IoU
All	**69.22**
w/o SlotAttn	67.39
w/o CrossAttn	67.81
w/o $\mathcal{L}_{orth}$	68.33

number of aggregation stages L_s in HSA to 2, with corresponding slot numbers of $(10, 20)$. Similarly, the number of stages L_c in HCA is set to 2 with $(32, 64)$ slots. The number of sampled pseudo-prototypes P in OPSB is set to 20. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

4.2 Comparison with State-of-the-art Methods

In Tab. 1, we compare our method with existing FSS and CD-FSS methods, achieving remarkable performance advantages under both 1-shot and 5-shot settings. Specifically, our method surpasses the state-of-the-art method HSL [33] by 2.69% and 1.94% under the 1-shot and 5-shot settings, respectively, and our 1-shot performance even exceeds the 5-shot performance of all other methods except HSL. Moreover, our method shows more pronounced advantages on target domains with larger category differences from the source domain. For instance, the 1-shot performance on Deepglobe and ISIC surpasses HSL by 4.02% and 4.89%, respectively. These results clearly demonstrate the effectiveness of our method. Additionally, we conduct a fair comparison with some CD-FSS methods under the setup of IFA [28]. Please refer to our supplementary material for more details. We also present some qualitative results of our method for 1-way 1-shot setting in Fig. 3(a).

4.3 Ablation Studies

Effects of Each Component. As shown in Tab. 2, we analyze each component under the 1-shot setting to validate the effectiveness of each design. The baseline builds on SSP [10] by applying global style randomization from [24] and target-domain fine-tuning to alleviate style gaps, and adopts the adaptive thresholding strategy in HSL for segmentation. HSA mitigates semantic over-alignment by constructing hierarchical semantic features, improving the baseline's average performance by 1.90%. HCA alleviates attribute over-alignment caused by hard channel-wise alignment via hierarchical channel aggregation, yielding a 1.37% average gain over the baseline. HSA and HCA are complementary, jointly bringing a 2.69% mIoU gain. In addition, OPSB extracts high-confidence features from the query prediction and builds a class semantic bank to provide extra support information, further improving performance by 1.71%. These results fully demonstrate the effectiveness of each component.

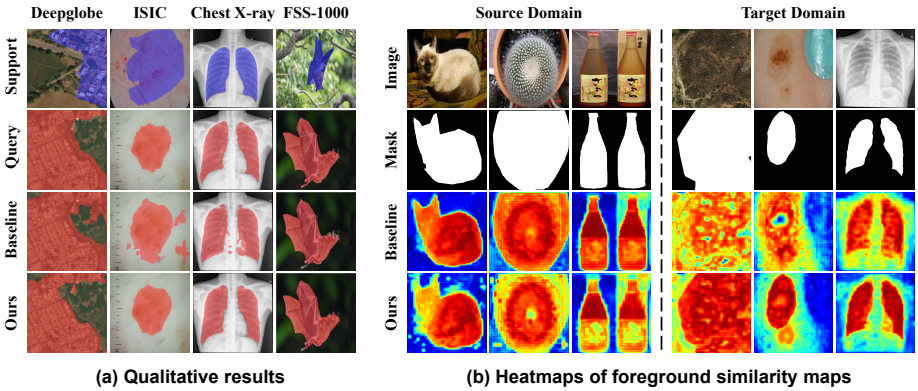

Fig. 3: (a) Qualitative results of our model for 1-way 1-shot setting. (b) The heatmaps of the foreground similarity maps show that our method can extract hierarchical semantic features to alleviate semantic over-alignment.

Table 4: Effect of the number of aggregation stages in HSA.

L_s	10	20	40	80	mean-IoU
1	✓	×	×	×	67.92
2	✓	✓	×	×	68.43
3	✓	✓	✓	×	**68.44**
4	✓	✓	✓	✓	68.11

Table 5: Effect of the number of aggregation stages in HCA.

L_c	32	64	128	256	mean-IoU
1	✓	×	×	×	68.84
2	✓	✓	×	×	**69.22**
3	✓	✓	✓	×	69.04
4	✓	✓	✓	✓	68.93

Effects of Components in HSA and HCA. We conduct ablation studies to investigate the importance of each component in the aggregation module, as shown in Tab. 3. Slot attention partitions features into distinct groups and aggregates information from them. Without it, there is no guarantee that individual slots aggregate distinct features, leading to performance degradation. Cross attention injects sample priors into the slots, enabling the aggregated slots to establish support–query correspondences. Removing cross attention causes unstable channel alignment in HCA, thus compromising performance. The orthogonal loss is applied to encourage the model in capturing differences among image features corresponding to different slots. Discarding this loss results in limited semantic discriminability for the features and a subsequent drop in performance. The results demonstrate the importance of these components.

