A Transformer-based Adaptive Prototype Matching Network for Few-Shot Semantic Segmentation

Sihan Chen[1] , **Yadang Chen**[1] , **Yuhui Zheng**[2*] , **Zhi-Xin Yang**[3] and **Enhua Wu**[4]
[1]School of Computer Science, Nanjing University of Information Science and Technology
[2]College of Computer, Qinghai Normal University
[3]State Key Laboratory of Internet of Things for Smart City, University of Macau
[4]Key Laboratory of System Software and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
{melody, adamchen}@nuist.edu.cn, zhengyh@vip.126.com, zxyang@um.edu.mo, ehwu@umac.mo

Abstract

Few-shot semantic segmentation (FSS) aims to generate a model for segmenting novel classes using a limited number of annotated samples. Previous FSS methods have shown sensitivity to background noise due to *inherent bias*, *attention bias*, and *spatial-aware bias*. In this study, we propose a Transformer-Based Adaptive Prototype Matching Network to establish robust matching relationships by improving the semantic and spatial perception of query features. The model includes three modules: target enhancement module (TEM), dual constraint aggregation module (DCAM), and dual classification module (DCM). In particular, TEM mitigates *inherent bias* by exploring the relevance of multi-scale local context to enhance foreground features. Then, DCAM addresses *attention bias* through the dual semantic-aware attention mechanism to strengthen constraints. Finally, the DCM module decouples the segmentation task into semantic alignment and spatial alignment to alleviate *spatial-aware bias*. Extensive experiments on PASCAL-5^i and COCO-20^i confirm the effectiveness of our approach.

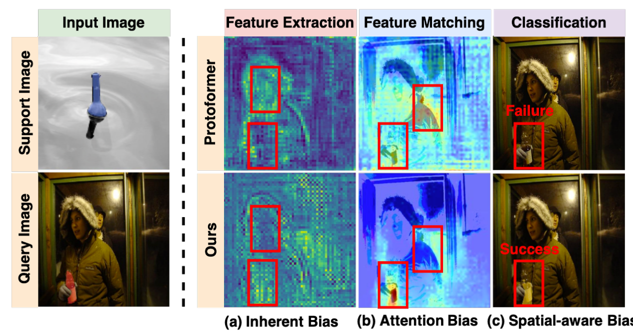

Figure 1: (a) *Inherent Bias:* During the feature extraction stage, our method effectively enhances features associated with the target class "bottle" while suppressing interference from the category "person" in the background. In contrast, the baseline erroneously focuses on the background "person". (b) *Attention Bias:* In the feature matching stage, our approach concentrates on the regions relevant to the target category "bottle", effectively mitigating matching inconsistencies caused by intra-class differences. Contrastingly, the baseline scatters attention, neglecting crucial features of the target class 'bottle' and encompassing irrelevant background regions. (c) *Spatial-aware Bias:* In the feature classification stage, our method achieves precise segmentation of the target class, while the baseline fails to segment the target object.

1 Introduction

In recent years, traditional semantic segmentation [Wang *et al.*, 2019] has made significant progress due to the rapid advancements in deep learning within the field of computer vision [Chen *et al.*, 2022b; Ye *et al.*, 2021]. However, this task has long struggled with challenges such as dense annotation requirements and limited generalization. In such case, few-shot semantic segmentation (FSS) [Tian *et al.*, 2020] has been proposed to simulate real-world scenarios with limited data and multiple categories.

FSS follows the framework of meta-learning, conducting training in the form of episodes that consist of a support set and a query set. The model execution process can be divided into three stages. First, both the support set and the query set are sent synchronously to the parameter-sharing backbone network for feature extraction. Next, in the feature matching stage, interaction occurs between the annotated support features and the query features using either a prototype-pixel mechanism [Zhang *et al.*, 2020; Zhang *et al.*, 2021a; Fan *et al.*, 2022; Liu *et al.*, 2022c; Cao *et al.*, 2022] or a pixel-pixel mechanism [Xie *et al.*, 2021; Min *et al.*, 2021; Zhang *et al.*, 2021b; Shi *et al.*, 2022]. Finally, in the classification stage, FSS predicts the segmentation mask for the target category in the query image.

Existing FSS models [Cao *et al.*, 2022; Zhang *et al.*, 2021b; Liu *et al.*, 2022c; Chen *et al.*, 2024] have shown impressive results. However, as illustrated in Fig.5, previous works still face challenges due to background interference, including adjacent regions, seen classes, and analogs. The reasons can be explained from three aspects: (i) in the fea-

*Corresponding author.

ture extraction stage, previous works [Tian *et al.*, 2020; Chen *et al.*, 2022a; Cao *et al.*, 2022] have relied on features directly extracted from the pretrained backbone networks. However, these pretrained backbones exhibit *inherent bias*. As shown in Fig.1(a), they tend to prioritize extracting features related to "person" rather than the specific target category "bottle" for the current task. (ii) In the feature matching stage, previous works [Cao *et al.*, 2022; Liu *et al.*, 2022c; Shi *et al.*, 2022] have often utilized attention mechanisms to establish the single-layer relationship between the support set and the query set for category information transfer. However, this subtle relationship proves to be insufficient when there exists significant intra-class variability within the target category, leading to *attention bias* (see in Fig.1(b)). (iii) In the classification stage, existing methods [Tian *et al.*, 2020; Zhang *et al.*, 2021a; Liu *et al.*, 2022b] predominantly rely on semantic relevance to make predictions, overlooking the exploration of spatial information, which we refer to as *spatial-aware bias*. As depicted in Fig.1(c), when faced with complex scenes, relying solely on semantic consistency may fail to accurately segment the target category "bottle."

The aforementioned concerns motivate us to introduce a novel Transformer-Based Adaptive Prototype Matching Network. This model mitigates background interference in FSS by strategically and efficiently interacting during the three stages of model execution. The central concept involves leveraging both semantic and spatial perceptions of query features to provide a more comprehensive understanding of category information, ultimately enhancing the robustness of the model.

