Computing and Informatics, Vol. 44, 2025, 92–123, doi: 10.31577/cai_2025_1_92

FEW-SHOT SEMANTIC SEGMENTATION WITH FREQUENCY PROTOTYPE LEARNING

Chunlin Wen

College of Information, Mechanical and Electrical Engineering
Shanghai Normal University, Shanghai, China

Huang Hui*, Yan Ma, Feiniu Yuan

College of Information, Mechanical and Electrical Engineering
Shanghai Engineering Research Center of Intelligent Education and Bigdata
Shanghai Normal University, Shanghai, China
e-mail: huanghui@shnu.edu.cn

Hongqing Zhu

School of Information Science and Engineering
East China University of Science and Technology
Shanghai, China

Peng Zhu

Shanghai Vixdetect Inspection Equipment Co., Ltd
Shanghai, China

Abstract. Few-shot semantic segmentation is a challenging task aimed at segmenting new objects in the query image with only a few annotated support images. Most advanced methods for this task mainly focus on either global or local prototype learning through global average pooling (GAP) or clustering. However, due to

* Corresponding author

the limitation of average and cluster operation, these methods still fail to exploit the object information from support images entirely. To address these limitations, we propose a generalization of prototype learning in the frequency domain through multi-frequency pooling (MFP) to mine both local and global object information. Based on the MFP, we further build a Frequency Prototype Network (FPNet) consisting of three novel designs. Firstly, the Frequency Prototype Generation Module (FPGM) extracts frequency prototypes by MFP in the DCT domain to provide complete object guidance information. Then, the Prior Attention Mask Module (PAMM) produces a prior attention mask to identify a query target more precisely and retain high generalization. Finally, the Frequency Prototype Selection Module (FPSM) selects the most effective support prototypes to reduce redundancy. Extensive experiments on PASCAL-5^i and COCO-20^i demonstrate that our model achieves state-of-the-art performances in both 1-shot and 5-shot settings.

Keywords: Few-shot segmentation, few-shot learning, prototype learning, frequency domain learning

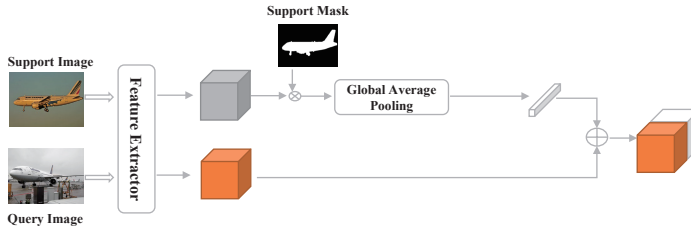
1 INTRODUCTION

In recent years, significant progress has been made in various computer vision tasks, particularly in semantic segmentation [1, 2, 3], owing to the remarkable advancements of deep convolutional neural networks [4, 5, 6]. Semantic segmentation aims to assign each pixel to a specific class. However, traditional semantic segmentation heavily relies on large amounts of annotated images [7], which are time-consuming and labour-intensive. Furthermore, when there is a lack of densely annotated training images, the performance of these frameworks drops dramatically.

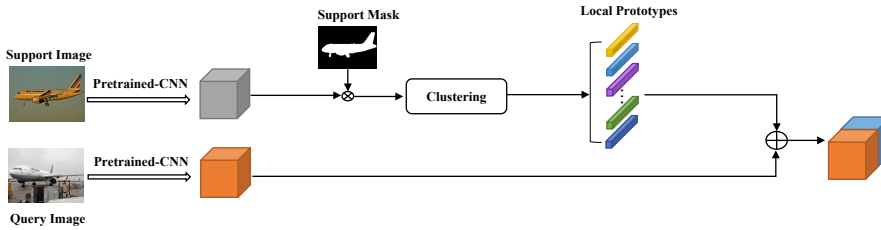
To tackle this issue, the few-shot image segmentation [8, 9, 10, 11, 12] is proposed to segment novel class objects in a query image with one or a few annotated support images. The keys to the few-shot image segmentation are:

1. How to mine the support object information to guide the segmentation of query images;
2. How to generalize the trained model to the novel classes.

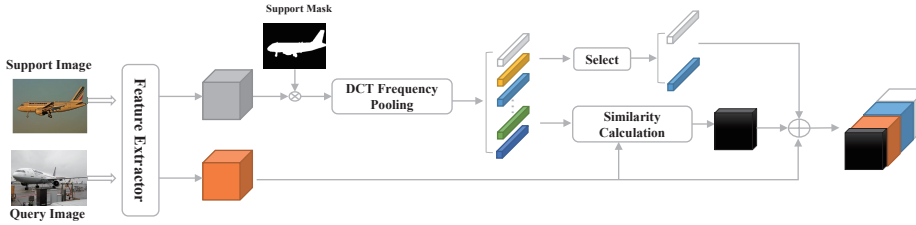
For the first key, most current methods [13, 14, 15, 16, 17] usually extract only a single prototype by masked global average pooling (GAP) from support images to store the object information and guide the segmentation process, as shown in Figure 1 a). However, the single prototype is prone to losing object details, essential clues for complete segmentation. Later, as shown in Figure 1 b), some methods [18, 19] extract prototypes by clustering the local features. Unfortunately, they ignore the global recognition of the object. Besides, other models [20, 21, 22] concurrently make predictions for both the support and query images to find the co-occurrent objects or mine lost critical information. However, the support image prediction



a) The global prototype model



b) The local prototype model

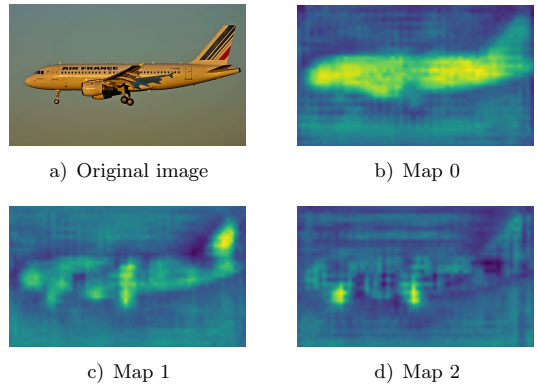


c) The multi-frequency prototype model

Figure 1. Illustrating the model difference between a) the global prototype model, b) the local prototype model, and c) the proposed multi-frequency prototype model. The global prototype model produces a single prototype by average pooling from the global perspective, and the local prototype model produces multi prototypes by clustering from the local perspective. In contrast, we utilize different frequency components to generate multi prototypes by multi-frequency pooling from global and local perspectives.

often increases the burden of the model and leads to more parameters toward base classes. For the second key, some existing models [14, 23, 24] retain the generalization with the attention map. However, these attention maps also have several limitations, like positioning imprecisely [14], introducing excessive learnable parameters [23], incurring background noises [24], etc. This paper investigates the two keys from the frequency domain to address the above problems.

