Dual-Guided Frequency Prototype Network for Few-Shot Semantic Segmentation

Chunlin Wen, *Member, IEEE*, Hui Huang, *Member, IEEE*, Yan Ma, *Member, IEEE*, Feiniu Yuan, *Senior Member, IEEE*, and Hongqing Zhu, *Member, IEEE*

***Abstract*—Few-shot semantic segmentation is a challenging task that aims to segment novel classes in the query images given only a few annotated support samples. Most existing prototype-based approaches extract global or local prototypes by global average pooling (GAP) or clustering to represent all object information. Subsequently, the prototype information is employed as guidance for query image segmentation. However, these frameworks fail to fully mine the object details and ignore information from query images. Consequently, we propose a Dual-Guided Frequency Prototype Network (DGFPNet) to solve these issues. Specifically, to mine the global and local object information, a Frequency Prototype Generation Module (FPGM) is first proposed to extract more comprehensive frequency prototypes by multi-frequency pooling (MFP) in the DCT domain. Then, with the guidance of support and query information, a Dual-Guided Selection Module (DGSM) is presented to produce the query attention mask and select more effective prototypes. Based on the query attention mask and support information, the generalized object information is integrated into the feature with the proposed Feature Generalization Module (FGM). Finally, we propose a Multi-Dimension Feature Enrichment Decoder Module (MDFEDM) to capture multi-dimension object information and tackle hard pixels for refining the final segmentation results. Extensive experiments on PASCAL-5^i and COCO-20^i show that our model achieves new state-of-the-art performances.**

***Index Terms*—Few-shot segmentation, few-shot learning, prototype learning, frequency domain learning, dual-guidance.**

Manuscript received 5 October 2023; revised 23 February 2024; accepted 24 March 2024. Date of publication 29 March 2024; date of current version 21 August 2024. This work was supported by the National Nature Science Foundation of China under Grant 61872143. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Song Wang. *(Corresponding authors: Hui Huang; Feiniu Yuan.)*

Chunlin Wen and Yan Ma are with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China (e-mail: w614057218@gmail.com; ma-yan@shnu.edu.cn).

Hui Huang is with the College of Information, Mechanical and Electrical Engineering, Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai 201418, China (e-mail: huanghui@shnu.edu.cn).

Feiniu Yuan is with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China, also with Research Base of Online Education for Shanghai Middle and Primary Schools, Shanghai, China, and also with Shanghai Engineering Research Center of Intelligent Education and Big Data, Shanghai 201418, China (e-mail: yfn@ustc.edu).

Hongqing Zhu is with the School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China (e-mail: hqzhu@ecust.edu.cn).

Our code will be released at https://github.com/ChunLinWen/DGFPNet.

Digital Object Identifier 10.1109/TMM.2024.3383276

I. Introduction

DEEP learning [1], [2] has been widely applied to various computer vision tasks, especially in semantic segmentation, which classifies each pixel into a particular class. Benefiting from the large pixel-level annotated datasets, semantic segmentation [3], [4], [5], [6], [36] has made tremendous progress in recent years. However, it is labor-intensive and time-consuming to manually annotate a large amount of pixel-wise labeled images. Furthermore, given a few annotated images, semantic segmentation hardly generalizes to novel classes. In contrast, humans can easily recognize a new concept from the image when only seeing a few examples. Inspired by this, Few-shot Semantic Segmentation (FSS) is investigated to solve this problem.

Few-shot semantic segmentation [7] aims to produce pixel-level predictions of novel category samples in the query image using only a few labeled samples. Following [8], [9], most previous methods adopt the prototype-based network to extract the global prototype from the support image by global average pooling (GAP), as shown in Fig. 1(a). However, the global prototype unavoidably tends to lose some intrinsic support object parts, causing incomplete segmentation for query targets, like the lost plane empennage in the prediction of Fig. 1(a). Then, some studies [10], [11], [12] further generate local prototypes by clustering the support features, as shown in Fig. 1(b). These studies mine local detail object information while ignoring the global understanding of the objects. In this case, these studies fail to discriminate the target objects and induce background noise in the prediction, such as the misclassified toolbox in the prediction of Fig. 1(b). Hence, global and local prototypes are essential to guide the target segmentation in the query images. Besides, there are many different attributions between query and support images, like quantities, views, illumination intensities, etc. Taking only the support information as the guidance usually leads to inaccurate matching between support and query features. Therefore, some methods are designed to build relation maps between support and query samples, such as graph attention [13], [37], cross-reference [14], 4D convolutions [15], Transformers [16], etc. Nevertheless, the entire features they used usually contain irrelevant background context and introduce too much noise, leading to performance decline.

To address the abovementioned problems, we propose Dual-Guided Frequency Prototype Network (DGFPNet), as shown in Fig. 1(c). The model prototypes are generated from various frequency components to capture

1520-9210 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

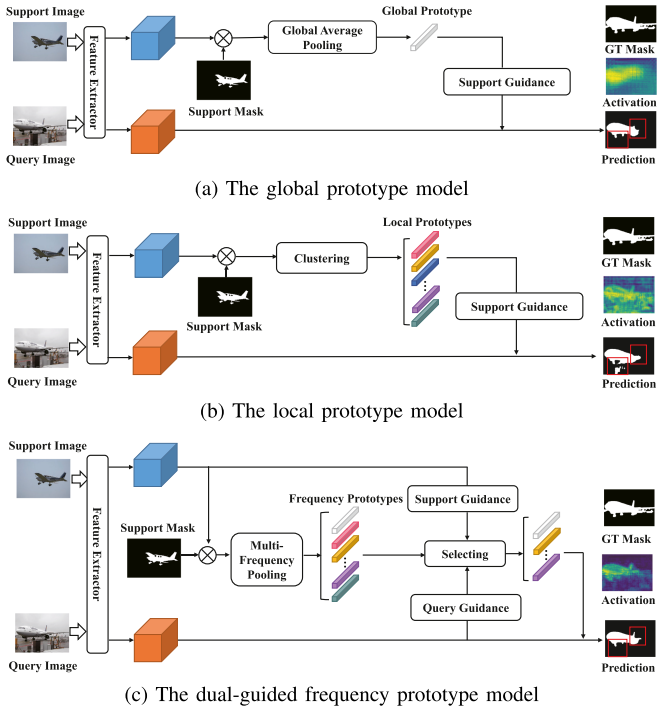


Fig. 1. Comparison of our DGFPNet and previous works. (a) The global prototype model produces single prototype by GAP from the global perspective, where the activation and prediction are generated by BAM [18]. (b) The local prototype model extracts multiple prototypes by clustering from the local perspective, where the activation and prediction are generated by ASGNet [11]. In contrast, (c) our DGFPNet proposes the multi-frequency pooling to generate multiple frequency prototypes in the DCT domain from both global and local perspective. Moreover, with the guidance of support and query information, DGFPNet selects more effective prototypes to bridge the gap between the support and query sets.

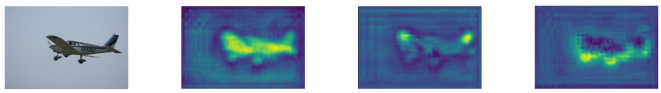


Fig. 2. Visualization of activation maps generated by different frequency prototypes. (a) denotes the original image. (b), (c), and (d) denote the activation map generated by frequency prototype 0, 9, and 27, respectively. The activation map (b) focuses on the global features like the fuselage of the airplane, while the activation map (c) and (d) pay attention to the local features like the empennage and wheel of the airplane. Please note that the frequency prototype 0 is equal to the prototype generated by GAP. Best viewed in color.

global and local object information in a unified way. Subsequently, we select effective prototypes by query and support information to diminish the domain gap between source and target. Specifically, we first propose a Multi-Frequency Pooling (MFP) to generalize prototype learning in the frequency domain and design a Frequency Prototype Generation Module (FPGM) based on MFP. FPGM extracts informative frequency prototypes in the frequency domain from the global and local perspectives to capture the essential features lost by GAP or clustering. As shown in Fig. 2, frequency prototypes contain global and local features for representing the object comprehensively and completely. So, compared to GAP and clustering, our model can mine information about the plane empennage part and filter out background noise well in the prediction of Fig. 1(c). Then, guided with support and query object information, effective prototypes are selected from all frequency prototypes to bridge the gap between the support set and query set, named Dual-Guided Selection Module (DGSM). Based on these effective prototypes, DGSM produces the training-free query attention mask, as shown in the activation of Fig. 1(c). Our attention mask is more accurate and less noisy than previous single-guided prior attention masks in Fig. 1(a) and Fig. 1(b). Technically, we select the support ground-truth mask as the support-guided mask, and integrate the mid-level attention mask, high-level attention mask, and base learner result [18] to generate the query-guided mask. Therefore, different from the previous dual-guided methods [13], [14], [15], [16], our method captures more target object information by guided masks that filter out most irrelevant information. Next, benefiting from the generalization ability of the query attention mask, the generalized feature is obtained by fusing this mask and the support information to focus more on the specific class in the proposed Feature Generalization Module (FGM). Finally, the previous decoder [11], [18], [20] is difficult to mine horizontal or vertical banding object features and classify some hard pixels. Therefore, we propose Multi-Dimension Feature Enrichment Decoder Module (MDFEDM) that integrates the asymmetric kernels into the FEM [20] to mine the object banding details and pre-segments the query attention mask to refine the final prediction.

In summary, the main contributions of this paper are listed as follows:

- We propose the Dual-Guided Frequency Prototype Network (DGFPNet) for few-shot semantic segmentation. DGFPNet mines global and local object-guided information in the frequency domain and bridges the gap between the support and query sets.
- We introduce the FPGM that extracts comprehensive frequency prototypes by MFP in the DCT domain to address the limitation of GAP and clustering. Then, DGSM is proposed to select effective prototypes and generate the final query attention mask with the guidance of support and query information.
- We introduce FGM and MDFEDM. FGM integrates query attention mask and support object information into query features to enhance the feature generalization. MDFEDM captures the horizontal and vertical banding object features by asymmetric kernels and refines the final prediction by handling hard pixels.
- Our DGFPNet achieves new state-of-the-art results on PASCAL-5^i and COCO-20^i datasets. Extensive experiments validate the effectiveness of each component in our DGFPNet.

II. Related Work

A. Semantic Segmentation

Semantic segmentation has made significant progress based on large-scale labeled datasets, aiming to provide pixel-level predictions for an input image. Benefiting from the advantage of fully convolutional networks (FCNs) [3], various robust networks have been designed successively. They focus on the larger

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

receptive field [5], [21], multi-scale feature aggregation [4], [22], and encoder-decoder architecture [6], [23]. Specifically, Deeplab [5] employs dilated convolution to extend the receptive field and capture more content. PSPNet [4] utilizes pyramid pooling to concatenate multi-scale features. Meanwhile, some researches [24], [25] show that the attention mechanism can capture long-distance dependencies and achieve state-of-the-art performance. Our work also introduces the multi-scale feature aggregation and the encoder-decoder architecture to get more informative and valuable representations.