Firstly, in the feature extraction stage, we draw inspiration from [Lin *et al.*, 2023] and introduce the target enhancement module (TEM) to alleviate *inherent bias*. TEM focuses on exploring the relevance of multi-scale local context in an computationally efficient manner to enhance foreground features. Secondly, in the feature matching stage, to address *attention bias*, we devise a dual constraint aggregation module (DCAM). This module emulates the human visual matching process, identifying the most similar target regions in the image based on prior information and using these regions as references for self-retrieval. Finally, in the feature classification stage, addressing the *spatial-aware bias*, we propose a dual classification module (DCM). This module aims to decouple the segmentation task into two subtasks: semantic alignment and spatial alignment. It leverages semantic consistency for target category identification and spatial consistency for precise localization, collectively improving performance.

In summary, our contributions are as follows:

- We propose a TEM to effectively mitigate *inherent bias*.
- We propose a DCAM to alleviate *attention bias*.
- We propose a DCM to individually achieve semantic alignment and spatial alignment, addressing the *spatial-aware bias*.
- Extensive experiments on two benchmark datasets, PASCAL-5^i and COCO-20^i, demonstrate that the proposed model outperforms other existing competitors using the same metrics.

2 Related Work

2.1 Few-Shot Semantic Segmentation

FSS predicts the mask for an unseen category by using a small number of annotated support images. Metric learning-based FSS can be divided into two main categories: prototype-based methods and pixel-based matching methods.

The prototype-based mechanism has emerged as a predominant method in FSS. The pioneering work by [Dong and Xing, 2018] introduced the prototype concept to FSS, utilizing a prototype to represent information about the target class in the support set and performing matching on query features to predict the segmentation mask. Subsequent studies extended this approach from different perspectives. Firstly, some studies [Yang *et al.*, 2020; Li *et al.*, 2021; Yang *et al.*, 2021; Zhang *et al.*, 2021a; Liu *et al.*, 2022a] proposed generating multiple foreground prototypes to fully utilize the support foreground information. Secondly, considering the potential presence of new classes in the background, [Yang *et al.*, 2021; Chen *et al.*, 2022a; Liu *et al.*, 2022b] generated one or more background prototypes by mining the query background.

Due to the inevitable loss of spatial information associated with the prototype-based approach, recent studies [Zhang *et al.*, 2019; Min *et al.*, 2021; Shi *et al.*, 2022; Zhang *et al.*, 2021b; Peng *et al.*, 2023] proposed a pixel-based matching strategy. Unlike the prototype-based method, pixel-based matching seeks to establish a dense association between the support pixels and the query pixels.

Despite the commendable performance achieved by the pixel-based matching method, it is prone to overfitting on the training set. Furthermore, the computational demands of the pixel-based matching method surpass those of the prototype-based approach. After comprehensive consideration, our approach selects a prototype-based matching scheme.

2.2 Transformer

Transformers have shown incredible success from the field of natural language processing (NLP) [Vaswani *et al.*, 2017] to computer vision (CV) [Huang *et al.*, 2022; Ouyang *et al.*, 2023] due to their ability to capture long-range correlations. Some recent works [Lu *et al.*, 2021; Zhang *et al.*, 2021b; Shi *et al.*, 2022; Cao *et al.*, 2022; Liu *et al.*, 2022c] have explored the use of transformers in FSS. [Lu *et al.*, 2021] introduced the classifier weight converter to dynamically adjust classifier weights for each query image. [Zhang *et al.*, 2021b] proposed a cyclic consistent attention mechanism to filter out task-irrelevant pixels from the support set. [Shi *et al.*, 2022] proposed dense pixel cross-query and support attention-weighted mask aggregation to predict the segmentation mask by aggregating multi-level support masks weighted by pixel relevance. [Cao *et al.*, 2022] utilized traditional vanilla attention to strengthen the discriminant of class prototypes. While these methods have been successful, they have certain limitations. The first issue is inherent bias of the backbone. Despite using self-alignment in [Zhang *et al.*, 2021b] to enhance target category features, the quadratic complexity of the input length remains a concern. The second issue is attention bias caused by insufficient constraints. [Liu *et*