For the first key, we generalize prototype learning in the frequency domain and propose multi-frequency pooling (MFP). Compared to GAP and clustering, MFP entirely mines the support object information with some frequency components of



a) Original image b) Map 0

c) Map 1 d) Map 2

Figure 2. Visualization of activation maps generated by multi-frequency pooling. Map 0 generated from DCT^0 focuses on the global features like the airplane's fuselage, but Map 1 and Map 2 generated from DCT^2 and DCT^5 pay more attention to the local features like the empennage and wheel of the airplane. Note that Map 0 is equal to the map generated by GAP. Best viewed in color.

discrete cosine transform (DCT) from both a global and local perspective. As shown in Figure 2, MFP guides the activation of global and local features in the target image, e.g., the fuselage, empennage, and wheel in the airplane, while GAP only guides the activation of the fuselage. Moreover, MFP also focuses on the object edge and background information ignored by GAP and clustering (Figure 9 shows more details about MFP). Based on the MFP, we propose Frequency Prototype Generation Module (FPGM) to generate frequency prototypes containing complete and rich support image information.

For the second key, we propose the Prior Attention Mask Module (PAMM) to obtain a prior attention mask for query images by uniting the different frequency prototypes. Different from previous attention maps [14, 23, 24], our prior attention mask is training-free, more precise, and less noisy. Specifically, we utilize the mid-level features and some low background similarity frequency prototypes to produce the fine attention mask for the query image, which contains both the global and local information of the object. However, this mask also incurs background noise similar to the target features. Therefore, we use the high-level features to obtain the coarse prior attention mask to filter the noise in the background. By combining the fine prior attention mask and the coarse prior attention mask, we get the final mask to enhance the model's generalization capability and improve the model's segmentation performance.

Based on the two modules proposed above, we establish the Frequency Prototype Network (FPNet), as shown in Figure 1 c). Moreover, to reduce model complexity and computational burden, we introduce the Frequency Prototype Selection Module (FPSM) in FPNet to select the most representative prototypes (base prototype and

complement prototype) from all frequency prototypes. The base prototype contains the global object information from the low-frequency component and the complement prototype compresses the object details from the other frequency components. Given the prior attention mask, query features, and prototypes, we use the FEM [24] to predict the final query results.

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

- To overcome the limitation of average and cluster operation, we generalize traditional GAP in the frequency domain and propose a novel multi-frequency pooling (MFP) method. It not only inherits the merits of GAP to capture inherent global features of the object but also focuses on local details.
- We propose the Prior Attention Mask Module (PAMM) to generate a prior attention mask established on frequency prototypes. This mask can help to focus on the query target more precisely and enhance generalization capability. To improve the model efficiency, we propose Frequency Prototype Selection Module (FPSM) that selects the most effective guidance prototypes from all the frequency prototypes.
- Our FPNet achieves the state-of-the-art results on PASCAL-5^i and COCO-20^i datasets. Extensive experiments validate the effectiveness of each component in our FPNet.

2 RELATED WORK

2.1 Semantic Segmentation

Semantic segmentation has achieved astonishing success in recent years, aiming to classify each image pixel based on large-scale labeled datasets [25, 26]. Most semantic segmentation methods are based on a fully convolutional network (FCN) [1] that replaces the fully connected layer with the convolution layer. Then, the state-of-the-art models are typically attentive to the larger receptive field [27, 28], the multi-scale feature aggregation [2, 29], and the encoder-decoder architecture [3, 30, 31, 32]. For example, DeepLab [27, 28] uses dilated convolutions to capture a larger context and design an Atrous Spatial Pyramid Pooling (ASPP). PSPNet [2] utilizes a Pyramid Pooling Module (PPM) to fuse multi-scale features. U-Net [3] proposes a symmetric encoder-decoder network to reconstruct high-resolution prediction maps. However, powerful traditional segmentation models cannot generalize well with a few annotated samples when segmenting objects of unseen classes.

2.2 Few-Shot Segmentations

In contrast to traditional semantic segmentation, few-shot segmentation is established to segment new classes in an image with only a few support examples. OSLSM [8] firstly proposes a two-branch network to solve the task, which consists of the condition branch to extract object information from the support images

and the segmentation branch to segment the object in the query images with the guidance of extracted information. Inspired by prototypical networks, PL [13] learns the prototype by GAP from the support set to segment query targets. The subsequent works [14, 15, 16, 17, 20] follow the prototype-based methods. SG-One [17] computes the cosine similarity to build the relationship between the guidance information and the query features to improve the prediction. PFENet [14] overcomes spatial inconsistency by designing the Feature Enrichment Module (FEM) to enrich query features with support features and prior masks adaptively. However, all the above approaches use a single prototype from the GAP, representing only the global support image information. Later studies [24, 33, 18, 34] focus on multiple prototypes to enhance the single prototype model. Based on the expectation-maximization (EM) algorithm, RPMM [33] generates multiple prototypes correlating diverse image regions. REF [24] harvests sufficient and prosperous guidance from global, peak, and adaptive embedding. However, they fail to recognize the object from global and local perspectives. Besides, some studies [21, 22] generate prototypes from other perspectives to capture more information. SCL [21] mines the lost critical information and aggregates both primary and auxiliary support vectors for better segmentation performance. DPCN [22] utilizes the visual reasoning results of support images to derive many different ?proxies? from a broader episodic perspective. However, the additional prototypes are generated by predicting the support image, leading to excessive learnable parameters and generalization ability reduction. Unlike all the existing few-shot semantic segmentation methods, we produce multiple prototypes in the frequency domain that could capture rich and complete object features without learnable parameters.

2.3 Frequency Domain Learning

Frequency analysis has been widely adopted to process image tasks in recent years. Some studies [35, 36] introduce frequency information into the model to reduce complexity or boost reasoning. In [36], the proposed method applies a learning-based dynamic channel selection strategy to remove the trivial frequency components in the input. Gueguen et al. [35] directly compute the block-wise DCT coefficients as part of the JPEG codec to accelerate image processing in the training stage. Other studies [37, 38, 39, 40] focus on the performance of the low-frequency and high-frequency components. Wang et al. [39] show that CNN may capture high-frequency components (HFC) to improve accuracy, which is misaligned with human visual preference. The research in [40] demonstrates that CNN shows a more significant bias towards learning low-frequency local features than humans. SF-TransT [37] jointly models Gaussian spatial prior and low-/high- frequency information for visual object tracking. MSFS-Net [38] restores the high-frequency features and retains the low-frequency features during the process of deblurring. Besides, FcaNet [41] proposes a novel multi-spectral channel attention in the frequency domain. DoG-LSTM [42] integrates a pyramid of Difference of Gaussians

(DOG) to attenuate high-frequency local components in the feature space and reduce the inductive texture bias on CNNs. SSAH [43] introduces the low-frequency constraint to limit perturbations within high-frequency components to ensure perceptual similarity between adversarial examples and origin. In this paper, our work focuses on prototype learning in the frequency domain to enhance the prototype representation.

3 METHOD

In this section, we introduce the proposed FPNet for few-shot image segmentation. Firstly, we introduce the task definition and illustrate the overall network architecture of FPNet from a global perspective. Then, we describe three modules of FPNet in detail, respectively: the Frequency Prototype Generation Module (FPGM), the Prior Attention Mask Module (PAMM), and the Frequency Prototype Selection Module (FPSM). Finally, we elaborate on the k-shot setting in FPNet.