B. Few-shot Semantic Segmentation

Few-shot semantic segmentation aims to segment novel classes in query images with only a few annotated examples. OSLSM [7] first proposes a two-branch network to solve the task. It generates classifier weights in the conditioning branch and uses the weights to segment the object in the segmentation branch. Then, PL [8] uses the global average pooling (GAP) to learn the prototypes from the support set for segmentation. Following PL, some prototype-based methods are proposed, such as SG-One [9], CANet [26], CRNet [14], PFENet [20], and SMCP [39]. They typically concatenate the expanded prototype with the query feature and densely match for segmentation. However, the global prototype usually contains limited object information. Therefore, recent studies focus on local prototypes to enhance the few-shot segmentation model. RPMM [10] leverages multiple prototypes generated by the expectation-maximization (EM) algorithm to correlate diverse image regions. ASGNet [11] introduces a superpixel-guided clustering method to extract multiple local prototypes from the support image. Moreover, some studies generate prototypes from other perspectives to further mine the object information. DPNet [27] builds other pseudo-prototype based on foreground features in the query image. DCP [28] proposes the divide-and-conquer proxies to derive different prototypes from a broader perspective. DPCN [19] utilizes dynamic kernels from the support foreground to fully capture the intrinsic details. Differently, we analyze the multiple prototypes in the frequency domain and further select effective prototypes based on both support and query images. Motivated by BAM [18], which applies an additional branch (base learner) to explicitly identify the targets of base classes, we also integrate the base learner into our model.

C. Frequency Domain Learning

The frequency analysis has been considered a powerful method to capture rich representation and process complicated tasks in recent computer vision research. In [29], the spectral bias is analyzed from the frequency perspective to identify and remove the trivial frequency components without classification accuracy loss. FcaNet [17] generalizes the compression of the channel attention mechanism in the frequency domain. ChfL [30] transforms the density map into the frequency domain to process the crowd-counting task. Meanwhile, some studies focus on the low-frequency and high-frequency components in the images. Geirhos et al. [31] demonstrated that CNN is more biased towards learning low-frequency local features than humans. Moreover, Wang et al. [32] showed that high-frequency components (HFC) can explain the generalization of convolutional neural networks. Afterward, SSAH [33] introduces the low-frequency constraint to limit perturbations within high-frequency components and ensure perceptual similarity between adversarial examples and origin. DecoupleSegNets [34] explicitly models the object body and edge for semantic segmentation, corresponding to the low and high frequency. Similarly, our work mines comprehensive image information and extracts representative prototypes based on various frequency components.

III. Task Description

Following the episode-based meta-learning paradigm [7], [8], [20], we apply the standard few-shot semantic segmentation setting. The dataset is divided into the training set D_{train} and the testing set D_{test}, where the class set in D_{train} is C_{base}, and the class set in D_{test} is C_{novel}. Note that the training class set and testing class set are disjoint, $C_{base} \cap C_{novel} = \varnothing$. Based on D_{train} and D_{test}, a series of episodes are randomly sampled from them. For a specific class C, each episode is formed with a query set $Q = \{(I_q, M_q)\}$ and a support set $S = \{(I_s^i, M_s^i)\}_{i=1}^k$. It is called the k-shot segmentation learning task. Specifically, $I_s^i \in \mathbb{R}^{H\times W\times 3}$ represents the i-th image, and $M_s^i \in \{0,1\}^{H\times W}$ indicates its binary mask of class C in the support image. Similarly, I_q is the query image, and M_q is the ground-truth mask for class C in the query image, which is only used for training. Our goal is to optimize the model on D_{train} during training and subsequently generalize it to D_{test} during testing without further optimization.

IV. Method

A. Overview

To mine more comprehensive support information and diminish the domain gap, we propose the Dual-Guided Frequency Prototype Network (DGFPNet), as shown in Fig. 3. The main idea of the model is to generate multiple prototypes from various frequency components in a unified way and segment the target under the guidance of support and query information. Firstly, the support image I_s and query image I_q are fed into the pre-trained backbone to extract mid-level features ($F_{s_2}^M$, $F_{s_3}^M$, $F_{q_2}^M$, and $F_{q_3}^M$) and high-level features (F_s^H and F_q^H), where F_{s_i/q_i}^M is the feature extracted from Block-i, and $F_{s/q}^H$ is the feature extracted from Block-4. Then, Frequency Prototype Generation Module (FPGM) uses $F_{s_3}^M$ to obtain all frequency prototypes P_{All} by multi-frequency pooling (MFP). These prototypes contain global and local information. Given the base learner result M^B, P_{All}, $F_{s_3}^M$, $F_{q_3}^M$, F_s^H, and F_q^H, we employ Dual-Guided Selection Module (DGSM) to select more effective prototypes and produce the query attention mask M_q^A, main prototype P_{main}, and complemental prototype P_{comple}, bridging the domain gap. Next, the generalized feature F^G is generated by Feature Generalization Module (FGM) with the input of M_q^A, $F_{s_2}^M$, $F_{s_3}^M$, $F_{q_2}^M$, and $F_{q_3}^M$. Finally, the results of the above modules are

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

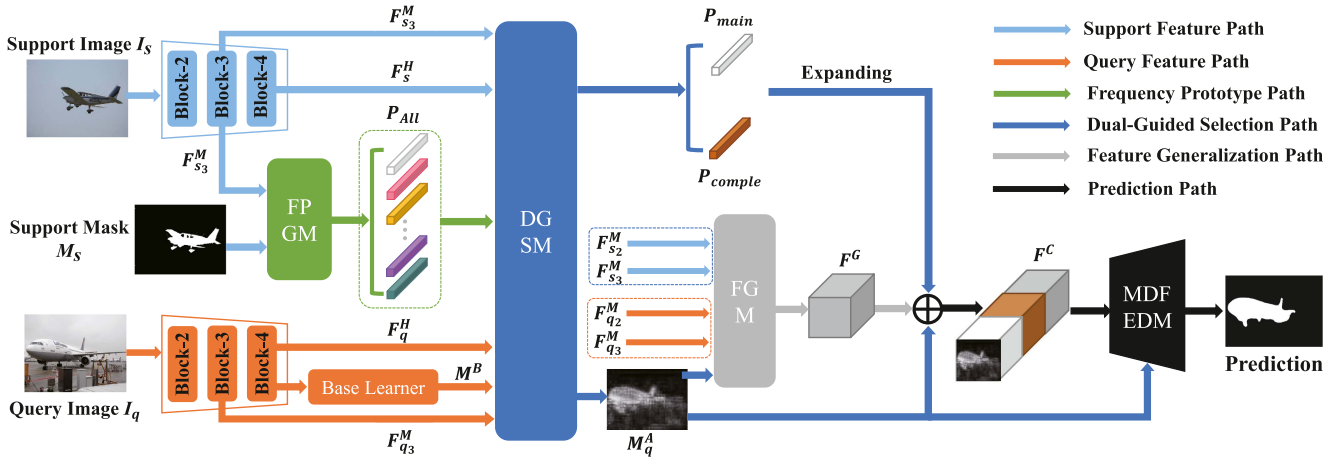


Fig. 3. Overall architecture of our proposed Dual-Guided Frequency Prototype Network (DGFPNet). Firstly, Frequency Prototype Generation Module (FPGM) takes the mid-level support feature $F^M_{s_3}$ and corresponding ground-truth mask M_s as the input to extract all frequency prototypes P_{All} by multi-frequency pooling (MFP). Then, with the guidance of support and query information, Dual-Guided Selection Module (DGSM) selects more effective prototypes for producing the query attention mask M^A_q, main prototype P_{main}, and complemental prototype P_{comple}. Next, Feature Generalization Module (FGM) fuses the query attention mask information and mid-level features to generate the generalized feature F^G. Finally, the concatenated features F^C are fed into the Multi-Dimension Feature Enrichment Decoder Module (MDFEDM) to capture multi-dimension features and segment objects.

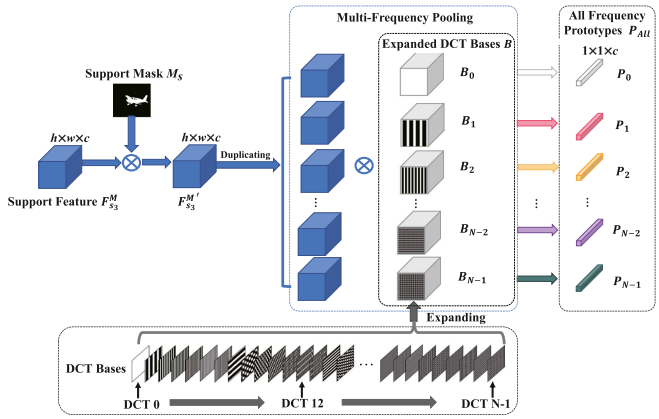


Fig. 4. Illustration of our Frequency Prototype Generation Module.

concatenated along the channel dimension, named concatenated features F^C, and taken as the input of the Multi-Dimension Feature Enrichment Decoder Module (MDFEDM) to segment the query object.

B. Frequency Prototype Generation Module

The previous few-shot segmentation methods usually extract the global prototype by GAP [8], [9], [20] or local prototypes by clustering [10], [11], [12], resulting in losing some intrinsic object information. Therefore, to address the limitation of GAP and clustering, we propose multi-frequency pooling (MFP) that generates global and local frequency prototypes in the DCT domain by sufficiently mining the object information, named Frequency Prototype Generation Module (FPGM).