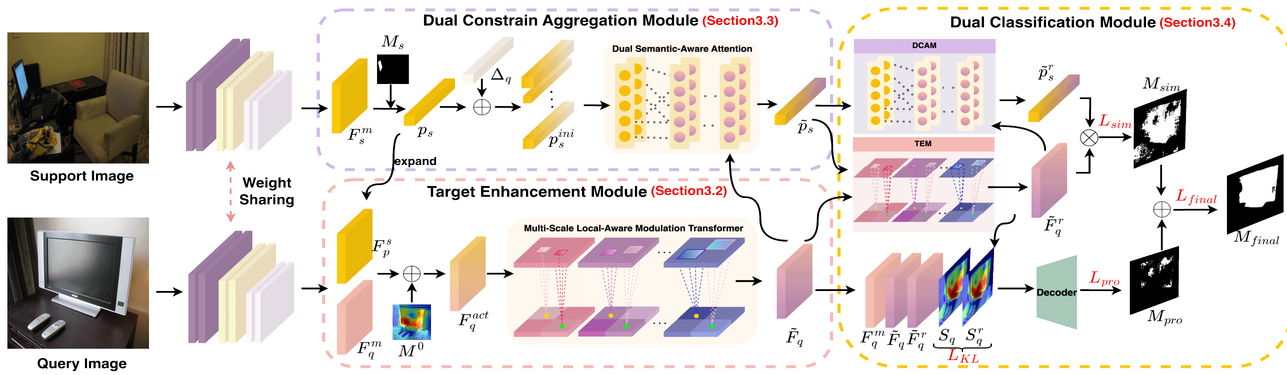

Figure 2: The overall architecture of our proposed network. First, target enhancement module(TEM) is proposed to generate enhanced query foreground feature $\tilde{F}_q$ using middle-level query features F_q^m, a support prototype feature F_p^s, and a prior mask M^0. Then, dual constraint aggregation module(DCAM) takes $\tilde{F}_q$, the support prototype p_s, and intra-class difference representation Δ_q as input to generate the discriminative category prototype $\tilde{p}_s$. Finally, in the upper half of the dual classification module(DCM), the above operations are repeated to obtain the refined category prototype $\tilde{p}_s^r$ and the refined query foreground feature $\tilde{F}_q^r$ and then the semantic similarity-based mask M_{sim} is generated by $\tilde{p}_s^r$ and $\tilde{F}_q^r$. In the lower half of the DCM module, the process starts by utilizing $\tilde{F}_q$ and $\tilde{F}_q^r$ to generate corresponding spatial distribution guidence S_q and S_q^r. Then the spatial distribution probability-based mask M_{pro} is generated by the decoder with F_q^m,$\tilde{F}_q$,$\tilde{F}_q^r$, S_q and S_q^r as input. The final query segmentation mask M_{final} is generated by combining M_{sim} and M_{pro}.

al., 2022c] employed mask attention to filter out background noise. However, in cases where there are significant variations within the same class, the accuracy of the mask is worrying.

In our work, we pioneer the integration of the convolutional transformer [Lin *et al.*, 2023] into FSS and enhance foreground features through a local adaptive strategy. Additionally, we introduce a dual semantics-aware attention mechanism to explicitly model the cross-consistency between the support set and the query set as well as the self-consistency within the query set, in order to achieve sufficient mining of the query target class information.

3 Method

3.1 Overview

In this work, we aim to mitigate the susceptibility to background noise caused by *inherent bias*, *attention bias*, and *spatial-aware bias*. To this end, we propose a novel Transformer-Based Adaptive Prototype Matching Network that establishes robust matching relationships between the support set and the query set during the three model execution phases. In the initial feature extraction stage, we introduce a target enhancement module to alleviate *inherent bias* (see in Section 3.2). Subsequently, in the feature matching stage, to address *attention bias*, we design a dual constraint aggregation module (see in Section 3.3). Lastly, in the feature classification stage, to tackle *spatial-aware bias*, we propose a dual classification module (see in Section 3.4). Without loss of generality, we demonstrate the entire network architecture in a 1-shot setting (see in Fig.2).

3.2 Target Enhancement Module

Alleviating the *inherent bias* of the backbone is crucial for FSS. Despite significant progress in recent work [Zhang *et al.*, 2021b], challenges persist due to the quadratic complexity of the input length. In TEM, to reduce computational costs, we introduce a Multi-Scale Local-Aware Modulation Transformer based on the convolutional transformer architecture [Lin *et al.*, 2023] for performing multi-scale feature extraction. Differently, we employ multi-scale self-adaptive local attention to enhance foreground information and mitigate background interference. Furthermore, we replace the standard multi-layer perception (MLP) [Vaswani *et al.*, 2017] with an invertible neural network (INN) [Dinh *et al.*, 2016] to preserve more fine-grained features in the feedforward process.

Specifically, we first follow previous work [Tian *et al.*, 2020] to get the initial activated query feature F_q^{act} using the middle-level query feature $F_q^m \in \mathbb{R}^{H\times W\times C}$, the support prototype feature $F_p^s \in \mathbb{R}^{H\times W\times C}$, and the prior mask M^0 as inputs. F_p^s is obtained from the support prototype extension, where the support prototype $p_s \in \mathbb{R}^{1\times C}$ is obtained by applying masked average pooling($\mathcal{MAP}$) on the middle-level of the support feature $F_s^m \in \mathbb{R}^{H\times W\times C}$. M^0 is obtained from high-level support and query features. Then, as shown in Fig.3, F_q^{act} is treated as the input to the Multi-Scale Local-Aware Modulation Transformer. We follow the paradigm of the multi-head attention mechanism, where features are divided into N groups $F_q^{act} = [F_{q1}^{act}, F_{q2}^{act}, ..., F_{qN}^{act}]$ by the channel dimension, and a convolutional layer $Conv_{K\times K}$ with a kernel size of $K \times K$ is applied to each group of features $F_{qi}^{act} \in \mathbb{R}^{H\times W\times (C/N)}$ to generate local attention weights $Attn(F_{qi}^{act}) \in \mathbb{R}^{H\times W\times (K\times K)}$. The weights are then normalized by a softmax function. In turn, the weighted features $\widehat{F}_{qi}^{act} \in \mathbb{R}^{H\times W\times (C/N)}$ are integrated by a convolutional layer $Conv_{K\times K}$ with the same kernel size of $K \times K$. We then concatenate each group of the weighted features along the channel dimension to obtain the convolutional modulator

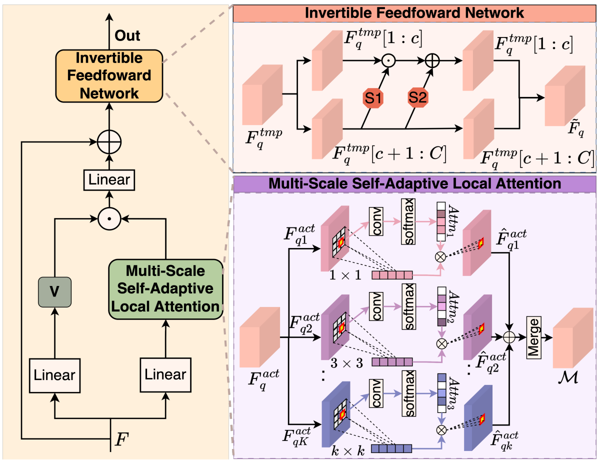

Figure 3: Illustration of the Multi-Scale Local-Aware Modulation Transformer in the target enhancement module(TEM).