3.1 Task Definition

We follow the standard few-shot semantic segmentation setting [8, 14]. The dataset is divided into a training set D_{train} and a testing set D_{test}. The classes in the training set D_{train} are named base classes C_{base}, and the classes in the testing set D_{test} are named novel classes C_{novel}, $C_{base} \cap C_{novel} = \oslash$. In episodes, the model is trained on C_{base}. After that, it is directly tested on C_{novel}. Each episode is formed by a support set S and a query set Q of class C. The support set S has k support images I_s, and the query set Q has only one query image I_q, which is called the k-shot segmentation task. Specifically, $S = \{(I_s^i, M_s^i)\}_{i=1}^k$ contains the support images $I_s^i \in \mathbb{R}^{H\times W\times 3}$ with their corresponding ground truth annotations $M_s^i \in \{0,1\}^{H\times W}$, and $Q = \{(I_q, M_q)\}$ contains the query image I_q and its mask $M_q \in \{0,1\}^{H\times W}$. Note that M_q is invisible to the model and is used to evaluate the prediction on the query image in each episode. During the training, the model is trained on $D_{train} = \{(I_{s/q}, M_{s/q})\}$ in C_{base}. Then, the trained model is generalized to C_{novel} with $D_{test} = \{(I_{s/q}, M_{s/q})\}$ in the testing stage.

3.2 FPNet

In this paper, we propose the frequency prototype network (FPNet) for few-shot semantic segmentation, as shown in Figure 3. FPNet addresses the incomplete support-guided information problem caused by the limitation of GAP and clustering. To this end, FPNet generates multiple prototypes in the DCT domain from global and local perspectives. Based on comprehensive and informative frequency prototypes, FPNet further produces a prior attention mask and selects the most representative frequency prototype to enhance the generalization ability and improve the segmentation performance.

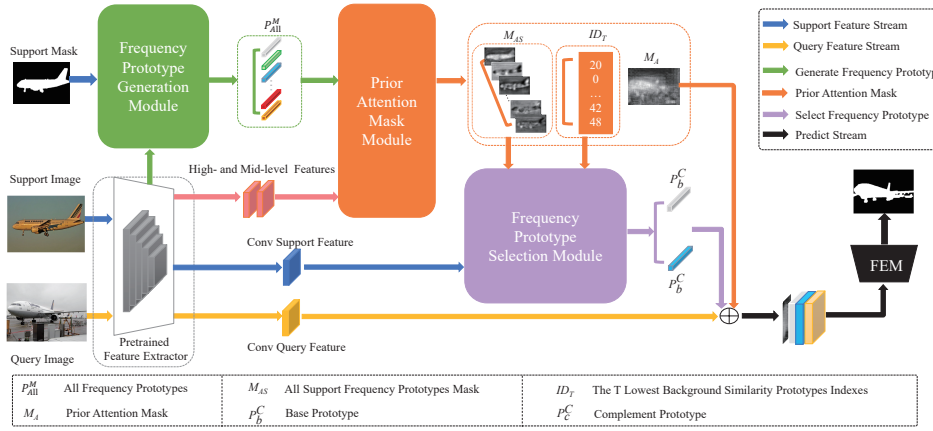


Figure 3. Overview of our proposed FPNet for few-shot segmentation. Our model mainly consists of a pretrained feature extractor, a Frequency Prototype Generation Module (FPGM), a Prior Attention Mask Module (PAMM), a Frequency Prototype Selection Module (FPSM), and a FEM [14]. Specifically, FPGM (green) first generates all frequency prototypes by multi-frequency pooling in the DCT domain from the global and local perspectives. Then, a prior attention mask is obtained by PAMM (orange) to retain the high generalization. Next, FPSM (purple) selects the most representative prototype from all frequency prototypes. Finally, the fusion of the prior attention mask, support guided prototypes, and query feature are fed into the FEM (black) for the final query segmentation mask prediction.

Firstly, the support image I_s and query image I_q are fed into the feature extractor, like the backbone of VGG [5] and ResNet [4], to extract multiple-level features. Similar to the previous few-shot segmentation work [14, 44], the feature extractor is pre-trained on ImageNet [45] and fixed during the training and testing process.

Then, the feature of different levels in the backbone is sent to different modules, respectively. CANet [16] proposes that *block*3 in the feature extractor performs the best when a single block is used. As a result, we feed the support features of *block*3 (F_{s3}^{M}) and the support mask M_s into the Frequency Prototypes Generation Module (FPGM) to obtain all frequency prototypes P_{All}^{M}, described in Section 3.3.

Previous studies prove that the semantic information contained in the high-level feature is more class-specific [14], and the *block*3 generalizes better to unseen classes [16]. Therefore, the *block*4's high-level feature ($F_{s/q}^{H}$) and the *block*3's mid-level feature ($F_{s3/q3}^{M}$) are inputted into the Prior Attention Mask Module (PAMM). Meanwhile, we also feed P_{All}^{M} and M_s into PAMM. After that, we get the outputs of this module: the prior attention mask M_A, all support frequency prototype masks M_{AS}, and the T lowest background similarity prototype indexes ID_T. Section 3.4 shows the details of PAMM.

CANet [16] also proposes that when multiple blocks are used for comparison, the combination of *block*2 and *block*3 achieves the best segmentation result. So we adopt the mid-level features of *block*2 ($F^M_{s2/q2}$) and *block*3 ($F^M_{s3/q3}$) to generate the Conv Feature $F^C_{s/q}$ for bringing more object information. With the input: M_{AS}, ID_T, M_s, and F^C_s, the Frequency Prototype Selection Module (FPSM) generates the base prototype P^C_b and the complement prototype P^C_c. Section 3.5 illustrates this procedure.

Finally, we expand P^C_b and P^C_c to the same shape as the F^C_q and concatenate them with M_A under the channel dimension to get the final concatenated feature F_f. It is taken as the input of the Feature Enrichment Module (FEM) [14] to produce the final predicted query output M'_q.

During training, the segmentation network is optimized end-to-end, driven by the Binary Cross Entropy (BCE) loss between the ground truth and the segmentation mask $L(M_q, M'_q)$.

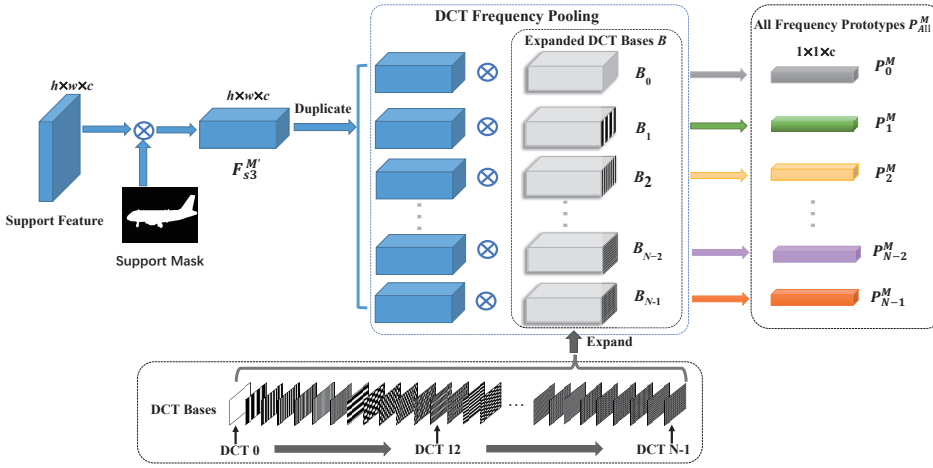


Figure 4. Illustration of our Frequency Prototype Generation Module

3.3 Frequency Prototype Generation Module

To fully mine support object information from the global and local perspective, we introduce multi-frequency pooling and propose FPGM. Figure 4 illustrates the details of our FPGM. Specifically, the support feature $F^M_{s3} \in \mathbb{R}^{h\times w\times c}$ is first multiplied with the ground-truth support mask $M_s \in \{0,1\}^{h\times w}$ to filter out background features. The resulting feature $F^{M'}_{s3} \in \mathbb{R}^{h\times w\times c}$ is generated as:

$$F^{M'}_{s3} = F^M_{s3} \cdot [M_s = 1], \tag{1}$$

where the M_s is resized to the same size as F_{s3}^{M}. The $M_s = 1$ denotes that the pixel at the corresponding spatial position is class C. The $[\cdot]$ is the Iverson bracket that equals 1, if the condition in square brackets is satisfied, otherwise, it equals 0.

Then, the expanded DCT bases B, the variant of the basic DCT formula, is computed as:

$$B = \{B_0, B_1, \ldots, B_{N-1}\}, \tag{2}$$

$$\begin{aligned} B_t = \text{Expand}(\text{DCT}(u, v)), \quad & t \in \{0, 1, \ldots, N-1\}, \\ & u \in \{0, 1, \ldots, K-1\} \times \bar{h}, \quad v \in \{0, 1, \ldots, K-1\} \times \bar{w}, \\ & \bar{h} = \lfloor h/K \rfloor, \quad \bar{w} = \lfloor w/K \rfloor, \quad N = K \times K, \end{aligned} \tag{3}$$

where $B_t \in \mathbb{R}^{h \times w \times c}$ means the t^{th} of the bases $B \in \mathbb{R}^{N \times h \times w \times c}$; h and w represent the height and width of B_t, same as $F_{s3}^{M'}$. The 2D DCT frequency space is divided into $K \times K$ parts (as illustrated in [22], $K = 7$), so h and w are also resized into $\bar{h}$ and $\bar{w}$. Moreover, Expand$(\cdot)$ denotes expanding to the same channel as $F_{s3}^{M'}$. DCT(u, v) is the value of DCT at the corresponding generalized frequency (u, v):

$$\text{DCT}(u, v) = \alpha(u)\alpha(v) \cos\left(\frac{\pi u}{h}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi v}{w}\left(j + \frac{1}{2}\right)\right), \tag{4}$$

$$\begin{aligned} \alpha(x) = & \begin{cases} 1, & x = 0, \\ \sqrt{2}, & x \neq 0, \end{cases} \\ & i \in \{0, 1, \ldots, h-1\}, j \in \{0, 1, \ldots, w-1\}, \end{aligned} \tag{5}$$

where $\alpha(u)$ and $\alpha(v)$ are the coefficients of DCT(u, v), and their value are obtained by Equation (5). i and j represent the spatial position in the B_t.

Finally, $F_{s3}^{M'}$ is duplicated to n copies and multiplied with the corresponding expanded DCT bases B respectively to produce all support frequency prototypes P_{All}^{M}, which is called the multi-frequency pooling (MFP) in the DCT domain:

$$P_{All}^{M} = \{P_0^M, P_1^M, \ldots, P_{N-1}^M\}, \tag{6}$$

$$\begin{aligned} P_t^M = & \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} F_{s3}^{M'}(i, j) \cdot B_t(i, j)}{\sum_{i=1}^{h} \sum_{j=1}^{w} [M_s(i, j) = 1]}, \\ & t \in \{0, 1, \ldots, N-1\}, \end{aligned} \tag{7}$$

$$P_0^M = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} F_{s3}^{M'}(i,j)\cdot B_0(i,j)}{\sum_{i=1}^{h}\sum_{j=1}^{w}[M_s(i,j)=1]}$$
$$= \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} F_{s3}^{M}(i,j)\cdot [M_s(i,j)=1]}{\sum_{i=1}^{h}\sum_{j=1}^{w}[M_s(i,j)=1]}, \quad (8)$$
$$B_0 = \text{Expand}(\text{DCT}(0,0)),$$

where the $P_t^M \in \mathbb{R}^{1\times1\times c}$ is the t^{th} of the prototypes $P_{All}^M \in \mathbb{R}^{N\times1\times1\times c}$. P_0^M is the lowest frequency support prototype for P_{All}^M, equal to the prototype generated by the masked GAP. Equation (8) also indicates that the prototype generated from the masked GAP only preserves the lowest frequency component while discarding the other frequency components. Different from GAP, FPGM utilizes the information of the lowest frequency and the other frequencies to produce more informative support frequency prototypes that guide the following feature matching and learning.

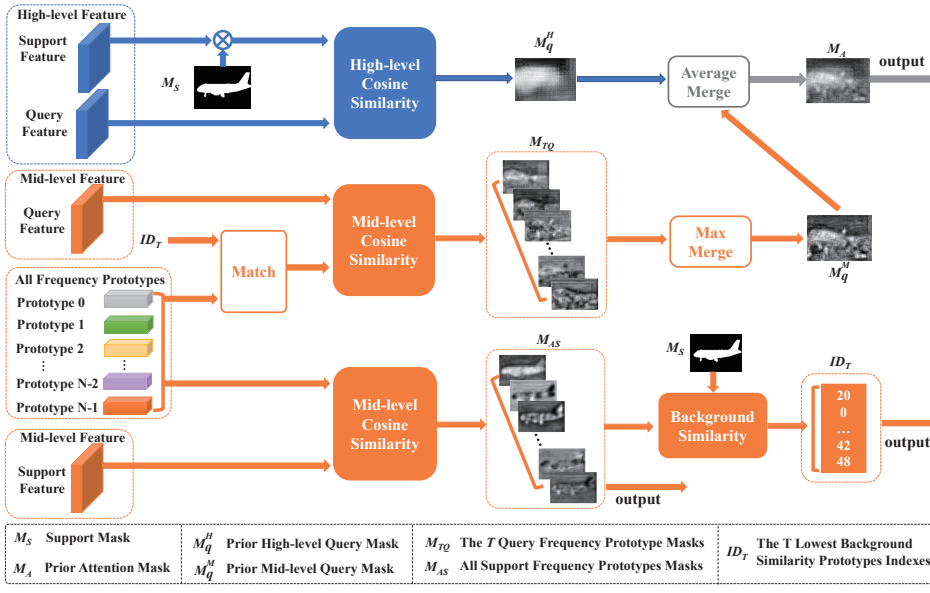


Figure 5. Illustration of Prior Attention Mask Module (PAMM)

3.4 Prior Attention Mask Module

Most previous methods [14, 23, 24] apply the attention map to achieve generalization. However, these attention maps always introduce many learnable parameters and background noises, resulting in generalization reduction and incorrect

predictions. Thus, we propose the PAMM to fuse rich object information from frequency prototypes into the attention map and produce a training-free, more precise, and less noisy prior attention mask M_A. Figure 5 illustrates the process of PAMM.