Since the mid-level feature constitutes object parts shared by unseen classes [26], we apply the mid-level feature to extract our frequency prototypes. As shown in Fig. 4, the mid-level support feature $F^M_{s_3} \in \mathbb{R}^{h\times w\times c}$ is first multiplied by the support mask M_s to generate the foreground support feature $F^{M'}_{s_3} \in \mathbb{R}^{h\times w\times c}$. Then, the expanded DCT bases $B = \{B_t\}_{t=0}^{N-1}$, which is the variant of the basic DCT formula, is computed as:

$$\begin{aligned} B_t &= Expand(DCT(u,v)), t \in \{0,1,\ldots,N-1\}, \\ u &= \lfloor t/K \rfloor \times \bar{h}, v = (t\ (mod\ K)) \times \bar{w}, \\ u &\in \{0,1,\ldots,K-1\} \times \bar{h}, v \in \{0,1,\ldots,K-1\} \times \bar{w}, \\ \bar{h} &= \lfloor h/K \rfloor, \bar{w} = \lfloor w/K \rfloor, N = K \times K, \end{aligned} \tag{1}$$

where $B_t \in \mathbb{R}^{h\times w\times c}$ is the t-th DCT base in $B \in \mathbb{R}^{N\times h\times w\times c}$, h and w represent the height and width of B_t. $\lfloor\cdot\rfloor$ is the floor bracket that rounds the number to the lower integer, and mod is the modulo operation. (u, v) is the corresponding 2D frequency component. The whole 2D DCT frequency space is divided into $K \times K$ parts (as illustrated in [17], $K = 7$). $Expand(\cdot)$ denotes duplicating and expanding $DCT(u, v) \in \mathbb{R}^{h\times w\times 1}$ to the same channel as $F^{M'}_{s_3}$. $DCT(u, v)$ is the generalized DCT component in (u, v):

$$DCT(u,v) = \alpha(u)\alpha(v)\cos\left(\frac{\pi u}{h}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi v}{w}\left(j+\frac{1}{2}\right)\right), \tag{2}$$

$$\alpha(x) = \begin{cases} 1, & x = 0, \\ \sqrt{2}, & x \neq 0, \end{cases}$$
$$i \in \{0,1,\ldots,h-1\}, j \in \{0,1,\ldots,w-1\}, \tag{3}$$

where $\alpha(u)$ and $\alpha(v)$ are the coefficients of $DCT(u, v)$ calculated according to (3), and i and j represent the spatial position of B_t.

Finally, $F^{M'}_{s_3}$ is duplicated to N copies and computed element-wise multiplication with the corresponding expanded DCT bases B to produce all support frequency prototypes $P_{All} = \{P_t\}_{t=0}^{N-1}$. This procedure is called multi-frequency pooling (MFP) in the

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

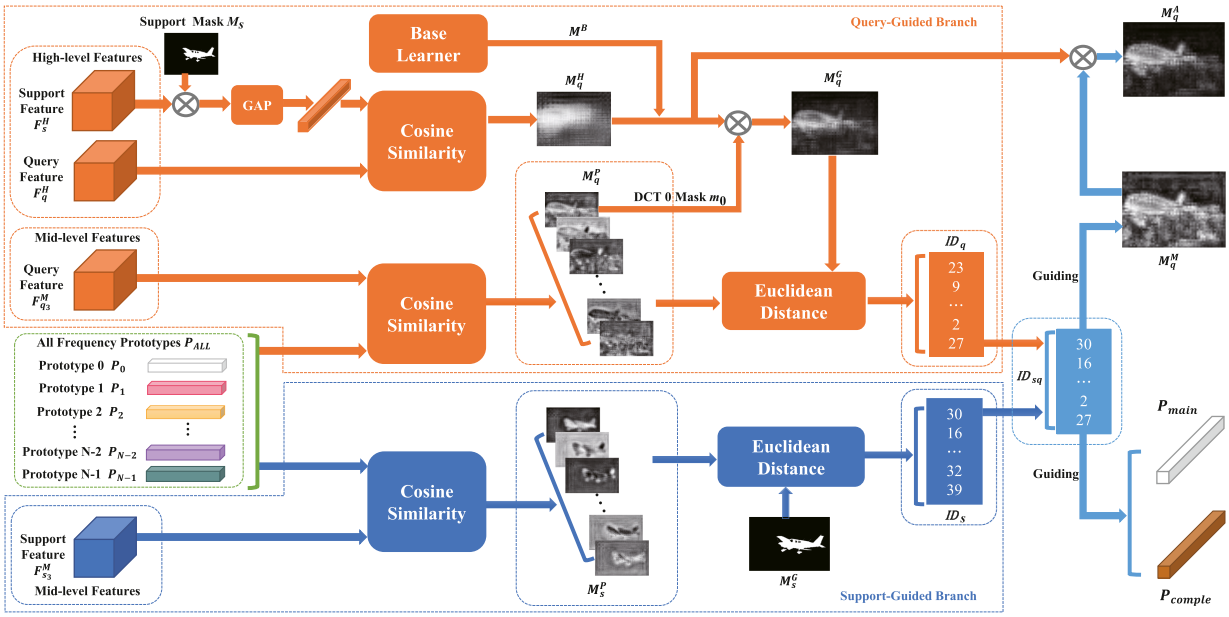


Fig. 5. Illustration of our Dual-Guided Selection Module. The above is the query-guided branch, and the below is the support-guided branch.

DCT domain:

$$P_t = \frac{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} F_{s_3}^{M'}(i,j)\cdot B_t(i,j)}{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[M_s(i,j)=1]}.$$
$$t \in \{0,1,\ldots,N-1\}. \tag{4}$$

When we set the t to 0, the lowest frequency support prototype P_0 is obtained:

$$\begin{aligned} P_0 &= \frac{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} F_{s_3}^{M'}(i,j)\cdot B_0(i,j)}{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[M_s(i,j)=1]} \\ &= \frac{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} F_{s_3}^{M}(i,j)\cdot [M_s(i,j)=1]}{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[M_s(i,j)=1]} \\ &= \text{masked } GAP(F_{s_3}^M, M_s). \end{aligned} \tag{5}$$

(5) proves that P_0 is equal to the prototype generated by the masked GAP. So, the prototype generated from the masked GAP only preserves the lowest frequency component object information while discarding the others. Moreover, since clustering compresses the partial object information into the local prototypes, it lacks the global recognition of the object, i.e., the lowest frequency component object information. In contrast, MFP utilizes the lowest and higher frequency information to capture global and local object features lost by GAP or clustering, generating comprehensive and informative frequency prototypes.

C. Dual-Guided Selection Module

Among all generated frequency prototypes, there are inferior and inefficient prototypes for query target segmentation, leading to model performance degradation. Moreover, it is different in many attributions between query and support images. Therefore, we propose the Dual-Guided Selection Module (DGSM) that selects superior and efficient prototypes for query target segmentation under support and query information guidance, bridging the gap between the support set and query set. Based on the selected prototypes, we obtain the final prototypes and query attention mask, where the prototypes are complementary, and the mask is more accurate and less noisy than the previous prior attention masks [11], [18], [20]. The process of DGSM is illustrated in Fig. 5.

DGSM mainly comprises two branches: query-guided branch and support-guided branch. The key idea of the guided branch is to obtain efficient prototype indexes by comparing the guided mask and corresponding prototype masks. The more efficient prototypes can generate prototype masks that are more similar to the guided mask. So, we apply the Euclidean distance between the guided mask and the prototype mask to judge the prototype quality. Moreover, to this end, we also need to obtain high-quality query and support guided masks. Previous works [15], [19], [40], [48], [49] usually fuse the prior attention information from both high-level and mid-level features to perform dense pixel predictions and refine the final results. Motivated by this, our query-guided mask M_q^G is generated by the base learner result M^B, the high-level attention mask M_q^H, and the DCT 0 mask m_0, where M^B is defined in BAM [18] that introduces an additional branch to explicitly predict the regions of base classes in the query images. M^B filters out base class objects in the background, M_q^H suppresses most background noise and locates targets, and m_0 represents global object information. As for the support-guided mask M_s^G, it is directly using the support ground-truth mask M_s. Different from the previous methods [13], [14], [15], [16], our guided masks filter out most background noise, focusing on mining more target object information from support and query features.

In the query-guided branch, the high-level query feature $F_q^H \in \mathbb{R}^{h\times w\times c}$ and the support prototype extracted by GAP are first fed into the cosine similarity to generate the high-level query attention mask $M_q^H \in \mathbb{R}^{h\times w\times 1}$. Then, given all frequency

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

prototypes P_{All} and the mid-level query feature $F_{q_3}^M \in \mathbb{R}^{h\times w\times c}$, we also apply the cosine similarity to produce the query frequency prototype masks $M_q^P = \{m_t\}_{t=0}^{N-1} \in \mathbb{R}^{N\times h\times w\times 1}$:

$$m_t = \mathcal{C}(F_{q_3}^M, P_t), t \in \{0, 1, \ldots, N-1\}, \tag{6}$$

where the $\mathcal{C}(\cdot)$ denotes the cosine similarity. After that, to obtain the high-quality guided mask, the query-guided mask $M_q^G \in \mathbb{R}^{h\times w\times 1}$ is produced by multiplying $M^B \in \mathbb{R}^{h\times w\times 1}$, M_q^H and m_0:

$$M_q^G = M^B \otimes M_q^H \otimes m_0, \tag{7}$$

where $\otimes$ denotes element-wise multiplication. The more effective prototype generates the more accurate and less noisy mask m_t. In other words, the Euclidean distance between m_t and M_q^G is smaller. So, based on all Euclidean distances $E_q = \{e_t\}_{t=0}^{N-1}$ between M_q^P and M_q^G, the query-guided indexes ID_q are the first Λ indexes with minimum distances, where Λ means the number of selected indexes:

$$e_t = \mathcal{T}(m_t, M_q^G), \tag{8}$$

$$ID_q = \mathcal{F}_{\min}(E_q, \Lambda), \tag{9}$$

where $\mathcal{T}(\cdot)$ represents the Euclidean distance, and $\mathcal{F}_{\min}(\cdot, \Lambda)$ represents sorting the Euclidean distances in ascending order and then taking out the indexes of the first Λ distance values.

In the support-guided branch, the support frequency prototype masks $M_s^P \in \mathbb{R}^{N\times h\times w\times 1}$ are first generated according to (6) with the input of $F_{s_3}^M$ and P_{All}. Then, the support ground-truth mask is resized to the same spatial size as M_s^P and taken as the support-guided mask $M_s^G \in \mathbb{R}^{h\times w\times 1}$. Finally, given the M_s^P and M_s^G, we obtain the support-guided indexes ID_s according to (8) and (9).

After obtaining ID_s and ID_q, the guided indexes ID_{sq} is generated by the union of ID_s and ID_q to fuse the guidance information from both support and query images and bridge the gap between them. Based on ID_{sq}, we select the corresponding query frequency prototype masks. Then, the pixel-level maximum value of all selected masks is taken as the mid-level query attention mask $M_q^M \in \mathbb{R}^{h\times w\times 1}$:

$$ID_{sq} = ID_s \cup ID_q, \tag{10}$$

$$M_q^M = \max(\mathcal{F}_{\text{select}}(M_q^P, ID_{sq})), \tag{11}$$

where $\mathcal{F}_{\text{select}}(\cdot, ID_{sq})$ denotes the selection based on ID_{sq}, and $\max(\cdot)$ denotes computing the pixel-level maximum value. Finally, the query attention mask $M_q^A \in \mathbb{R}^{h\times w\times 1}$ is obtained by multiplying the M^B, M_q^H, and M_q^M to absorb the prior information from high-level feature, mid-level feature, and base learner:

$$M_q^A = M^B \otimes M_q^H \otimes M_q^M. \tag{12}$$

With the guidance of ID_{sq}, we further obtain complementary frequency prototypes. For the main prototype $P_{\text{main}} \in \mathbb{R}^{1\times 1\times c}$, we directly use the prototype generated by DCT 0, which contains the global object information. As for the complemental prototype $P_{comple} \in \mathbb{R}^{1\times 1\times c}$, it is generated from the average of the remaining selected prototypes, which contains the local object information:

$$P_{comple} = avg(\mathcal{F}_{\text{select}}(P_{All}, ID_{sq}) - P_{\text{main}}), \tag{13}$$

where $avg(\cdot)$ represents the average value. Note that the P_{main} and P_{comple} are generated by FPGM with the input of support features derived from Block-2 and Block-3, according to [11], [18], [20], [28], [43].