$\mathcal{M} \in \mathbb{R}^{H \times W \times C}$. We initialize the kernel size with 1×1 and gradually increase it by 2 per head. The process is described by the following equation:

$$\begin{aligned} \widehat{F}_{qi}^{act} &= Conv_{K \times K}(\Gamma_1(softmax(Attn(F_{qi}^{act}))) \odot \Gamma_2(F_{qi}^{act})), \\ \mathcal{M} &= Concat(\widehat{F}_{q1}^{act}, \widehat{F}_{q2}^{act}, ..., \widehat{F}_{qN}^{act}), \end{aligned} \tag{1}$$

where $\Gamma_1(\cdot)$ denotes the transformation of $Attn(F_{qi}^{act})$ to $\mathbb{R}^{H \times W \times 1 \times (K \times K)}$, and $\Gamma_2(\cdot)$ denotes the transformation of F_{qi}^{act} to $\mathbb{R}^{H \times W \times (C/N) \times (K \times K)}$. $\odot$ represents the element-wise product.

With the convolutional modulator, we obtain features $F_q^{tmp} \in \mathbb{R}^{H \times W \times C}$ fused by spatial dimension. Finally, for preserving more detailed features, we employ the invertible feedforward network for inter-channel information fusion. This process is described as:

$$\begin{aligned} F_q^{tmp} &= Conv_{1\times1}(\mathcal{M} \odot W_v F_q^{act}) + F_q^{act}, \\ F_q^{tmp}[1:c] &= exp(S_1(F_q^{tmp}[c+1:C])) \odot F_q^{tmp}[1:c] + S_2(F_q^{tmp}[c+1:C]), \\ \tilde{F}_q &= Concat(F_q^{tmp}[1:c], F_q^{tmp}[c+1:C]), \end{aligned} \tag{2}$$

where W_v denote parameters of the linear mapping. $Conv_{1\times1}$ is a convolutional operation with a kernel size of 1×1. $\tilde{F}_q \in \mathbb{R}^{H \times W \times C}$ indicates the output feature. $F_q^{tmp}[1:c] \in \mathbb{R}^{H \times W \times c}$ is the 1st to the cth channels of the input feature F_q^{tmp}. S_i is the residual function in th INN layer.

3.3 Dual Constraint Aggregation Module

To enhance the discriminant of category cues, existing approach [Cao *et al.*, 2022] mines more category information by utilizing vanilla attention to explore the correlation between the support prototype and query features. However, in scenarios with significant intra-class variations of the target category, this single-layer constraint proves to be insufficient, resulting in the issue of *attention bias* (shown in Fig.4(c)). To address this limitation, we propose a DCAM as illustrated in Fig.4(b) , which consists of two key components: intra-class difference representation and a dual semantic-aware attention mechanism.

Intra-Class Difference Representation. To alleviate intra-class variations, we propose utilizing a set of learnable vectors, namely intra-class difference representation, to model the variance between the support set and the query set. Specifically, we expand the support prototype p_s to a size of $N \times C$, and assign an intra-class difference representation Δ_q to each of them separately. The generated prototype can be denoted as p_s^{ini}. To explain the validity of the proposed intra-class difference representation, we formulated its role in the subsequent attention process. In the role of the intra-class difference representation, the computational process of the attention map can be extended as follows:

$$\begin{aligned} W &= (Q + \Delta_q)K^T \\ &= (QK^T) + (\Delta_q K^T) \\ &= W_{supp} + W_{intra}, \end{aligned} \tag{3}$$

where W_{supp} is the attention weight obtained from query Q and key K, and W_{intra} is the weight obtained from the intra-class difference representation Δ_q and key K. W_{intra} can be considered as a factor to adjust W_{supp} according to the intra-class variation.

Dual Semantic-Aware Attention Mechanism. As illustrated in Fig.4(d), our proposed dual semantic-aware attention mechanism consists of two layers of constraints. In the first layer, we use the support prototype as a reference to select points with high matching confidence in the query feature. These selected points are then used as guidance in the second layer to find points with high feature similarity in the entire query feature map. Throughout this process, we refer to p_s^{ini} as Q_1, $\tilde{F}_q$ as Q_2, K, and V.

To begin, we calculate the matching confidence map $S_{mat} \in \mathbb{R}^{N \times HW}$ between Q_1 and K. The matching score can be obtained by

$$S_{mat} = softmax(Q_1 K^T). \tag{4}$$

Then we use the indices of the M points with the highest matching confidence in S_{mat} to guide the selection of corresponding points $TopM(Q_2)$ in Q_2 as references for the second layer constraints. It is worth noting that we employ a soft conditioning approach, where the number of M is set to be $1/4$ of the average size of all support foreground regions in each batch.

In the second layer, we calculate the feature similarity matrix between the guide points $TopM(Q_2)$ and K, and perform the maximization operation $max()$ along the guide points dimension. The process can be described as:

$$S_{sim} = max(softmax(TopM(Q_2)K^T)), \tag{5}$$

We then use the obtained attention map $S_{sim} \in \mathbb{R}^{1 \times HW}$ to weigh the aggregation of the category prototype from the query feature, and send it to the feedforward layer (FFN) to generate the robust support category prototype $\tilde{p}_s$.