To produce high-quality M_A, we introduce the High-level Cosine Similarity and the Mid-level Cosine Similarity, which generate the prior masks of different levels by frequency prototypes. Specifically, the prior high-level query mask M_q^H could filter out the background noise, and the prior mid-level query mask M_q^M could capture more object information. Therefore, we first produce M_q^H and M_q^M, then generate M_A.

Prior High-Level Query Mask. Firstly, the binary mask M_s is bilinearly downsampled to the same spatial size as the feature map F_s^H. According to Equation (1), the masked high-level support feature $F_s^{H'}$ is extracted from F_s^H. Given the feature vectors $f_{q'}^H \in F_q^H$ and $f_{s'}^H \in F_s^{H'}$, we compute the pixel-wise cosine similarity map by the High-level Cosine Similarity:

$$\cos\left(f_{q'}^H, f_{s'}^H\right) = \frac{\left(f_{q'}^H\right)^T f_{s'}^H}{\|f_{q'}^H\| \|f_{s'}^H\|}, \quad q', s' \in \{0, 1, \ldots, hw-1\}. \tag{9}$$

For each $f_{q'}^H$, we take the maximum similarity among all support nodes as the value of the prior high-level query probability map $map_{q'}^H \in \mathbb{R}$:

$$map_{q'}^H = \max_{s' \in \{0,1,\ldots,hw-1\}} \left(\cos\left(f_{q'}^H, f_{s'}^H\right)\right), \tag{10}$$

$$M_q^H = \left\{map_0^H, map_1^H, \ldots, map_{hw-1}^H\right\} \in \mathbb{R}^{hw \times 1}. \tag{11}$$

Then, all the values in M_q^H are normalized to $[0, 1]$ by min-max normalization:

$$M_q^H = \frac{M_q^H - \min\left(M_q^H\right)}{\max\left(M_q^H\right) - \min\left(M_q^H\right) + \epsilon}, \tag{12}$$

where ϵ is set to $1e-7$. Finally, M_q^H is resized into $h \times w \times 1$.

Prior Mid-Level Query Mask. To capture complete and rich object information and fuse them into M_q^M, we analyze and process all frequency prototypes P_{All}^M. Based on the object guidance information in prototypes, we produce all support frequency prototype masks M_{AS} for the support image. Because the object guidance information in different prototypes can predict different results, we judge the quality of prototypes by the segmentation results (M_{AS}). To this end, we apply background similarity as the evaluation criterion and select the better prototypes from P_{All}^M to guide the generation of M_q^M. The specific process is described below.

Given $f_{s'}^M \in F_{s3}^M$ and $P_t^M \in P_{All}^M$, we apply Mid-level Cosine Similarity to produce the single similarity map $map_{ASt} \in M_{AS}$:

$$map_{ASt} = \cos\left(f_{s'}^M, P_t^M\right) = \frac{\left(f_{s'}^M\right)^T P_t^M}{\|f_{s'}^M\| \|P_t^M\|} \in \mathbb{R}^{hw \times 1}, \tag{13}$$

$$s' \in \{0, 1, \ldots, hw-1\}, \quad t \in \{0, 1, \ldots, N-1\}. \tag{14}$$

$$M_{AS} = \{map_{AS0}, map_{AS1}, \ldots, map_{ASN-1}\}. \tag{15}$$

Then, we normalize map_{ASt} according to Equation (12) and reshape them to get final $M_{AS} \in N \times h \times w \times 1$.

PAMM aims to fuse more object information and less background noise into the prior attention mask. However, the features extracted by the foreground similarity always contain many background noises similar to the object. Therefore, we apply the background similarity to select M_{AS}. Firstly, we filter out foreground by pixel-wised multiplication with the ground-truth support mask M_s and get M_{AS}^B:

$$M_{AS}^B = M_{AS} \cdot [M_s = 0], \tag{16}$$

where M_s is reshaped and expanded into $N \times h \times w \times 1$. Then, we compute background similarity S_B:

$$S_B = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} M_{AS}^B(i,j)}{\sum_{i=1}^{h} \sum_{j=1}^{w} [M_s(i,j) = 0]}. \tag{17}$$

After obtaining S_B, we take out the indexes of the T lowest background similarity prototypes, denoted ID_T. The specific function is as follows:

$$ID_T = \mathcal{F}_{Index}(\mathcal{F}_{Low}(S_B, T)), \tag{18}$$

where $\mathcal{F}_{Low}(\cdot, T)$ represents selecting the T lowest background similarities, and $\mathcal{F}_{Index}(\cdot)$ represents taking out the corresponding background similarity prototype indexes.

Based on ID_T, we can get the prior mid-level query mask M_q^M. Specifically, we first feed the matched prototypes and mid-level query feature into the Mid-level Cosine Similarity. Then, we obtain the T query frequency prototype masks $M_{TQ} \in \mathbb{R}^{T \times h \times w \times 1}$, which contain rich object information from the frequency prototype. Finally, $M_q^M \in \mathbb{R}^{h \times w \times 1}$ is generated by extracting the maximum pixel value on each position of the M_{TQ}:

$$M_q^M = \max(M_{TQ}), \tag{19}$$

where $Max(\cdot)$ is the pixel-level maximum value operation.

Prior Attention Mask. M_q^H is more coarse and whole so that it can identify the approximate position of the object. Moreover, we fuse rich frequency prototype information into M_q^M. Therefore, M_q^M contains fine and detailed object information. In the end, we produce M_A by averaging the M_q^H and M_q^M to boost object information and suppress background noises:

$$M_A = \frac{M_q^H + M_q^M}{2}. \tag{20}$$

In our model, we also integrate the result of the base learner in BAM [46] to produce a more precise prior attention mask.

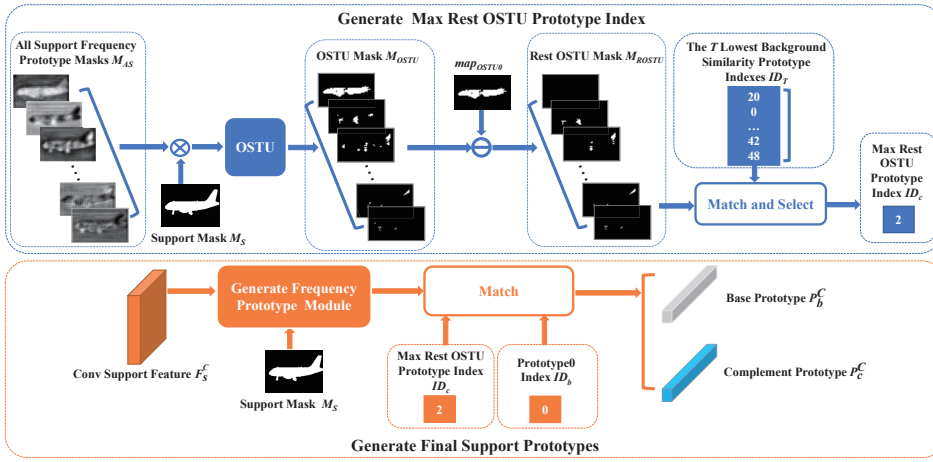


Figure 6. Illustration of Frequency Prototype Selection Module (FPSM). The upper part is the process of generating the max rest OSTU prototype index. The lower part is the process of generating final support prototypes.