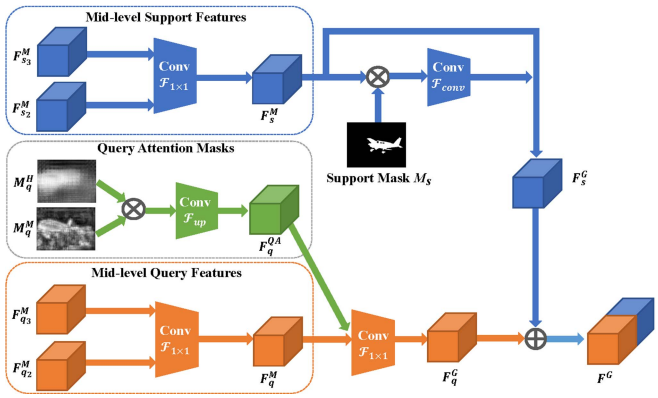


Fig. 6. Illustration of our Feature Generalization Module.

D. Feature Generalization Module

Query features are usually extracted by the frozen pre-trained backbone in the previous works [10], [11], [20]. However, some semantic clues contained in query features are unrelated to the objects shown in support images, bringing unexpected obstacles to segment novel classes in FSS. Therefore, we propose the Feature Generalization Module (FGM) to fuse more object semantic information into the features and further enhance the model generalization ability, as shown in Fig. 6.

Specifically, we first concatenate the features derived from Block-2 and Block-3, as shown in Fig. 3. Then, the support feature $F_s^M \in \mathbb{R}^{h\times w\times c}$ and query feature $F_q^M \in \mathbb{R}^{h\times w\times c}$ are generated by 1×1 convolution, respectively, where the convolution reduces the channel number of the concatenated features:

$$F_s^M = \mathcal{F}_{1\times 1}(F_{s_2}^M \oplus F_{s_3}^M), F_q^M = \mathcal{F}_{1\times 1}(F_{q_2}^M \oplus F_{q_3}^M), \tag{14}$$

where $\oplus$ denotes the concatenation along the channel dimension.

Given the F_s^M, the support generalization feature $F_s^G \in \mathbb{R}^{h\times w\times c}$ is generated by the residual structure inspired by [35]:

$$F_s^G = \mathcal{F}_{conv}(F_s^M \otimes M_s) + F_s^M, \tag{15}$$

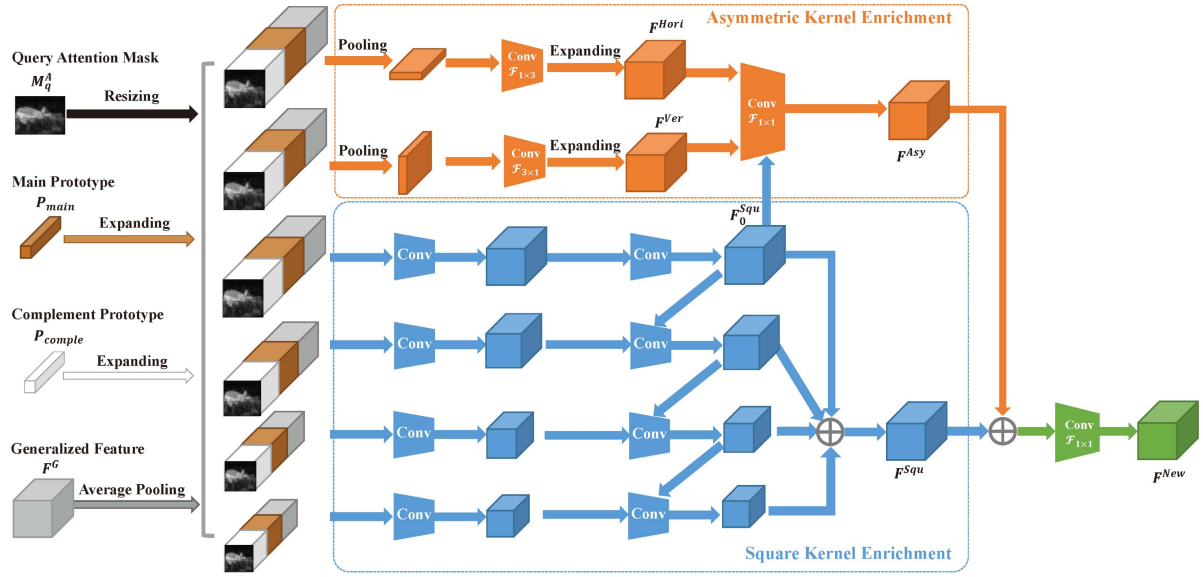
where $F_{conv}(\cdot)$ denotes the convolution network.

As for the query generalization feature F_q^G, we integrate the query attention feature F_q^{QA} into F_q^M to produce it, where F_q^{QA} is generated by fusing the semantic clues of M_q^H and M_q^M:

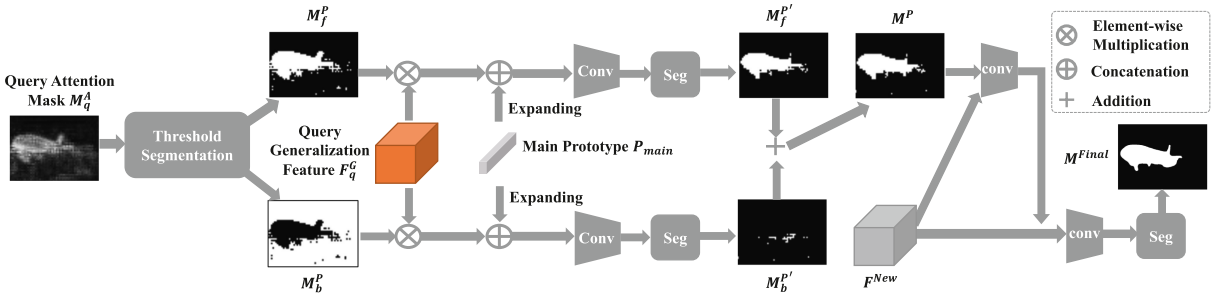
$$F_q^{QA} = \mathcal{F}_{up}(M_q^H \otimes M_q^M), \tag{16}$$

$$F_q^G = \mathcal{F}_{1\times 1}(F_q^{QA} \oplus F_q^M), \tag{17}$$

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.



(a) The feature enrichment part.



(b) The segmentation part.

Fig. 7. Illustration of our Multi-Dimension Feature Enrichment Decoder Module. The above is the feature enrichment part, and the below is the segmentation part.

where $\mathcal{F}_{up}(\cdot)$ consists of two convolutions: the 1×1 convolution for increasing the channel number and the 3×3 convolution for refining the mask features.

Finally, the generalized feature F^G is generated by concatenating F_s^G and F_q^G and subsequently fed into the decoder module for feature matching and target segmentation.

E. Multi-Dimension Feature Enrichment Decoder Module

FEM (Feature Enrichment Module) [20] is first proposed in the PFENet to overcome spatial inconsistency by the square kernels. Then, some researches [40], [41], [48] adopt FEM for its efficiency. However, we find it difficult for FEM to capture the horizontal or vertical banding object features. Therefore, asymmetric kernels are integrated into the FEM to mine the subtle object details, as shown in Fig. 7(a). Although the feature enrichment part captures multi-scale and banding object features, there are still some hard pixels (the false negative pixels and the false positive pixels) in the segmentation result. To tackle the hard pixels, the segmentation part is introduced to refine the segmentation result, as shown in Fig. 7(b). The segmentation part applies threshold segmentation to obtain the foreground and background pseudo-segmentation results. Based on the pseudo-segmentation results, the pseudo foreground segmentation branch can filter out the background pixels (false positive pixels) in the object, like the background pixels at the bottom of the plane nose, and the pseudo background segmentation branch can mine the foreground pixels (false negative pixels) in the background, like the pixels at the bottom of the plane fuselage.

As for the feature enrichment part, we first follow FEM [20] to construct the square kernel enrichment. Then, for the asymmetric kernel enrichment, the concatenated features $F^C \in \mathbb{R}^{h \times w \times (4c+1)}$ are processed with the convolution network to generate the horizontal or vertical banding object feature:

$$F^{Hori} = \text{expand}(\mathcal{F}_{1\times3}(\text{avgpool}_h(F^C))) \in \mathbb{R}^{h\times w\times c},$$
$$F^{Ver} = \text{expand}(\mathcal{F}_{3\times1}(\text{avgpool}_w(F^C))) \in \mathbb{R}^{h\times w\times c}, \quad (18)$$

where $avgpool_h(\cdot)$ and $avgpool_w(\cdot)$ denote averaging the feature values under the high and width dimension, respectively. $\mathcal{F}_{1\times3}(\cdot)$ and $\mathcal{F}_{3\times1}(\cdot)$ denote 1×3 and 3×1 asymmetric kernel convolution, respectively. Moreover, $expand(\cdot)$ represents expanding operations. Given the F_0^{Squ}, F^{Hori}, and F^{Ver}, we fuse

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

them to generate the final asymmetric feature F^{Asy}:

$$F^{Asy} = \mathcal{F}_{1\times1}(F^{Hori} \oplus F^{Ver} \oplus F_0^{Squ}) \in \mathbb{R}^{h\times w\times c}, \quad (19)$$

where F_0^{Squ} represents the first output of square kernel enrichment. Finally, the new query feature F^{New} is produced by the F^{Asy} and F^{Squ}:

$$F^{New} = \mathcal{F}_{1\times1}(F^{Asy} \oplus F^{Squ}). \quad (20)$$

In the segmentation part, we first set a threshold τ to segment the query attention mask $M_q^A \in \mathbb{R}^{h\times w\times1}$. Then, we generate the pseudo foreground query mask $M_f^P \in \mathbb{R}^{h\times w\times1}$ and pseudo background query mask $M_b^P \in \mathbb{R}^{h\times w\times1}$. Next, the pseudo masks multiply the query generalization feature F_q^G and concatenate with the expanded main prototype P'_{main} to segment the object. The segmentation results from $M_f^{P'}$ and $M_b^{P'}$ are fused to obtain the pseudo query mask $M^P \in \mathbb{R}^{h\times w\times1}$:

$$M_f^{P'} = \text{segfore}((M_f^P \otimes F_q^G) \oplus P'_{\text{main}}),$$
$$M_b^{P'} = \text{segback}((M_b^P \otimes F_q^G) \oplus P'_{\text{main}}), \quad (21)$$
$$M^P = M_f^{P'} + M_b^{P'}, \quad (22)$$

where $segfore(\cdot)$ and $segback(\cdot)$ are the segment head. Finally, M^P is integrated into F^{New} by the residual structure to refine the final segmentation result M^{Final}:

$$M^{Final} = seg(\mathcal{F}_{conv}(F^{New} \oplus M^P) + F^{New}). \quad (23)$$

F. Extension to k-Shot Setting

In the case of the k-shot setting, more support images are given. To extend 1-shot segmentation to k-shot, we propose the appropriate way for the above modules. As for the high-level query attention mask M_q^H in DGSM, PFENet [20] proves that the high-level mask helps the model identify the target in query images. Therefore, to better identify targets and filter out background noise, we focus on the same region of different high-level masks in the k-shot setting by average fusion strategy. As for the mid-level query attention mask M_q^M in DGSM, CANet [26] demonstrates that Block-3 performs the best when the single block is used for comparison. In other words, the mid-level query attention mask contains rich object information. In the k-shot setting, different support images contain different support guiding information, so using average fusion to address the mid-level masks can weaken the information from different images. Therefore, we apply the max fusion strategy for the mid-level query attention mask to capture more object information. For other module components, we follow the previous works [20], [26], [27], [40], [43] to simply use the average fusion strategy to address them.