Number of Aggregation Stages. We conduct ablation studies under the 1-shot setting to validate the necessity of multi-level aggregation. Specifically, we gradually increase the number of aggregation stages in HSA and HCA, respectively, with the number of slots increasing progressively at each stage. The results are shown in Tab. 4 and Tab. 5. The best performance in HSA is achieved with 3 stages of spatial aggregation. However, since 2 stages yield almost identical performance with lower computational overhead, we set the number of aggregation

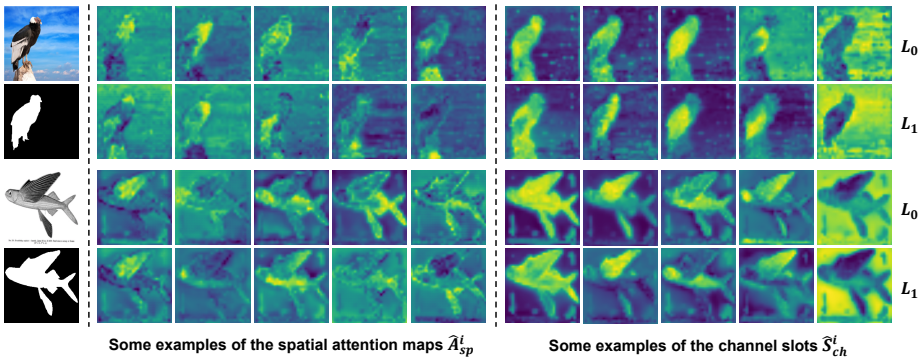

Fig. 4: Visualization of spatial and channel slots demonstrates that multi-stage slots facilitate constructing features with hierarchical semantic and attribute enhancement.

Table 6: Comparison of different query prediction utilization strategies.

	Deepglobe	ISIC	Chest	FSS	Average
Fine-tuning	44.75	57.29	86.41	83.07	67.88
Deterministic prototype	47.87	62.89	85.78	83.26	69.95
Probabilistic prototype	49.79	64.25	86.06	83.61	**70.93**

Table 7: Model efficiency and performance under 1-shot setting.

Method	FLOPs (G)	FPS	mean-IoU
HSL [33]	186.52	27.25	68.24
Ours	176.48	30.77	70.93

stages L_s to 2. Similarly, since 2 stages of channel aggregation in HSA achieve the best performance, we also set the number of aggregation stages L_c to 2. These results demonstrate that hierarchical aggregation is necessary. However, performance does not continue to improve as the number of aggregation stages further increases. A possible reason is that excessively fine-grained aggregation and orthogonal constraints may instead disrupt the original category hierarchy, leading to negative effects.

Query Prediction Utilization Strategies. To validate the effectiveness of our probabilistic semantic bank, we compare different strategies for utilizing query information during testing, as reported in Tab. 6. Existing CD-FSS methods [4, 19] typically leverage predicted pseudo query masks to further fine-tune the model during inference. However, such online fine-tuning incurs additional computational overhead and is prone to knowledge forgetting. In contrast, directly computing pseudo prototypes from pseudo query masks and maintaining as well as updating this deterministic prototype online can provide stable additional support information. Our method instead constructs a probabilistic prototype and samples multiple pseudo prototypes, thereby offering more diverse support information and achieving significant performance gains.

4.4 Analysis and Visualizations

Model Efficiency. We compare model efficiency with the current state-of-the-art method HSL [33] under the 1-shot setting, including computational com-

plexity (FLOPs) and inference speed (FPS). As shown in Tab. 7, our method demonstrates both better model efficiency and segmentation performance.

Visualization of Slots. To validate the construction process of our hierarchical features, we visualize the spatial and channel slots at each stage. Specifically, we visualize the attention maps of spatial slots and the spatial activation maps of channel slots after restoring their spatial structure, as shown in Fig. 4. Here, L_0 denotes the coarse-grained stage (with fewer slots), and L_1 denotes the fine-grained stage. Different spatial slots attend to distinct semantic regions, with those in the fine-grained stage capturing more detailed semantics. Furthermore, channel slots activate regions with similar attributes. Since the coarse-grained channel slots aggregate more channels, their activated regions are typically larger than those of the fine-grained slots. These visualizations intuitively demonstrate that multi-stage slots facilitate the construction of features with hierarchical semantic and attribute enhancement.