3.4 Dual Classification Module

Existing methods [Tian *et al.*, 2020; Zhang *et al.*, 2021a; Liu *et al.*, 2022b] predominantly predict based on semantic consistency, often neglecting the spatial consistency of the target object, leading to failures in locating target categories. In the DCM, our objective is to decouple the segmentation task into two subtasks: semantic alignment and spatial alignment. The semantic similarity-based mask serves to identify the target category, while the spatial distribution probability-based mask assists in precise localization, synergizing for improved performance. Firstly, we optimize $\tilde{F}_q$ and $\tilde{p}_s$ using TEM and DCAM. Then, we generate the semantic similarity-based map M_{sim} by element-wise product of refined prototype $\tilde{p}_s^r$ and $\tilde{F}_q^r$. Moreover, taking into account the intrinsic guidance provided by query features, we exploit the spatial consistency within the target object itself in the query set to obtain the spatial distribution probability-based mask. Specifically, we apply spatial attention [Woo *et al.*, 2018] to $\tilde{F}_q$ and $\tilde{F}_q^r$, obtaining spatial distribution guidences, S_q and S_q^r, reflecting foreground position information. We then concatenate F_q^m, $\tilde{F}_q$, $\tilde{F}_q^r$, S_q, S_q^r and the prior mask M^0 along the channel dimension and feed them into a decoder $FEM(\cdot)$[Tian *et al.*, 2020] to estimate the spatial distribution probability-based query foreground mask M_{pro}. The process can be described as:

$$\begin{aligned} M_{sim} &= \tilde{p}_s^r \odot \tilde{F}_q^r. \\ M_{pro} &= FEM(F_q, \tilde{F}_q, \tilde{F}_q^r, M^0, S_q, S_q^r). \end{aligned} \tag{6}$$

Finally, we combin M_{sim} and M_{pro} through simple addition to obtain the ultimate query foreground segmentation map M_{final}:

$$M_{final} = M_{sim} + M_{pro}^f, \tag{7}$$

where M_{pro}^f denotes the foreground prediction map in M_{pro}.

3.5 Total Training Loss

Our training loss is composed of three parts. Firstly, we employ two dice losses, $\mathcal{L}_{final}$ and $\mathcal{L}_{sim}$, to supervise the training of the final prediction map M_{final} and the semantic similarity-based mask M_{sim}. Second, we utilize the binary cross-entropy loss $\mathcal{L}_{pro}$ to supervise the training of the spatial distribution probability-based map M_{pro}. Finally, we utilize KL(Kullback-Leibler) divergence loss $\mathcal{L}_{KL}$ to distill the foreground distribution information of the target obeject in the query to students S_q^r and S_q using the query ground truth M_q and S_q^r as the teacher, respectively. In summary, our overall objective function is:

$$\mathcal{L}_{total} = \mathcal{L}_{final} + (1 - e/epoch)(\mathcal{L}_{sim} + \mathcal{L}_{pro}) + \lambda \mathcal{L}_{KL}, \tag{8}$$

where $epoch$ represents the total number of training rounds, e represents the current round, and λ is a adjustable loss weights, here we set λ to 10.

4 Experiments

In the experiments, we leverage two popular FSS benchmarks, i.e., PASCAL-5^i [Shaban *et al.*, 2017] and COCO-20^i [Nguyen and Todorovic, 2019], to evaluate the proposed approach. We adopt mean intersection-over-union (mIoU) as the evaluation metric for experiments.

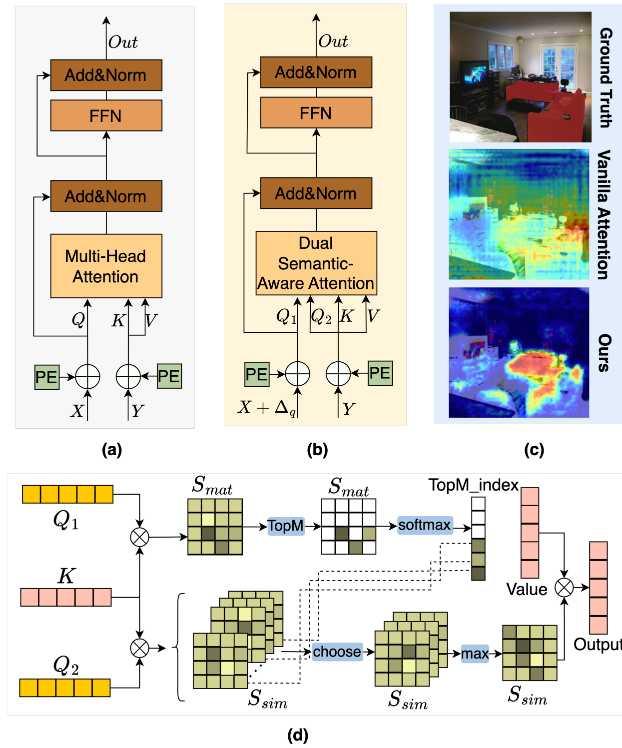

Figure 4: Detailed architectures. (a)Vanilla Transformer block, (b) DCAM block, (c) Visualization of correspondence maps, and (d) Dual semantic-aware attention mechanism in DCAM.

4.1 Implementation Details

Following [Lang *et al.*, 2022], we first train the PSPNet to obtain a backbone based on the seen training classes for each fold, i.e., 16/61 training classes (including background) for PASCAL-5^i/COCO-20^i. Subsequently, the parameters of the trained backbone are frozen, and a meta-learning strategy is employed to train the remaining structures. Optimization of these structures is conducted using the Adam optimizer with a learning rate of 10e-3, involving 50 epochs on PASCAL-5^i and 100 epochs on COCO-20^i. All images are resized directly to 473×473, and the channel dimension of the image features is set to 64. The training batch size is configured as 20 for the 1-shot setting and 15 for the 5-shot setting. No data augmentation strategies are applied during training. All experiments are executed on a single 24GB RTX3090 GPU.