G. Training Loss

During training, the cross entropy loss is selected as the loss function of DGFPNet. Following FEM [20], we apply the intermediate supervision for the feature enrichment part to generate n losses $\mathcal{L}_1^i$, $i \in \{1, 2, \ldots, n\}$. $\mathcal{L}_1^i$ contains the $n-1$ different losses from square kernel enrichment and the 1 loss from asymmetric kernel enrichment. As for the segmentation part, we calculate the loss of foreground and background segmentation results, respectively. So, the pseudo loss $\mathcal{L}_2$ consists of the pseudo foreground loss and the pseudo background loss. Then, the final loss $\mathcal{L}_3$ is generated by the final segment result M^{Final}. In summary, the total loss is:

$$\mathcal{L} = \frac{\alpha}{n}\sum_{i=1}^{n}\mathcal{L}_1^i + \beta\mathcal{L}_2 + \mathcal{L}_3, \quad (24)$$

where n is 5 in the model. The α and β are trade-off parameters set to 1.0 in all experiments.

V. Experiments

A. Implementation Details

Datasets: Our model is evaluated in the datasets of PASCAL-5^i [7] and COCO-20^i [45]. PASCAL-5^i consists of PASCAL VOC 2012 dataset [46] and augmented SDS dataset [47], while COCO-20^i is built from MSCOCO [44]. We evenly divide the object categories of both datasets into four folds. For each fold, we randomly sample 1,000 pairs of support and query images for validation. The cross-validation experiment evaluates the proposed model: three for training and one for testing.

Experimental Setting: We construct our model on PyTorch. For a fair comparison with existing FSS methods, two different backbone networks are chosen, including VGG-16 [1] and ResNet-50 [2]. The backbones are pre-trained on ImageNet [42] and fixed during the training and testing stage. The SGD is adopted as the optimizer for training, where the momentum and weight decay are set to 0.9 and 0.0001, respectively. The learning rate is fixed at 0.005 for 200 epochs on PASCAL-5^i and 50 epochs on COCO-20^i. Meanwhile, we also apply the data augmentation strategies like random scale, Gaussian filtering, horizontal flip, and random rotation. All images are cropped into 473 $\times$ 473(PASCAL-5^i) and 641 $\times$ 641(COCO-20^i) for training and testing. The final testing result is computed by averaging the results of 5 trials with different random seeds. Our experiments run on Nvidia Tesla V100 GPUs.

Evaluation Metrics: As in [7], [18], [45], we adopt the mean Intersection-over-Union (mIoU) as the major performance evaluation metric because of its comprehensiveness and objectivity. For class C, the IoU is defined as $IoU_c = TP_c/(TP_c + FP_c + FN_c)$, where TP_c, FP_c, and FN_c are the number of true positives, false positives, and false negatives in segmentation masks. The mIoU is defined as the mean IoUs of all image classes. Moreover, we also report the foreground-background IoU (FB-IoU) for more comparisons. The FB-IoU is the mean of foreground IoU and background IoU over all test images, which ignores the image classes.

B. Comparison with State-of-the-arts

PASCAL-5^i: As shown in Table I, we build our model on two backbones, including VGG-16 and ResNet-50, respectively. It can be found that DGFPNet achieves new state-of-the-art performances on both 1-shot and 5-shot tasks with the two backbones.

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

TABLE I
MIOU AND FB-IOU PERFORMANCE OF 1-SHOT AND 5-SHOT SEGMENTATION ON PASCAL-5^i DATASET

Method	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
OSLSM [7](BMVC'17)	VGG-16	33.6	55.3	40.9	33.5	40.8	61.3	35.9	58.1	42.7	39.1	44.0	61.5
SG-One [9](TCYB'20)		40.2	58.4	48.4	38.4	46.3	63.9	41.9	58.6	48.6	39.4	47.1	65.9
RPMM [10](ECCV'20)		47.1	65.8	50.6	48.5	53.0	-	50.0	66.5	51.9	47.6	54.0	-
PFENet [20](TPAM'20)		56.9	68.2	54.4	52.4	58.0	72.0	59.0	69.1	54.8	52.9	59.0	72.3
MMNet [40](ICCV'21)		57.1	67.2	56.6	52.3	58.3	-	56.6	66.7	63.6	56.5	58.3	-
HSNet [15](ICCV'21)		59.6	65.7	59.6	54.0	59.7	73.4	64.9	69.0	64.1	58.6	64.1	76.6
APANet [38](TMM'22)		58.0	68.9	57.0	52.2	59.0	61.6	59.8	70.0	62.7	57.7	62.6	66.0
DCP [28](IJCAI'22)		59.7	68.7	63.8	53.1	61.3	74.9	64.3	70.7	67.4	61.1	65.8	79.4
DPCN [19](CVPR'22)		58.9	69.1	63.2	55.7	61.7	73.7	63.4	70.7	68.1	59.0	65.3	77.2
BAM [18](CVPR'22)		63.2	70.8	66.1	57.5	64.4	77.3	67.4	73.1	70.6	64.0	68.8	81.1
Baseline		54.7	71.4	57.4	60.6	61.0	74.3	63.8	71.8	67.6	62.3	66.4	78.0
DGFPNet(ours)		**66.9**	**72.4**	**68.7**	**60.8**	**67.2**	**79.1**	**71.4**	**74.7**	**73.3**	**66.3**	**71.4**	**82.6**
CANet [26](CVPR'19)	ResNet-50	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.6
RPMM [10](ECCV'20)		55.2	66.9	52.6	50.7	56.3	-	56.3	67.3	54.5	51.0	57.3	-
PPNet [12](ECCV'20)		48.6	60.6	55.7	46.5	52.8	69.2	58.9	68.3	66.8	58.0	63.0	75.8
PFENet [20](TPAM'20)		61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
ASGNet [11](CVPR'21)		58.8	67.9	56.8	53.7	59.3	69.2	63.7	70.6	64.2	57.4	64.0	74.2
HSNet [15](ICCV'21)		64.3	70.7	60.3	60.5	64.0	76.7	70.3	73.2	67.4	67.1	69.5	80.6
CyCTR [16](NeurIPS'21)		67.8	72.8	58.0	58.0	64.2	-	71.1	73.2	60.5	57.5	65.6	-
APANet [38](TMM'22)		62.2	70.5	61.1	58.1	63.0	63.6	63.3	72.0	68.4	60.2	66.0	66.7
DPNet [27](AAAI'22)		60.7	69.5	62.8	58.0	62.7	-	64.7	70.8	69.0	60.1	66.2	-
DPCN [19](CVPR'22)		65.7	71.6	69.1	60.6	66.7	78.0	70.0	73.2	70.9	65.5	69.9	80.7
BAM [18](CVPR'22)		69.0	73.6	67.6	61.1	67.8	79.7	70.6	75.1	70.8	67.2	70.9	82.2
Baseline		66.2	73.1	66.6	60.2	66.5	76.6	69.2	73.8	67.3	63.2	68.4	79.1
DGFPNet(ours)		**70.5**	**74.5**	**70.0**	**61.9**	**69.2**	**80.1**	**74.0**	**76.3**	**72.3**	**67.5**	**72.5**	**83.1**

The best performances are highlighted in bold. "Baseline" means the meta learner proposed in BAM [18], where the ASPP in it is replaced with FEM [20].

TABLE II
MIOU AND FB-IOU PERFORMANCE OF 1-SHOT AND 5-SHOT SEGMENTATION ON COCO-20^i DATASET

Method	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
FWB [45](ICCV'19)	ResNet-101	17.0	18.0	21.0	28.9	21.2	-	19.1	21.5	23.9	30.1	23.7	-
RPMM [10](ECCV'20)	ResNet-50	29.5	36.8	28.9	27.0	30.6	-	33.8	42.0	33.0	33.3	35.5	-
PPNet [12](ECCV'20)	ResNet-50	28.1	30.8	29.5	27.7	29.0	-	39.0	40.8	37.1	37.3	38.5	-
PFENet [20](TPAM'20)	ResNet-101	36.8	41.8	38.7	36.7	38.5	63.0	40.4	46.8	43.2	40.5	42.7	65.8
ASGNet [11](CVPR'21)	ResNet-50	-	-	-	-	34.6	60.4	-	-	-	-	42.5	67.0
CyCTR [16](NeurIPS'21)	ResNet-50	38.9	43.0	39.6	39.8	40.3	-	41.1	48.9	45.2	47.0	45.6	-
HSNet [15](ICCV'21)	ResNet-101	37.2	44.1	42.4	41.3	41.2	69.1	45.9	53.0	51.8	47.1	49.5	72.4
APANet [38](TMM'22)	ResNet-101	40.7	44.6	42.5	39.6	41.9	64.8	45.7	49.7	47.4	42.8	46.4	69.8
DCP [28](IJCAI'22)	ResNet-50	40.9	43.8	42.6	38.3	41.4	-	45.8	49.7	43.7	46.6	46.5	-
NTRENet [43](CVPR'22)	ResNet-50	36.8	42.6	39.9	37.9	39.3	68.5	38.2	44.1	40.4	38.4	40.3	69.2
DPCN [19](CVPR'22)	ResNet-50	42.0	47.0	43.2	39.7	43.0	63.2	46.0	54.9	50.8	47.4	49.8	67.4
BAM [18](CVPR'22)	ResNet-50	**43.4**	50.6	47.5	43.4	46.2	-	**49.3**	54.2	51.6	49.6	51.2	-
baseline	ResNet-50	41.1	52.2	46.1	45.7	46.3	70.8	43.4	54.9	48.8	48.0	48.8	72.5
DGFPNet(ours)	ResNet-50	43.1	**56.1**	**48.0**	**48.0**	**48.8**	**73.0**	48.4	**61.8**	**54.2**	**53.0**	**54.4**	**76.8**

The best performances are highlighted in bold. "Baseline" means the meta learner proposed in BAM [18], where the ASPP in it is replaced with FEM [20].