Mitigating semantic over-alignment. Semantic over-alignment makes the learned features excel at distinguishing categories only at the source-domain granularity, while lacking the ability to discriminate target-domain semantics. We utilize ground truth masks to extract foreground prototypes and calculate similarity maps, with the corresponding heatmaps visualized in Fig. 3(b). The baseline overfits to extracting highly consistent features for source-domain categories, failing to effectively distinguish foreground from background in the target domain. In contrast, our method focuses on extracting more hierarchical features, enabling generalization to diverse domains.

Mitigating attribute over-alignment. Attribute over-alignment causes the degradation of source-insensitive attributes, where all channels tend to represent source-sensitive attributes, leading to attribute redundancy, *i.e.*, high inter-channel correlation. Following [41], we use the Pearson correlation coefficient to measure the degree of inter-channel redundancy. As shown in Tab. 8, our hierarchical channel aggregation mitigates channel misalignment and reduces inter-channel correlation, enabling features to capture more diverse attributes for adaptation to different domains.

Table 8: Effects of our mehod on channel correlation.

	Deepglobe	ISIC	Chest	FSS
Baseline	0.2713	0.3149	0.3307	0.2817
Ours	0.2032	0.2366	0.2257	0.2318

5 Conclusion

In this paper, we propose a Dual Hierarchical Aggregation Network (DHANet) to simultaneously alleviate semantic over-alignment and attribute over-alignment in CD-FSS. Specifically, we propose a Hierarchical Spatial Aggregation (HSA) module, which aligns features at multiple granularities through hierarchical aggregation along the spatial dimension, thereby adapting to target domains with varying segmentation granularities. Furthermore, we introduce a Hierarchical Channel Aggregation (HCA) module that mitigates the rigid channel-wise hard

alignment via hierarchical aggregation along the channel dimension, thus preserving sensitivity to diverse attributes. HSA and HCA effectively alleviate semantic and attribute over-alignment, respectively. Finally, we present an Online Probabilistic Semantic Bank (OPSB) module to address the scarcity of support information during inference. Extensive experiments demonstrate that our DHANet achieves state-of-the-art performance across multiple CD-FSS datasets.

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. Bi, H., Feng, Y., Diao, W., Wang, P., Mao, Y., Fu, K., Wang, H., Sun, X.: Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation. IEEE transactions on pattern analysis and machine intelligence **47**(1), 131–148 (2024)
2. Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13979–13988 (2021)
3. 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)
4. Chen, J., Quan, R., Qin, J.: Cross-domain few-shot semantic segmentation via doubly matching transformation. arXiv preprint arXiv:2405.15265 (2024)
5. 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)
6. 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)
7. 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)
8. 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**, 303–338 (2010)
9. 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)

10. 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)
11. 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)
12. 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)
13. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in neural information processing systems **35**, 1140–1156 (2022)
14. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 international conference on computer vision. pp. 991–998. IEEE (2011)
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. 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)
19. 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)
20. Leng, J., Zhao, Z., Tian, C., Chen, Z., Wang, G., Liu, X.: Multi-modal few-shot semantic segmentation based on triple attention mechanism and hierarchical decoding transformer. IEEE Transactions on Circuits and Systems for Video Technology (2025)
21. 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)
22. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022)
23. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 142–158. Springer (2020)
24. 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)

25. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
26. Ma, C., Yuhuan, Y., Ju, C., Zhang, F., Zhang, Y., Wang, Y.: Attrseg: open-vocabulary semantic segmentation via attribute decomposition-aggregation. Advances in neural information processing systems **36**, 10258–10270 (2023)
27. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6941–6952 (2021)
28. 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)
29. 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)
30. 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)
31. Peng, S.F., Sun, G., Li, Y., Wang, H., Xie, G.S.: Sam-aware graph prompt reasoning network for cross-domain few-shot segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6488–6496 (2025)
32. 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)
33. 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)
34. 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)
35. 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)
36. 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)
37. 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)
38. 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)
39. 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)
40. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)

41. Wang, W., Fu, C., Guo, J., Cai, D., He, X.: Cop: Customized deep model compression via regularized correlation-based filter-level pruning. arXiv preprint arXiv:1906.10337 (2019)
42. 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)
43. 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)
44. 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)
45. Xu, S., Wei, S., Ruan, T., Liao, L.: Learning invariant inter-pixel correlations for superpixel generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6351–6359 (2024)
46. Yan, H., Wu, M., Zhang, C.: Multi-scale representations by varying window attention for semantic segmentation. arXiv preprint arXiv:2404.16573 (2024)
47. 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)
48. 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)
49. 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)
50. 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)
51. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)