3.5 Frequency Prototype Selection Module

Applying all the prototypes for object prediction consumes enormous computational power and storage space. So we propose FPSM to select more representative prototypes (the base prototype P_b^C and the complemnet prototype P_c^C) based on ID_T. P_b^C from B_0 captures the global information of the target, and P_c^C from $B_t (t > 0)$ guides the best local segmentation. Figures 2 and 9 show more details about the frequency prototype visualization. By concatenating and expanding the above prototypes, we introduce rich and complete support information to guide the query target segmentation while reducing model complexity and computational burden.

Firstly, we use M_s to filter out the background masks of M_{AS} (Equation 1) and get $M^{'}_{AS}$. Then, the OSTU is applied to segment $M^{'}_{AS}$:

$$M^{'}_{AS} = \{map^{'}_{AS0}, map^{'}_{AS1}, \ldots, map^{'}_{ASN-1}\}, \tag{21}$$

$$map_{OSTUt}(i,j) = [map^{'}_{ASt}(i,j) \geq \tau], \tag{22}$$
$$t \in \{0,1,\ldots,N-1\}, \quad i \in \{0,1,\ldots,h-1\},$$
$$j \in \{0,1,\ldots,w-1\},$$

$$M_{OSTU} = \{map_{OSTU0}, map_{OSTU1}, \ldots, map_{OSTUN-1}\}, \tag{23}$$

where the τ is the segmentation threshold from $map^{'}_{AS0}$, and the M_{OSTU} is the segmentation result of $M^{'}_{AS}$. By pixel-wised subtracting map_{OSTU0} from M_{OSTU}, all the rest threshold segmentation masks M_{ROSTU} is generated:

$$M_{ROSTU} = M_{OSTU} - map_{OSTU0}, \tag{24}$$

where the map_{OSTU0} is expanded into $N \times h \times w \times 1$ first. By matching the indexes of M_{ROSTU} with ID_T and selecting the max rest masks from them, we obtain the ID_c of P^C_c. In this way, we obtain two complemental prototypes to capture global and local object information without learnable parameters. Finally, given F^C_s and M_s, we employ the FPGM again to produce the final prototype P^C_b and P^C_c with the indexes of ID_b and ID_c ($ID_b = 0$).

3.6 k-Shot Setting

In the k-shot setting, k support images are given to extract the object information. For extending 1-shot segmentation to k-shot, we propose the appropriate way for different modules.

For PAMM, we apply two schemes because of the diversity of the support images: Masks Average Fusion (MAF) and Masks Max Fusion (MMF). MAF produces the final M_A by averaging operation. Since the same class objects in different images usually have different features, we apply MMF to capture more object information. Firstly, MMF takes out the pixel-wised maximum value of all M^M_q as the final M^M_q and the average of all M^H_q as the final M^H_q. Then, the average of M^M_q and M^H_q are taken as the final M_A. According to the ablation experiment results on the 5-shot setting in Section 4.3, we use MMF in this paper.

As for FPSM, we take the average of all the base and complement prototypes as the final P^C_b and P^C_c.

4 EXPERIMENTS

4.1 Implementation Details

Datasets. Our model has evaluated on PASCAL-5^i [8] and COCO-20^i [47] datasets, which are usually used for previous few-shot semantic segmentation [48, 19]. The PASCAL-5^i consists of PASCAL VOC 2012 [49] and augmented SDS [50] datasets. Following OSLSM [8], we evenly split the 20 classes of the PASCAL-5^i into 4 folds with 5 classes per fold. The cross-validation experiment evaluates the proposed model: three for training and one for testing. For COCO-20^i, following FWB [47], we evenly divide the 80 classes in MSCOCO [51] into 4 folds with the same cross-validation strategy, and each fold contains 20 classes.

Experimental Setting. We construct our framework on PyTorch. For a fair comparison, we apply different backbones, including VGG-16 [5] and ResNet-50 [4], where VGG is the original version, and the ResNet is the dilated version similar to previous works [11, 14, 16]. The SGD is adopted as the optimizer for training. The momentum and weight decay are set to 0.9 and 0.0001, respectively. During training, we also use the data augmentation strategies such as random scale, Gaussian filtering, horizontal flip, and random rotation. Then, the processed images are cropped into 473×473 (Pascal) or 641×641 (COCO) as training samples. We set the initial learning rate to 0.005 with batch size 4 on PASCAL-5^i and 0.005 with batch size 8 on COCO-20^i. The final testing result is computed by averaging the results of 5 trials with different random seeds to ensure its reliability. Our experiments run on Nvidia Tesla V100 GPUs.

Evaluation Metrics. Following the previous work in [14, 16, 47], we adopt the mean intersection-over-union (mIoU) as the performance evaluation metric. 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 numbers of true positives, false positives, and false negatives, respectively. The mIoU is defined as the mean IoUs of all image classes.

4.2 Comparison with State-of-the-Art

PASCAL-5^i. As shown in Table 1, we build our model on two backbones, including VGG-16 and ResNet-50, and compare our model with the state-of-the-art approaches on PASCAL-5^i. It can be seen that the FPNet achieves state-of-the-art performance on both 1-shot and 5-shot tasks with the two backbones. With VGG-16 as the backbone, our model outperforms previous state-of-the-art results with a margin of 1.4 % for 1-shot. With ResNet-50 as a backbone, FPNet is comparable to the state-of-the-art approaches, and the mIoU increases 0.3 % on the 5-shot task.

Method	Backbone	1-shot					5-shot				
		Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
OSLSM [8] (BMVC'17)	VGG-16	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
PANet [15] (ICCV'19)		42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
SG-One [17] (TCYB'20)		40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
PFENet [14] (TPAM'20)		56.9	68.2	54.4	52.4	58.0	59.0	69.1	54.8	52.9	59.0
HSNet [52] (ICCV'21)		59.6	65.7	59.6	54.0	59.7	64.9	69.0	64.1	58.6	64.1
DPCN [22] (CVPR'22)		58.9	69.1	63.2	55.7	61.7	63.4	70.7	68.1	59.0	65.3
BAM [46] (CVPR'22)		63.2	70.8	66.1	**57.5**	64.4	67.4	73.1	**70.6**	**64.0**	**68.8**
FPNet (ours)		**67.0**	**72.6**	**66.8**	56.8	**65.8**	**69.2**	**73.4**	70.1	62.6	**68.8**
CANet [16] (CVPR'19)	ResNet-50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet [44] (ICCV'19)		56.0	66.9	50.6	56.0	57.7	57.7	68.7	52.9	54.6	58.5
RPMM [33] (ECCV'20)		55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PFENet [24] (TPAM'20)		61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
ASGNet [18] (CVPR'21)		58.8	67.9	56.8	53.7	59.3	63.7	70.6	64.2	57.4	63.9
MMNet [53] (CVPR'21)		62.7	70.2	57.3	57.0	61.8	62.2	71.5	57.5	62.4	63.4
HSNet [52] (ICCV'21)		64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	67.1	69.5
DPNet [54] (AAAI'22)		60.7	69.5	62.8	58.0	62.7	64.7	70.8	69.0	60.1	66.2
DPCN [22] (CVPR'22)		65.7	71.6	69.1	60.6	66.7	70.0	73.2	70.9	65.5	69.9
BAM [46] (CVPR'22)		69.0	73.6	67.6	**61.1**	67.8	70.6	75.1	70.8	**67.2**	70.9
FPNet (ours)		**69.4**	**73.8**	**68.8**	60.0	**68.0**	**73.0**	**75.8**	**71.2**	64.5	**71.2**

Table 1. The mIoU performance of 1-shot and 5-shot segmentation on PASCAL-5^i. The best performances are highlighted in bold.