Specifically, with VGG-16 as the backbone, our model outperforms BAM [18], which holds the previous state-of-the-art results, with a margin of 2.8% and 2.6% for 1-shot and 5-shot tasks, respectively. With ResNet-50 as the backbone, DGFPNet surpasses 1.4% and 1.6% in 1-shot and 5-shot settings for BAM, respectively. Moreover, Table I shows the FB-IoU results with two different backbones. Compared to the BAM, our model improves the performances of all the settings, i.e., 1.8% (79.1% vs. 77.3%), 1.5% (82.6% vs. 81.1%), 0.4% (80.1% vs. 79.7%), and 0.9% (83.1% vs. 82.2%).

COCO-$\mathbf{20^i}$: The results of COCO-20^i are reported in Table II. We compare the segmentation results of our model and other models on COCO-20^i. As can be seen, our model still reaches new state-of-the-art results in both 1-shot and 5-shot settings, and it outperforms BAM [18] (previous SOTA) with mIoU gains of 2.6% and 3.2%, respectively. In Table II, using the FB-IoU as the evaluation metric, our model also achieves significant 3.9% (73.0% vs. 69.1%) and 4.4% (76.8% vs. 72.4%) FB-IoU improvement over others in both the 1-shot and 5-shot settings. These results prove that our model is capable of handling more challenging cases.

Segmentation Examples: To better understand our proposed model, we report some qualitative segmentation results generated from our DGFPNet, baseline model, ASGNet [11], and BAM [18] on the PASCAL-5^i and COCO-20^i, as shown in Fig. 8. BAM achieves the best results in all previous global

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

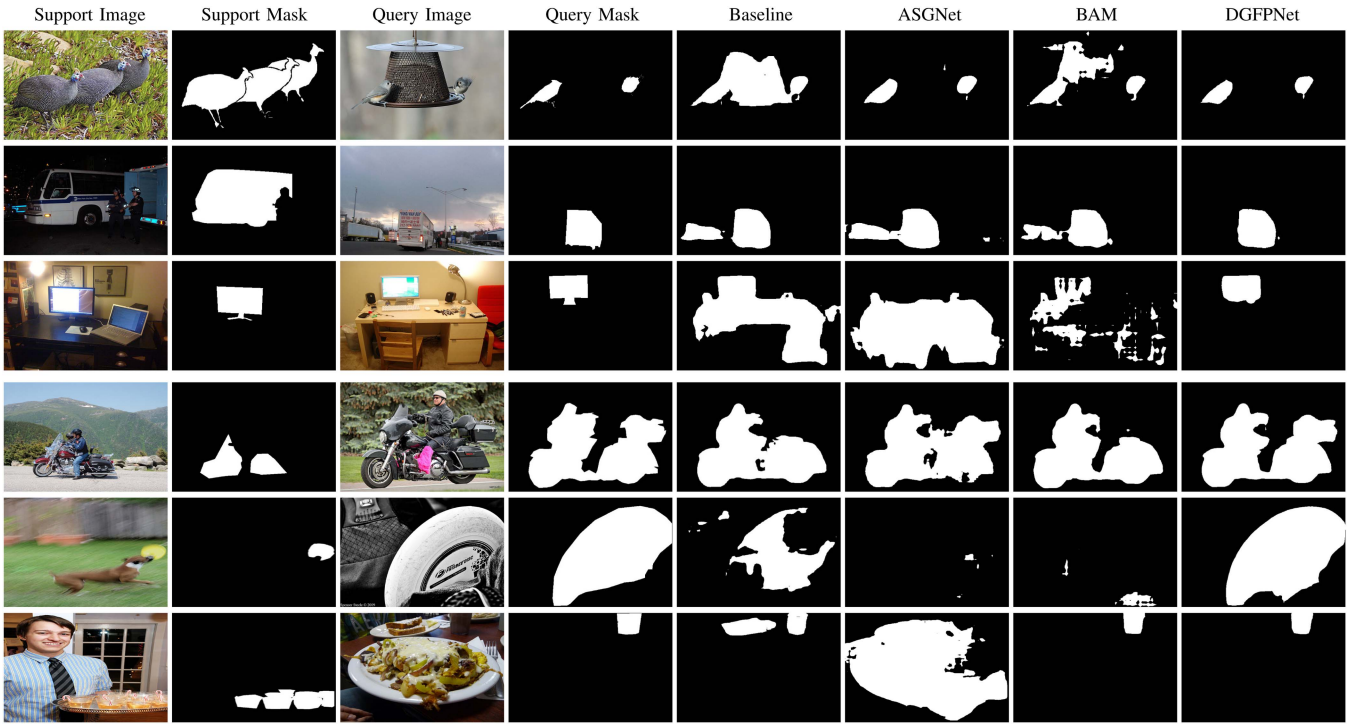


Fig. 8. Visual results of our model on PASCAL-5^i and COCO-20^i dataset. The top three samples are from PASCAL-5^i and the bottom three ones are from COCO-20^i. Each column from left to right represents the support image, support mask, query image, query mask, baseline prediction, ASGNet [11] prediction, BAM [18] prediction, and our DGFPNet prediction, respectively.

prototype models, and ASGNet reaches the best in all previous local prototype models. As for the global prototype, it mines the inherent and discriminative global features so that it can guide the model to distinguish the target object from other class objects. Therefore, in the last row of Fig. 8, our model and BAM can accurately segment the cup with the guidance of the global prototype. However, ASGNet suffers from the food and plate in the background by using local prototypes, leading to poor segmentation results. As for local prototypes, they extract the local information and details from the support images, especially the characteristics and edges of the object parts. In the first row of Fig. 8, the global texture of the bird in the support image is highly similar to the nest in the query image. So, with the guidance of global information in the support prototype, BAM classifies pixels in the nest as birds. In contrast, the local prototype can mine the characteristics and edges of the head and wing, guiding our model and ASGNet to suppress the disturbance from the nest and generate correct predictions. Compared to BAM and ASGNet, by applying frequency prototypes and bridging the gap between support and query images, our model shows superior capability in other more challenging scenarios, like complete segmentation, accurate classification, and barely background noise.

C. Ablation Study

We conduct a set of ablation experiments with ResNet-50 under the 1-shot setting on PASCAL-5^i to verify the effectiveness of the proposed modules.

Effect of MFP, FGM, and MDFEDM: To demonstrate the effectiveness of the proposed multi-frequency pooling (MFP), Feature Generalization Module (FGM), and Multi-Dimension Feature Enrichment Decoder Module (MDFEDM), we conduct ablation experiments on prototype generation, features, and decoder. Please note that we use the MFP to denote the Frequency Prototype Generation Module (FPGM) and Dual-Guided Selection Module (DGSM) and apply Generalized to denote the generalized feature F^G in FGM. In this ablation experiment, we set two baseline models that generate global or local prototypes by GAP or clustering. Specifically, we first apply the meta learner proposed in BAM [18] and replace the ASPP in the learner with FEM [20] to implement the global prototype baseline model. For the local prototype baseline model, the prototype generation module is established with the superpixel-guided clustering and guided prototype allocation proposed by ASGNet [11]; other modules are the same as the global prototype baseline model. Then, we orderly replace the corresponding module in the baseline models with our proposed module. As shown in Table III, when we only apply MFP to replace the prototype generation in two baseline models, class mIoU significantly increases by 0.8% (67.3% vs. 66.5%) and 3.0% (67.3% vs. 64.3%) compared to GAP and clustering, respectively. This proves that MFP captures more comprehensive and generalized object information than the other two methods. Besides, the results outperform the two baseline models by 0.7% (67.2% vs. 66.5%) and 1.4% (65.7% vs. 64.3%) when we only replace the features with F^G. These results show that FGM fuses more object information into the model than the original features. However, when

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III
ABLATION STUDIES ON PROTOTYPE GENERATION, FEATURES, AND DECODER

Prototype Generation			Features		Decoder		Fold-0	Fold-1	Fold-2	Fold-3	Mean
GAP	Clustering	MFP	Original	Generalized	FEM	MDFEDM					
✓			✓		✓		66.2	73.1	66.6	60.2	66.5
✓				✓	✓		66.5	73.5	67.4	61.2	67.2
✓			✓			✓	65.8	72.4	67.1	61.2	66.6
✓				✓		✓	65.9	73.2	66.8	**61.9**	67.0
	✓		✓		✓		63.6	71.5	63.6	58.5	64.3
	✓			✓	✓		65.5	72.1	65.4	59.6	65.7
	✓		✓			✓	62.8	71.5	64.7	57.6	64.2
	✓			✓		✓	66.3	72.1	65.3	59.7	65.9
		✓	✓		✓		69.1	73.7	66.6	59.8	67.3
		✓		✓	✓		69.6	74.4	66.6	60.5	67.8
		✓	✓			✓	70.4	73.9	69.6	60.6	68.6
		✓		✓		✓	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**

GAP/clustering, original, and FEM denote global average pooling/clustering, original query features, and feature enrichment module [20] in the global/local prototype baseline model, respectively. MFP, generalized, and MDFEDM denote multi-frequency pooling, generalized feature in feature generalization module, and multi-dimension feature enrichment decoder module, respectively.

FEM is replaced with MDFEM in two baseline models, the performance improves barely (66.6% vs. 66.5%) and even drops (64.2% vs. 64.3%). We ascribe this phenomenon to the abundant learned parameters introduced by MDFEM, which reduces the generalization ability of the model. In other words, without further capturing valuable object information into the model, only enhancing the decoder obtains little effect, or even worse. A similar phenomenon can also be found in the combination of GAP/Clustering, Generalized, and MDFEDM (67.0% vs. 67.2% or 65.9% vs. 65.7%). Therefore, when MFP further mines the complete and generalized object information, MDFEM brings 1.3% (68.6% vs. 67.3%) improvement in mIoU. Moreover, FGM also helps MFP to fuse more object information into the model, i.e., 67.8% vs. 67.3%. Finally, when all modules are replaced, the performance further improves and achieves new state-of-the-art.

Visualization of the FPGM Process: In Fig. 9, we visualize the process of the Frequency Prototype Generation Module (FPGM) to illustrate clearly the mining object information way and the advantage of multi-frequency pooling (MFP). From Fig. 9(a) to (c), we visualize the procedure results of MFP in the part DCT bases. Next, to thoroughly analyze the MFP's capability of capturing information, we take the 9-th DCT base as an example to show all the details about the MFP process in Fig. 9(d).