4.2 Comparison With State-of-the-Art Methods

PASCAL-5^i Results. Table 1 presents a performance comparison of mIoU on the PASCAL-5^i dataset between our method and several representative models. It is evident that (1) our method outperforms the previous state-of-the-art [Min *et al.*, 2021] by 2.4% and 0.6% in the 1-shot and 5-shot settings with VGG16 as the backbone network, respectively. (2) On ResNet50, our model achieves state-of-the-art performance [Bao *et al.*, 2023] with just 1/7 of the parameters. With a minimal increase of 0.5M parameters, we surpass the baseline [Cao *et al.*, 2022] by 3.9% and 3.1% in the 1-shot

Backbone	Method	1-shot Fold-0	Fold-1	Fold-2	Fold-3	Mean	5-shot Fold-0	Fold-1	Fold-2	Fold-3	Mean	# learnable params
VGG-16	PFENet[Tian *et al.*, 2020]	56.9	**68.2**	54.4	52.4	58.0	59.0	**69.1**	54.8	52.9	59.0	10.3M
	HSNet[Min *et al.*, 2021]	**59.6**	65.7	**59.6**	**54.0**	**59.7**	**64.9**	69.0	**64.1**	**58.6**	**64.1**	**2.5M**
	NTRENet[Liu *et al.*, 2022b]	57.7	67.6	57.1	53.7	59.0	60.3	68.0	55.2	57.1	60.2	19.9M
	ours	**62.0**	**69.8**	**59.8**	**56.8**	**62.1**	**62.3**	**72.1**	**62.7**	**61.6**	**64.7**	**1.0M**
Res-50	PFENet[Tian *et al.*, 2020]	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9	10.3M
	CWT[Lu *et al.*, 2021]	56.3	62.0	59.9	47.2	56.4	61.3	68.5	68.5	56.6	63.7	-
	CYCTR[Zhang *et al.*, 2021b]	65.7	71.0	59.5	59.7	64.0	69.3	73.5	63.8	63.5	67.5	15.4M
	HSNet[Min *et al.*, 2021]	64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	**67.1**	69.5	2.5M
	IPMT[Liu *et al.*, 2022c]	**72.8**	**73.7**	59.2	**61.6**	66.8	**73.1**	74.7	61.6	63.4	68.2	-
	SSP[Fan *et al.*, 2022]	61.4	67.2	**65.4**	49.7	60.9	68.0	72.0	**74.8**	60.2	68.8	8.7M
	DCAMA[Shi *et al.*, 2022]	67.5	72.3	59.6	59.0	64.6	70.5	73.9	63.7	65.8	68.5	47.7M
	NTRENet[Liu *et al.*, 2022b]	65.4	72.3	59.4	59.8	64.2	66.2	72.8	61.7	62.2	65.7	19.9M
	RiFeNet[Bao *et al.*, 2023]	**68.4**	73.5	**67.1**	59.4	**67.1**	70.0	74.7	**69.4**	64.2	**69.6**	7.7M
	Proformer[Cao *et al.*, 2022]	65.9	72.5	55.9	58.1	63.1	71.4	**75.2**	57.5	65.7	67.4	**0.6M**
	ours	67.9	**74.3**	61.1	**64.6**	**67.0**	**72.0**	**76.4**	64.5	**69.1**	**70.5**	**1.1M**

Table 1: Comparison with state-of-the-art methods on PASCAL-5^i with class Mean-IoU metric. **Red**/**Blue** indicates the best/2nd results.

Backbone	Method	1-shot Fold-0	Fold-1	Fold-2	Fold-3	Mean	5-shot Fold-0	Fold-1	Fold-2	Fold-3	Mean	# learnable params
VGG-16	PFENet[Tian *et al.*, 2020]	35.4	38.1	36.8	34.7	36.3	38.2	42.5	41.8	38.9	40.4	**10.3M**
	SAGNN [Xie *et al.*, 2021]	35.0	40.5	37.6	36.0	37.3	37.2	45.2	40.4	40.0	40.7	-
	DPCN[Liu *et al.*, 2022a]	**38.5**	**43.7**	**38.2**	**37.7**	**39.5**	**42.7**	**51.6**	**45.7**	**44.6**	**46.2**	-
	ours	**38.6**	**45.6**	**41.4**	**41.7**	**41.8**	**45.6**	**50.7**	**48.7**	**45.8**	**47.7**	**1.0M**
Res-50	CYCTR[Zhang *et al.*, 2021b]	38.9	43.0	39.6	39.8	40.3	41.1	48.9	45.2	47.0	45.6	15.4M
	HSNet[Min *et al.*, 2021]	36.3	43.1	38.7	38.7	39.2	43.3	51.3	48.2	45.0	46.9	2.5M
	CWT[Lu *et al.*, 2021]	30.3	36.6	30.5	32.2	32.4	38.5	46.7	39.4	43.2	42.0	-
	DCAMA[Shi *et al.*, 2022]	41.9	45.1	44.4	41.7	43.3	45.9	50.5	50.7	46.0	48.3	47.7M
	NTRENet[Liu *et al.*, 2022b]	36.8	42.6	39.9	37.9	39.3	38.2	44.1	40.4	38.4	40.3	19.9M
	IPMT[Liu *et al.*, 2022c]	41.4	45.1	45.6	40.4	43.0	43.5	49.7	48.7	47.9	47.5	-
	RiFeNet[Bao *et al.*, 2023]	39.1	47.2	44.6	45.4	44.1	44.3	52.4	49.3	48.4	48.6	7.7M
	MIANet[Yang *et al.*, 2023]	**42.5**	**53.0**	**47.8**	**47.4**	**47.7**	45.8	**58.2**	51.3	51.9	51.7	-
	Protoformer[Cao *et al.*, 2022]	42.4	48.5	46.3	45.5	45.7	**48.1**	57.8	**55.0**	**52.7**	**53.4**	**0.6M**
	ours	**44.2**	**51.5**	**47.8**	**46.5**	**47.5**	**49.3**	**58.6**	**56.9**	**53.8**	**54.7**	**1.1M**

Table 2: Comparison with state-of-the-art methods on COCO-20^i with class Mean-IoU metric. **Red**/**Blue** indicates the best/2nd results.

and 5-shot settings. Notably, in the 1-shot and 5-shot settings of fold2, we outperform the baseline by 5.2% and 7.0%, and in fold3, we achieve improvements of 6.5% and 3.4%, respectively. These results further emphasize the effectiveness of our method in mitigating background interference and achieving accurate query segmentation mask.