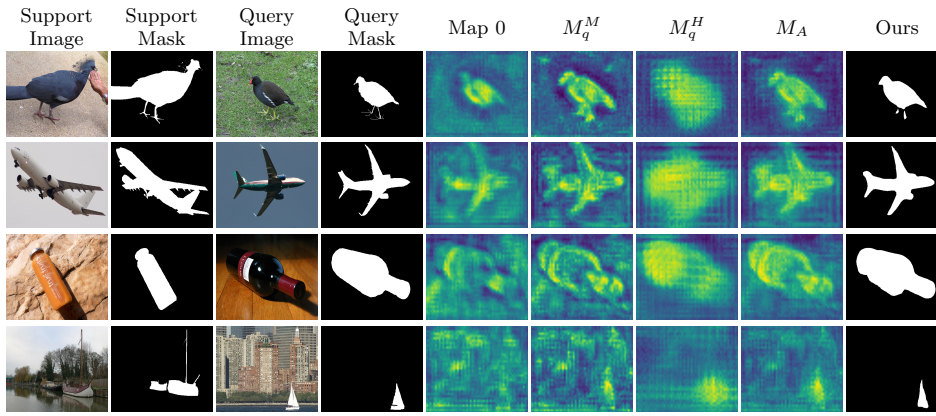


Figure 7. Visual results of our model on fold-0 of the PASCAL-5^i dataset. Each column from left to right represents the support image, support mask, query image, query mask, activation map generated by GAP, prior mid-level query mask, prior high-level query mask, prior attention mask, and prediction of our proposed FPNet, respectively. Best viewed in color and zoom-in.

COCO-20^i. In Table 2, we compare the segmentation results of our model and other state-of-the-art models on COCO-20^i. As can be seen, FPNet achieves state-of-the-art results in both 1-shot and 5-shot setting and it outperforms others with mIoU gains of 0.7 % and 0.8 %, respectively. These results prove the capability of our method to handle more challenging cases.

Segmentation Examples. To better understand our proposed method, we show some prediction results of the test stage and activation maps produced by frequency prototypes. As shown in Figures 7 and 8, M_q^M can capture global and local object information compared with Map 0, and M_q^H can suppress the background noise incurred by M_q^M. By combining the M_q^M and M_q^H, our model generates high-quality M_A to predict complete and accurate results on PASCAL-5^i and COCO-20^i datasets. In the first row of Figure 7, M_q^M mines the body and head information of the bird compared to Map 0, and M_q^H can filter the background noise around the bird in M_q^M. Moreover, we visualize some examples of the frequency prototype activation maps in Figure 9, where we randomly select 5 from all 49 maps. Different class maps are both obtained from the same 2D DCT bases. As we can see, different frequency prototypes represent different object features. Those prototypes not only capture the object inside information from global and local perspective, but also extract object outside information. In the third row of Figure 9, Map 0, Map 1 and Map 2 can capture the main body, head and tail of the bird respectively and Map 3, and Map 4 can capture the shape of the bird and the background respectively. These visual results verify the effectiveness of frequency prototype learning.

Method	Backbone	1-shot					5-shot				
		Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
FWB [47] (ICCV '19)	ResNet-101	19.9	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
RPMM [33] (ECCV '20)	ResNet-50	29.5	36.8	28.9	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PFENet [14] (TPAM '20)	ResNet-101	36.8	41.8	38.7	36.7	38.5	40.4	46.8	43.2	40.5	42.7
SCL [21] (CVPR '21)	ResNet-101	36.4	38.6	37.5	35.4	37.0	38.9	40.5	41.5	38.7	39.9
HSNet [52] (ICCV '21)	ResNet-101	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
DCP [55] (IJCAI '22)	ResNet-50	40.9	43.8	42.6	38.3	41.4	45.8	49.7	43.7	46.6	46.5
DPCN [22] (CVPR '22)	ResNet-50	42.0	47.0	43.2	39.7	43.0	46.0	54.9	50.8	47.4	49.8
BAM [46] (CVPR '22)	ResNet-50	**43.4**	50.6	**47.5**	43.4	46.2	**49.3**	54.2	**51.6**	49.6	51.2
FPNet(ours)	ResNet-50	40.8	**53.9**	47.1	**45.7**	**46.9**	46.6	**59.8**	50.9	**50.6**	**52.0**

Table 2. The mIoU performance of 1-shot and 5-shot segmentation on COCO-20^i. The best performances are highlighted in bold.

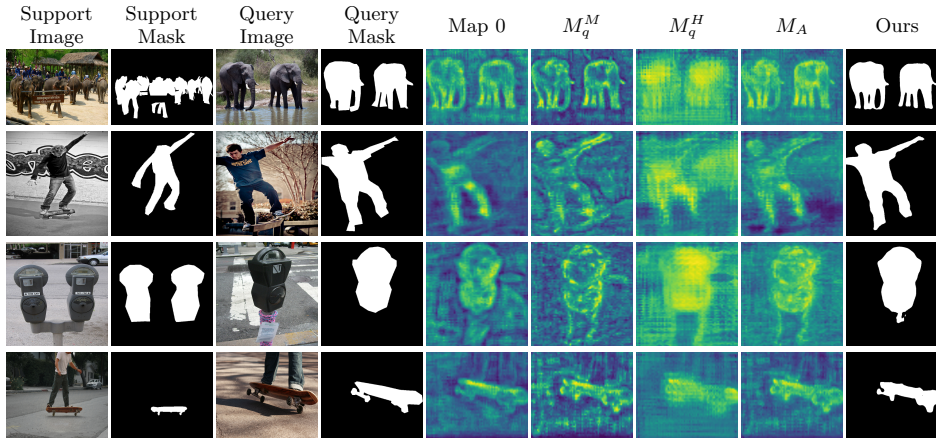


Figure 8. Visual results of our model on fold-0 of the COCO-20^i dataset. Each column from left to right represents the support image, support mask, query image, query mask, activation map generated by GAP, prior mid-level query mask, prior high-level query mask, prior attention mask, and prediction of our proposed FPNet, respectively. Best viewed in color and zoom-in.