As shown in Fig. 9(d), the MFP is mainly responsible for strengthening and suppressing objects. The positive values in the DCT base strengthen some object features, and the negative values in the DCT base suppress some object features. Meanwhile, the values close to 0 are neglected since they hardly work for feature extraction. By multiplying foreground features with the 9-th DCT base, we obtain the strengthened and suppressed object features from this DCT base. Then, these changed features are compressed into a prototype, which contains the dominant effect object information in this DCT base. Therefore, when the effect of this prototype is activated by computing the cosine similarity with the original masked object feature, we know this prototype primarily involves the positive information of plane empennage and plays a strengthening role in the plane empennage, as shown in the prototype activation map.

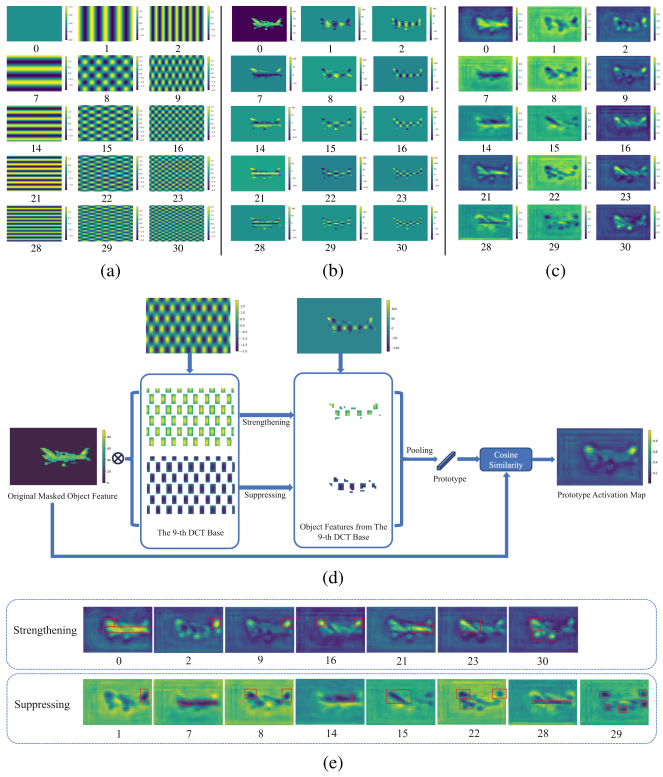


Fig. 9. Visualization of the FPGM Process, where (a), (b), and (c) denote the DCT bases, activation maps generated by multiplying foreground support features with DCT bases, and activation maps generated by frequency prototypes, respectively. Please note that the lower and right images contain higher-frequency information. (d) visualizes the process of MFP in the 9-th DCT base. (e) represents the effect of two class prototypes.

Letting $Z = \sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[M_s(i,j)=1]$, we can obtain:

$$P_t = \frac{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} F_{s_3}^{M'}(i,j)\cdot B_t(i,j)}{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[M_s(i,j)=1]}$$

$$= \frac{1}{Z}\alpha(u)\alpha(v)\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} F_{s_3}^{M'}(i,j)$$

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION STUDIES ON EACH BRANCH OF DGSM

QGB	SGB	Fold-0	Fold-1	Fold-2	Fold-3	Mean
		65.9	73.2	66.8	**61.9**	67.0
✓		69.8	74.0	69.4	61.1	68.6
	✓	69.2	74.0	69.0	61.3	68.4
✓	✓	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**

QGB and SGB denote query-guided branch and support-guided branch, respectively.

$$\cdot \cos\left(\frac{\pi u}{h}\left(i+\frac{1}{2}\right)\right) cos\left(\frac{\pi v}{w}\left(j+\frac{1}{2}\right)\right) = \frac{1}{Z}F_{dct}(u,v), \tag{25}$$

where $F_{dct}(u,v)$ denotes the traditional standard DCT. Therefore, P_t is proportional to the transform coefficient in (u,v), carrying the object features in this frequency. When analyzing the activation maps in Fig. 9(c) and dividing these maps into two classes (strengthening and suppressing) in Fig. 9(e), we can see that different frequency prototypes contain different positive or negative object information, such as fuselage, wing, empennage, and wheel. Another important observation is that the prototype contains information from global constituents to local details when the frequency increases from low to high, no matter the strengthening or suppressing effects.

Hence, our proposed MFP addresses the limitation of GAP and clustering, and generates comprehensive and informative frequency prototypes from both global and local perspectives in a unified way.

Moreover, we also find in Fig. 9(c) that if the prototypes represent correct global and local object information, these corresponding prototype masks are more similar to the ground truth mask than those with some background noise. Therefore, DGSM selects the more effective frequency prototypes based on the Euclidean distance values between the guided mask and prototype masks, including the lower or higher frequency prototypes.

Branches in DGSM: The Dual-Guided Selection Module (DGSM) mainly contains two branches: the query-guided branch and support-guided branch. Table IV presents our validation experiment on the effectiveness of each branch, where the first line result is generated by our model without QGB and SGB (only using GAP to extract support object information). When we only add QGB to help our model capture rich and generalized object features from query images, the performance improves from 67.0% to 68.6%, contributing to a 1.6% performance gain. Similarly, when applying SGB to mine more support object features by MFP, we also obtain an impressive performance improvement of 1.4% compared to using the GAP. Moreover, with the object information from both support and query images, we can obtain another 0.6% and 0.8% performance gain respectively, and improve the result to 69.2%. Obviously, with the guidance of the two above branches, DGSM effectively bridges the gap between the support set and query set and captures complete object information.

Prior Information in DGSM: We conduct ablation experiments to demonstrate the effectiveness of the different prior mask information in Dual-Guided Selection Module (DGSM). In Table V, M_q^H, M_q^M, and M^B denote the prior attention mask from the high-level feature, mid-level feature, and base learner, respectively. The first line result of the table is only generated by M_q^H (M_q^H is used in our baseline model), achieving 67.0% mIoU performance. Then, we add M_q^M or M^B to mine more target information or filter out the base object in the background, bringing 0.5% and 0.6% performance gain compared to only using M_q^H. However, the performance drops greatly without using M_q^H, especially for the Fold-3. This phenomenon strongly demonstrates the M_q^H is the essential part for suppressing the background noises and identifying the object. When applying all the M_q^H, M_q^M, and M^B to generate our final query attention mask M_q^A, it outperforms others with a large margin, i.e., 69.2% vs. 67.6%, 69.2% vs. 67.5%, 69.2% vs. 67.0%. So, we argue that these three masks are complementary and necessary, namely suppressing background noises by M_q^H, mining object information by M_q^M, and filtering out the base object by M^B.

TABLE V
ABLATION STUDIES ON THE PRIOR INFORMATION OF HIGH-LEVEL, MID-LEVEL, AND BASE LEARNER IN DGSM

M_q^H	M^B	M_q^M	Fold-0	Fold-1	Fold-2	Fold-3	Mean
✓			65.9	73.2	66.8	**61.9**	67.0
✓	✓		68.6	74.1	67.2	60.4	67.6
✓		✓	67.6	73.4	67.9	61.0	67.5
	✓	✓	67.0	74.2	67.8	58.9	67.0
✓	✓	✓	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**

M_q^H, M_q^H, and M^B denote the prior attention mask from the high-level feature, mid-level feature, and base learner, respectively.

TABLE VI
ABLATION STUDIES ON THE NUMBER (Λ) OF SELECTED INDEXES IN DGSM

Λ	Fold-0	Fold-1	Fold-2	Fold-3	Mean
5	70.3	73.8	69.0	61.2	68.6
10	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**
15	69.7	74.1	69.4	61.4	68.7

TABLE VII
ABLATION STUDIES ON THE FEATURES IN FGM

	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Original Feature	70.4	73.9	69.6	60.6	68.6
Original Feature + SGF	70.3	74.0	69.3	61.5	68.8
QGF	69.9	73.7	**70.3**	61.6	68.9
QGF + SGF	**70.5**	**74.5**	70.0	**61.9**	**69.2**

QGF and SGF denote query generalization feature and support generalization feature, respectively.

Number of Selected Indexes in DGSM: We conduct ablation experiments to analyze the number (Λ) of selected indexes with minimum distance in Dual-Guided Selection Module (DGSM). Table VI shows that our model performs best when Λ is set to 10. It suggests that we extract more informative complemental prototypes and generate a more accurate query attention mask in this setting. Therefore, we set Λ to 10 for our model.

Analysis of FGM: Feature Generalization Module (FGM) aims to integrate more generalized features into the model. Table VII shows the ablation study on our proposed FGM. It can be seen that the performance of original feature achieves

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
ABLATION STUDIES ON EACH COMPONENT OF MDFEDM

	Fold-0	Fold-1	Fold-2	Fold-3	Mean
SKE	69.6	74.4	66.6	60.5	67.8
SKE+AKE	69.0	74.4	68.4	60.7	68.1
SKE+SP	69.3	73.9	68.6	61.3	68.3
SKE+AKE+SP	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**

SKE, AKE and SPC denote square kernel enrichment, asymmetric kernel enrichment and segmentation part, respectively.

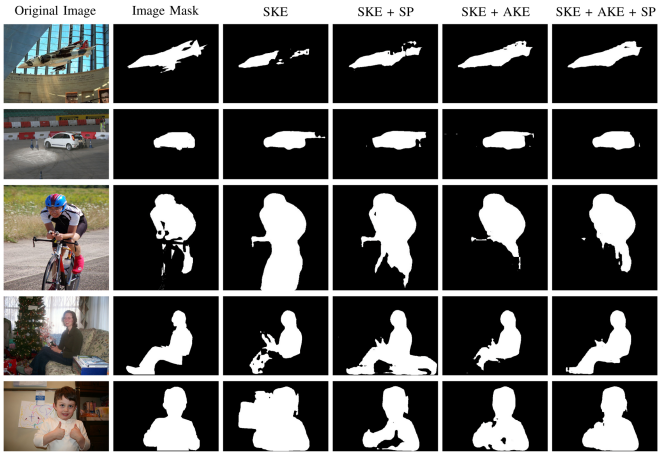


Fig. 10. Visual results by using the components in MDFEDM. SKE, AKE, and SP denote square kernel enrichment, asymmetric kernel enrichment, and segmentation part, respectively. The first and second rows show MDFEDM captures more horizontal banding features. The third and fourth rows show MDFEDM captures more vertical banding features. Moreover, the fifth row shows that MDFEDM also refines the segmentation, such as fliting out the background noise and complementing objects.