COCO-20^i Results. COCO-20^i is a more challenging dataset with a diverse range of categories and intricate backgrounds. Table 2 illustrates the mIoU performance comparison on the COCO-20^i benchmark. It can be seen that (1) our method built on VGG16 surpass the previous state-of-the-art [Liu *et al.*, 2022a] by 2.3% and 1.5% in the 1-shot and 5-shot settings, respectively. (2) Using ResNet50 as backbone, our image-only model has competitive performance with state-of-the-art [Yang *et al.*, 2023] that utilizes both text and image information under the 1-shot setting. Furthermore, in the 5-shot setting, we surpass [Yang *et al.*, 2023] by 3.0% using only 1.1M parameters. This underscores the robust generalization capability of our model to effectively cope with in handling bias problems.

Qualitative Results. We present qualitative results comparing our method with previous works, including CYCTR [Zhang *et al.*, 2021b], IPMT [Liu *et al.*, 2022c], and Protoformer [Cao *et al.*, 2022], on the PASCAL-5^i and COCO-20^i benchmarks. Our method demonstrates several advantages over previous works, as depicted in Fig.5. (1) Our approach successfully mitigates background interference from adjacent regions, a contrast to previous works that erroneously treat the background surrounding the foreground as the region for segmentation(see 1st to 3rd column). (2) Our method rectifies the misconception of considering known classes in the background as foreground, concentrating attention on the current target category(see 4th to 7th columns). (3) Our method effectively suppresses interference from analogs in the background, facilitating precise target category localization (see 8th to 10th columns).

4.3 Ablation Experiments

We conduct following ablation studies with ResNet-50 backbone under the 1-shot setting on PASCAL-5^i dataset.

Components Analysis. Our approach comprises three main modules: target enhancement module (TEM), dual constraint aggregation module (DCAM), and dual classification module (DCM). Table 3 presents our validation on the effectiveness of each component. Compared to the baseline, using TEM alone to enhance query foreground features and using DCAM alone to enhance the discriminant of category prototypes results in a 0.9% and 2.2% improvement, respectively. The synergistic effect of TEM and DCAM lead to a 2.6% improvement. Employing DCM to achieve semantic alignment and spatial alignment provides an extra growth of 1.3%. The results reveal a 3.9% improvement of our model over the baseline, indicating that the introduced modules effectively address three issues—namely, *inherent bias*, *attention bias*, and *spatial-aware bias*. This ultimately reduces background interference, leading to precise segmentation.

Target Enhancement Module. TEM aims to mitigates the inherent bias of the backbone and enhance the query foreground regions. To evaluate the performance of our proposed method, we perform experiments with other methods in terms

Figure 5: Qualitative results of our method and alongside previous works, including CYCTR , IPMT, and Protoformer, on PASCAL-5^i and COCO-20^i benchmarks. Each row from top to bottom represents the support images with ground-truth (GT) masks (**blue**), query images with GT masks (**red**), CYCTR results (**yellow**), IPMT results (**yellow**), Protoformer results (**yellow**), and our results (**yellow**), respectively. Zoom in for details.

TEM	DCAM	DC	Fold-0	Fold-1	Fold-2	Fold-3	Mean
			65.9	72.5	55.9	58.1	63.1
✓			67.0	72.6	58.0	58.5	64.0
	✓		67.6	**74.5**	58.6	60.3	65.3
✓	✓		67.8	73.5	60.0	61.4	65.7
✓	✓	✓	**67.9**	74.3	**61.1**	**64.6**	**67.0**

Table 3: Ablation studies of main model components.

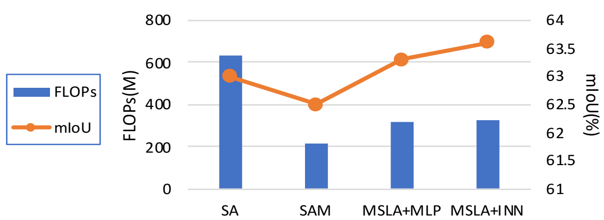

Figure 6: Comparison of target enhancement from different methods in terms of accuracy and efficience. SA: Self Alignmnet. SAM: Scale-Aware Modulation Transformer. MSAL+MLP: Our Multi-Scale Self-Adaptive Attention with MLP. MSAL+INN: Our Multi-Scale Local-Aware Modulation Transformer. FLOPs means floating point operations per second.

of computational effort and accuracy, respectively. We modified our model by (1) employing the self alignment module in [Zhang *et al.*, 2021b] as the attention mechanism for TEM (referred to as SA), (2) replacing SA with the convolutional transformer architecture [Lin *et al.*, 2023] (referred to as SAM), (3) substituting it with our Multi-Scale Self-Adaptive Local Attention (referred to as MSLA+MLP), and (4) utilizing INN[Dinh *et al.*, 2016] instead of MLP as feed-forward network (referred to as MSLA+INN). As illustrated in Figure 6, our approach maintains a high level of accuracy while reducing computational complexity. Moreover, our feedforward network retains more feature details at a slightly increased computational cost.

Dual Constraint Aggregation Module. We provide a comprehensive analysis of a crucial component in DCAM. We modified our model by (1) employing the original Vanilla Attention[Cao *et al.*, 2022] as the attention mechanism for DCAM (referred to as VA), (2) replacing VA with mask attention [Cheng *et al.*, 2022] (referred to as MA), (3) substituting it with our dual semantic-aware attention (referred to as DSAA), and (4) using intra-class difference representation (referred to as IDR). The results in Table 4 suggest that employing mask attention to mitigate background noise interference has negligible impact on performance enhancement. We attribute this to the fact that the mask derived from the similarity between support and query set suffers from the challenge of accuracy when there exists significant intra-class differences. In contrast, our dual semantic-aware attention mechanism copes with sensitivity to intra-class differences by mitigating background interference in a learnable manner(see in Fig. 4(c)). Moreover, our intra-class difference representation proves beneficial across three different attention mechanisms.

VA	MA	DSAA	IDR	mIoU(%)
✓				63.1
	✓			64.2
		✓		64.8
✓			✓	64.2
	✓		✓	65.1
		✓	✓	**65.3**

Table 4: Ablation studies on different attention mechanism settings. VA: Vanilla Attention. MA: Mask Attention. DSAA: Dual Semantic-Aware Attention. IDR: Intra-class Difference Representation.

M_{sim}	M_{pro}	mIoU(%)
		64.8
✓		66.3
	✓	62.5
✓	✓	**67.0**

Table 5: Ablation studies of main components in DCM. The baseline is equipped with TEM and DCAM. M_{sim} and M_{pro} denotes the semantic similarity-based map and the spatial distribution probability-based map respectively.

Dual Classification Module. To assess different DCM components, ablation experiments were conducted. Table 5 shows that using only the semantic similarity-based mask improves the model's performance by 1.5%, demonstrating the necessity of optimizing class prototypes and query features. However, when using only the spatial distribution probability-based segmentation map, performance decreases by 2.3%. We consider this because relying only on the foreground distribution of the query image itself causes the model to bias towards the areas of known classes, resulting in a failure of segmentation for unseen classes.

5 Conclusion

We introduce a novel transformer-based adaptive prototype matching network to counteract background interference arising from *inherent bias*, *attention bias*, and *spatial-aware bias*. Our method includes three modules: Target Enhancement Module (TEM) addresses *inherent bias* by leveraging multi-scale local context relevance to enhance foreground features. Dual Constraint Aggregation Module (DCAM) handles *attention bias* through a dual semantic-aware attention mechanism to reinforce constraints. Dual Classification Module (DCM) decouples the segmentation task into semantic alignment and spatial alignment to alleviate *spatial-aware bias*. Our experiments demonstrate that our method achieves the state-of-the-art performance with minimal parameters.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U20B2065, Grant U22B2056, Grant 62272468, Grant 62332015, and Grant 62072449; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20211539; in part by the Science and Technology Development Fund, Macau, SAR, under Grant 0075/2023/AMJ, Grant 0003/2023/RIB1, and Grant 001/2024/SKL; in part by the Zhuhai Science and Technology Innovation Bureau under Grant ZH2220004002524; and in part by the University of Macau under Grant MYRG2022-00059-FST and Grant MYRG-GRG2023-00237-FST-UMDF.

Contribution Statement

The contributions of each author to the manuscript are detailed below:

- **Sihan Chen and Yadang Chen (Co-First Authors)**: Both authors contributed equally to the Conceptualization, Methodology, Investigation, Visualization, and Writing - Original Draft.
- **Yuhui Zheng (Corresponding Author)**: Conceptualization, Funding Acquisition, Supervision, Writing - Review & Editing.
- **Zhi-Xin Yang**: Funding Acquisition, Resources.
- **Enhua Wu**: Resources, Supervision.

References

[Bao *et al.*, 2023] Xiaoyi Bao, Jie Qin, Siyang Sun, Yun Zheng, and Xingang Wang. Relevant intrinsic feature enhancement network for few-shot semantic segmentation. *arXiv preprint arXiv:2312.06474*, 2023.

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

[Chen *et al.*, 2022a] Jiacheng Chen, Bin-Bin Gao, Zongqing Lu, Jing-Hao Xue, Chengjie Wang, and Qingmin Liao. Apanet: Adaptive prototypes alignment network for few-shot semantic segmentation. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[Chen *et al.*, 2022b] Yadang Chen, Chuanyan Hao, Zhi-Xin Yang, and Enhua Wu. Fast target-aware learning for few-shot video object segmentation. *Science China Information Sciences*, 65(8):182104, 2022.

[Chen *et al.*, 2024] Yadang Chen, Ren Jiang, Yuhui Zheng, Bin Sheng, Zhi-Xin Yang, and Enhua Wu. Dual branch multi-level semantic learning for few-shot segmentation. *IEEE Transactions on Image Processing*, 33:1432–1447, 2024.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[Dinh *et al.*, 2016] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[Dong and Xing, 2018] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

[Fan *et al.*, 2022] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. *arXiv preprint arXiv:2207.11549*, 2022.

[Huang *et al.*, 2022] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 972–979. International Joint Conferences on Artificial Intelligence, 2022.

[Lang *et al.*, 2022] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022.

[Li *et al.*, 2021] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021.

[Lin *et al.*, 2023] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6015–6026, 2023.

[Liu *et al.*, 2022a] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022.

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

[Liu *et al.*, 2022c] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 35:38020–38031, 2022.

[Lu *et al.*, 2021] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8741–8750, 2021.

[Min *et al.*, 2021] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021.

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

[Ouyang *et al.*, 2023] Shuyi Ouyang, Hongyi Wang, Shiao Xie, Ziwei Niu, Ruofeng Tong, Yen-Wei Chen, and Lanfen Lin. Slvit: scale-wise language-guided vision transformer for referring image segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1294–1302, 2023.

[Peng *et al.*, 2023] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023.

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

[Shi *et al.*, 2022] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022.

[Tian *et al.*, 2020] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

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

[Wang *et al.*, 2019] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *IJCAI International joint conference on artificial intelligence*, 2019.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xie *et al.*, 2021] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5475–5484, 2021.

[Yang *et al.*, 2020] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 763–778. Springer, 2020.

[Yang *et al.*, 2021] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8721–8730, 2021.

[Yang *et al.*, 2023] Yong Yang, Qiong Chen, Yuan Feng, and Tianlin Huang. Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7131–7140, 2023.

[Ye *et al.*, 2021] Qiaolin Ye, Peng Huang, Zhao Zhang, Yuhui Zheng, Liyong Fu, and Wankou Yang. Multiview learning with robust double-sided twin svm. *IEEE transactions on Cybernetics*, 52(12):12745–12758, 2021.

[Zhang *et al.*, 2019] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019.

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

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

[Zhang *et al.*, 2021b] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021.