TABLE IX
ABLATION STUDIES ON THRESHOLD VALUE (τ) IN MDFEDM

Threshold Value τ	Fold-0	Fold-1	Fold-2	Fold-3	Mean
0.25	70.0	73.8	69.1	61.3	68.6
0.30	68.9	73.3	69.6	61.1	68.2
0.35	**70.5**	**74.5**	**70.0**	**61.9**	**69.2**
0.40	69.9	74.0	68.3	61.3	68.4
0.45	69.8	**74.5**	68.9	61.0	68.6

TABLE X
ABLATION STUDY ON k-SHOT FUSION STRATEGIES

	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Average	73.6	76.0	70.7	67.4	71.9
Max	72.5	**76.3**	71.8	66.6	71.8
Average + Max	**74.0**	**76.3**	**72.3**	**67.5**	**72.5**

Average denotes the average fusion strategy for all masks in DGSM. Max denotes the max fusion strategy for all masks in DGSM. Average + max denotes the average fusion strategy for M_q^H and the max fusion strategy for M_q^M in DGSM.

TABLE XI
ABLATION STUDIES ON THE TRADE-OFF PARAMETERS α OF INTERMEDIATE LOSS $\mathcal{L}_1^i$ AND β OF PSEUDO LOSS $\mathcal{L}_2$

α	β	Fold-0	Fold-1	Fold-2	Fold-3	Mean
1.0	0.5	69.7	**74.6**	68.4	61.2	68.5
0.5	1.0	70.4	74.0	69.4	61.0	68.7
1.0	1.0	**70.5**	74.5	**70.0**	**61.9**	**69.2**

improvement by concatenating the support generalization feature (SGF). Moreover, the query generalization feature (QGF) result is much better than the original. The performance is further improved from 68.9% to 69.2% when SGF is fused into QGF. The experiment results show the effectiveness of the FGM.

Components in MDFEDM: Multi-Dimension Feature Enrichment Decoder Module (MDFEDM) consists of square kernel enrichment, asymmetric kernel enrichment, and segmentation part. We evaluate the effectiveness of each component. As shown in Table VIII, when we only add the asymmetric kernel enrichment or the segmentation part for the mask prediction, it achieves 0.3% and 0.5% mIoU improvement over the square kernel enrichment, respectively. The performance of adding all parts further improves by 1.1% and 0.9%, respectively. These results show that the asymmetric kernel enrichment and segmentation part play an essential role in object segmentation. Moreover, Fig. 10 presents the segmentation results using the components in MDFEDM. It shows that asymmetric kernel enrichment captures more banding object features, and the segmentation part refines the object segmentation better by tackling hard pixels. With the square kernel enrichment, asymmetric kernel enrichment, and segmentation part, MDFEDM achieves more accurate and superior segmentation results.

Effects of Threshold Value: We explore the influence of threshold value τ that segments the query attention mask in the segmentation part. Table IX shows the results when τ is varied from 0.25 to 0.45. It achieves the best results when $\tau = 0.35$. Hence, τ is set to 0.35 in our model.

***k**-shot Fusion Strategies:* In the k-shot setting, we compare three fusion strategies in Table X. We can see that our strategy brings 0.6% and 0.7% performance gains compared to the average strategy and max strategy, respectively. This result indicates that, in Dual-Guided Selection Module (DGSM), it is suitable for the high-level query attention mask M_q^H to filter out the noise by average fusion strategy, and the mid-level query attention mask M_q^M is appropriate to mine more object information by max fusion strategy. Therefore, we apply the average-max fusion strategy in DGSM.

Effect of Loss Weight: We investigate the impact of the trade-off parameters α and β presented in (24), where α and β mean the balancing weights of intermediate loss $\mathcal{L}_1^i$ and pseudo loss $\mathcal{L}_2$, respectively. As shown in Table XI, we find that the best and most stable results are achieved when α and β are both set to 1.0. For example, when α and β are set to 1.0 and 0.5, respectively, we obtain the best result on Fold-1. However, the results on Fold-0 and Fold-2 are the worst. Similarly, when α and β are set to 0.5 and 1.0, respectively, the results are also unstable, where Fold-0 and Fold-2 are better while Fold-1 is worse. Another important observation is that these two losses are complementary to each other. As a result, we set α and β to the same value, i.e., 1.0.

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

VI. Conclusion

In this paper, we design DGFPNet for the few-shot semantic segmentation task. To address the limitation of GAP and clustering, we propose the FPGM that mines the global and local object features and extracts various frequency prototypes by MFP. With the input of frequency prototypes, our proposed DGSM generates the query attention mask and complementary prototypes under the support and query information guidances. Then, the obtained mask is incorporated into the features in FGM to improve its generalization. Given concatenated features, MDFEDM further mines the multi-dimension object features by square and asymmetric kernels and tackles hard pixels by pre-segmenting query attention mask to refine final segmentation results. Extensive experiments verify the superiority of DGFPNet. From the frequency domain perspective, our model generalizes the prototype-based method to mine complete object information. Meanwhile, our model diminishes the domain gap and retains high generalization. In the future, we will explore more possibilities about further utilizing the rich image information in the frequency domain.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," in *Proc. Int. Conf. Learn. Representations*, 2015.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[7] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–9.

[8] N. Dong and E.P. Xing, "Few-shot semantic segmentation with prototype learning," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 79–91.

[9] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-One: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020.

[10] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 763–778.

[11] G. Li et al., "Adaptive prototype learning and allocation for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8334–8343.

[12] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 142–158.

[13] C. Zhang et al., "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9587–9595.

[14] W. Liu, C. Zhang, G. Lin, and F. Liu, "CRNet: Cross-reference networks for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4165–4173.

[15] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6921–6932.

[16] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," in *Proc. Adv. Neural Inform. Process. Syst.*, 2021, vol. 34, pp. 21984–21996.

[17] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 783–792.

[18] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8057–8067.

[19] J. Liu et al., "Dynamic prototype convolution network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11553–11562.

[20] Z. Tian et al., "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022.

[21] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," Dec. 2017, *arXiv:1706.05587*.

[22] Z. Huang, C. Wang, X. Wang, W. Liu, and J. Wang, "Semantic image segmentation by scale-adaptive networks," *IEEE Trans. Image Process.*, vol. 29, pp. 2066–2077, 2019.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[25] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3146–3154.

[26] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5217–5226.

[27] B. Mao et al., "Learning from the target: Dual prototype network for few shot semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 1953–1961.

[28] C. Lang, B. Tu, G. Cheng, and J. Han, "Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 1024–1030.

[29] K. Xu et al., "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1740–1749.

[30] W. Shu, J. Wan, K. Tan, S. Kwong, and A. Chan, "Crowd counting in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19618–19627.

[31] R. Geirhos et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–20.

[32] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8684–8694.

[33] C. Luo et al., "Frequency-driven imperceptible adversarial attack on semantic similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15315–15324.

[34] X. Li et al., "Improving semantic segmentation via decoupled body and edge supervision," in*Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 435–452.

[35] L. Liu et al., "Dynamic extension nets for few-shot semantic segmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1441–1449.

[36] B. Kang, Y. Lee, and T. Q. Nguyen, "Depth-adaptive deep neural network for semantic segmentation," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2478–2490, Sep. 2018.

[37] W. Liu, C. Zhang, H. Ding, T.-Y. Hung, and G. Lin, "Few-shot segmentation with optimal transport matching and message flow," *IEEE Trans. Multimedia*, vol. 25, pp. 5130–5141, 2023, doi: 10.1109/TMM.2022.3187855.

[38] J. Chen et al., "APANet: Adaptive prototypes alignment network for few-shot semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 4361–4373, 2023, doi: 10.1109/TMM.2022.3174405.

[39] T. Chen et al., "Semantically meaningful class prototype learning for one-shot image segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 968–980, 2022.

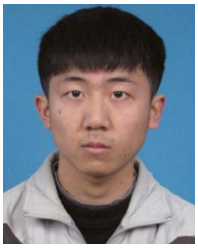
[40] Z. Wu, X. Shi, G. lin, and J. Cai, "Learning meta-class memory for few-shot semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 517–526.

[41] W. Liu, Z. Wu, H. Ding, F. Liu, J. Lin, and G. Lin, "Few-shot segmentation with global and local contrastive learning," Aug. 2021, *arXiv:2108.05293*.

[42] J. Deng et al., "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.

[43] Y. Liu et al., "Learning non-target knowledge for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11573–11582.
[44] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[45] K. Nguyen and S. Todorovic, "Feature weighting and boosting for fewshot segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 622–631.
[46] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[47] B. Hariharan, P. Arbel' aez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 991–998.
[48] X. Luo et al., "PFENet : Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1273–1289, Feb. 2024.
[49] Z. Zheng et al., "Quaternion-Valued correlation learning for few-shot semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2102–2115, May 2023.



Chunlin Wen (Member, IEEE) received the B.E. degree from Dalian University, Dalian, China, in 2018. He is currently working toward the M.E. degree with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. His research interests include computer vision and machine learning, specifically for prototype learning and few-shot learning.

Hui Huang (Member, IEEE) received the B.S. and M.S. degrees from the School of Biomedical Engineering, Southeast University, Nanjing, China, in 2004 and 2007, and the Ph.D. degree from the Department of Information and Image Processing from the University of Rennes, Rennes, France, in collaboration with Institute MINES-TELECOM, TELECOM Bretagne, and LaTIM Laboratory, INSERM U1101, Brest, France, in 2011. She is currently an Associate Professor of the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. She is also a member of the ACM. Her primary research interests include image processing, computer vision and pattern recognition. Her research interests include medical image processing, deep learning, computer vision, and signal processing.

Yan Ma (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. She is currently a Professor of the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai. Her research interests include data mining, image segmentation, and action recognition.

Feiniu Yuan (Senior Member, IEEE) received the B.Eng. and M.E. degrees in mechanical engineering from the Hefei University of Technology, Hefei, China, in 1998 and 2001, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the University of Science and Technology of China (USTC), Hefei, in 2004. From 2004 to 2006, he was a Postdoctor with the State Key Lab of Fire Science, USTC. From 2010 to 2012, he was a Senior Research Fellow with Singapore Bioimaging Consortium, Agency for Science, Technology and Research (A*STAR), Singapore. He is currently a Professor, a Ph.D. Supervisor and the Vice Dean with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. His research interests include deep learning, image segmentation, pattern recognition and 3D modeling. He is a senior member of CCF.

Hongqing Zhu (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000. She is currently a Professor with the East China University of Science and Technology, Shanghai, China. From 2003 to 2005, she was a Postdoctoral Fellow with the Department of Biology and Medical Engineering, Southeast University, Nanjing, China. Her current research interests include deep learning, computer vision, pattern recognition, and medical image processing. She is a member of IEICE.

Authorized licensed use limited to: Shanghai Normal University. Downloaded on September 24,2024 at 00:55:58 UTC from IEEE Xplore. Restrictions apply.