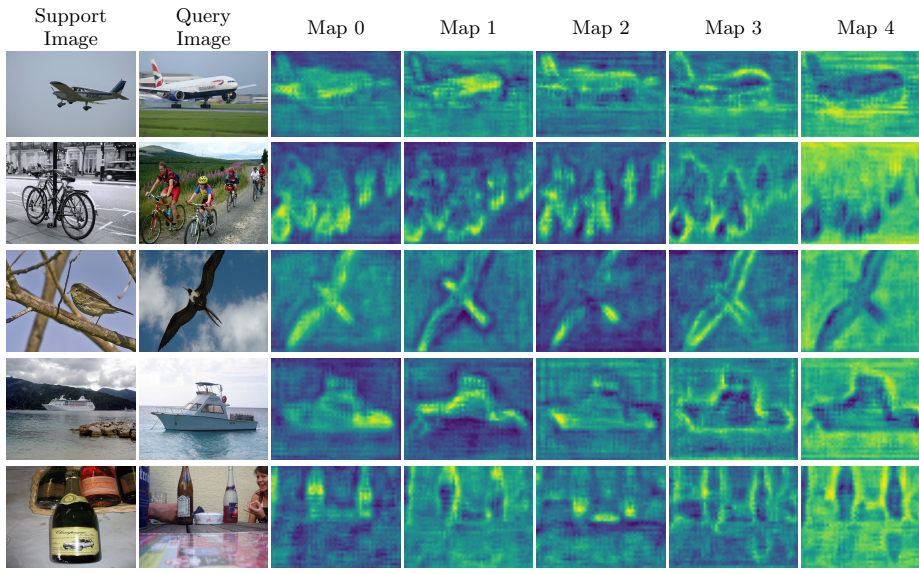


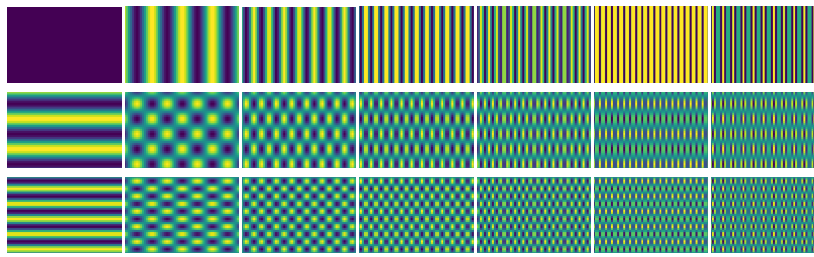
Figure 9. Visual results of the frequency prototype activation maps. We randomly select 5 from all 49 query activation maps, and all query activation maps are obtained from the support frequency prototypes. Map 0 represents the map generated by DCT^0, and Map 1 to Map 4 are generated by DCT^t ($t \in \{1, 2, \ldots, 48\}$). We can see that Map 0 highlights the main part of the object, and Map 1 and Map 2 highlight the local part of the object. Moreover, Map 3 and Map 4 highlight the object's edge and the background, respectively. Best viewed in color and zoom-in.

4.3 Ablation Studies

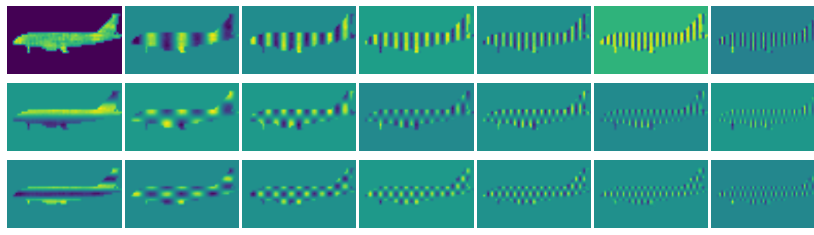
We conduct a set of ablation experiments with ResNet-50 on PASCAL-5^i to verify the effectiveness of the proposed modules.

Visualization of the Process of FPGM. The process of FPGM is visualized in Figure 10, showing the way of mining object information and the advantage of multi-frequency pooling (MFP). Different from the previous GAP and clustering, our proposed MFP captures both the global and local object information. By multiplying foreground support features with DCT 0 (the lowest frequency), the global object information is captured and compressed into the frequency prototype P_0, as shown in the left upper corner of Figure 10 b). So P_0 could activate the main part of the plane, like the fuselage in the left upper corner of Figure 10 c).

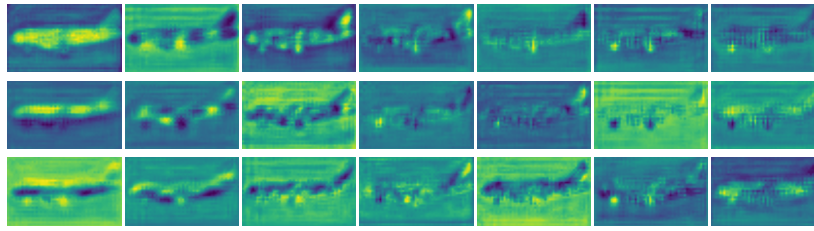
The higher frequency prototypes represent the local object information. As shown in Figure 10 b), these local information could be regarded as the horizontal and vertical banding or gridding features. In the first row of Figure 10 b), vertical banding object features of different sizes are activated by



a) DCT bases



b) Activation maps generated by multiplying foreground support features with DCT bases



c) Activation maps generated by frequency prototypes

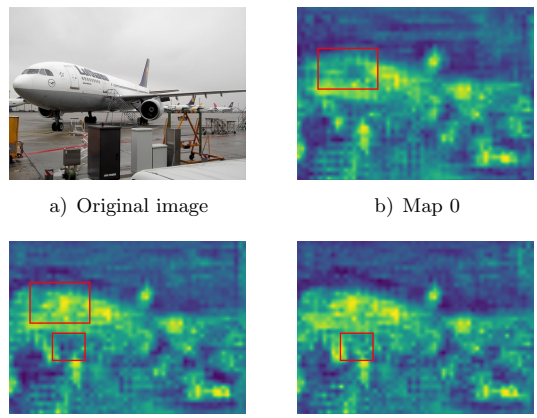
Figure 10. Visualization of the process of FPGM. Please note that the lower and right images contain higher-frequency domain information. Best viewed in color and zoom-in.

the corresponding DCT bases in Figure 10 a). Then these features are compressed into the corresponding frequency prototypes. Hence, these prototypes represent some vertical object parts, like the empennage and wheel in the first row of Figure 10 c). Similarly, horizontal banding object features are compressed into the corresponding frequency prototypes as shown in the first column of Figure 10 b). These prototypes can focus on the horizontal object part, like the fuselage in the first column of Figure 10 c). Moreover, the gridding object features of different sizes are mined by multiplying foreground support features with corresponding DCT bases, resulting in the corresponding frequency prototypes, as can be seen in the other rows or columns of Figure 10 b). In the third and seventh columns of Figure 10 c), plane parts of different sizes are activated by these prototypes, like the empennage, nose, and wheel.

Hence, our proposed MFP addresses the limitation of GAP and clustering and generates comprehensive and informative frequency prototypes.

Number of the Lowest Background Similarity Prototypes	1-shot				
	Fold0	Fold1	Fold2	Fold3	Mean
5	67.6	73.1	67.9	58.6	66.8
10	**69.4**	**73.9**	**68.8**	**60.0**	**68.0**
15	68.8	73.4	68.0	58.6	67.2

Table 3. Ablation studies for the lowest background similarity prototype numbers



a) Original image b) Map 0 c) Map 1 d) Map 2

Figure 11. Prior mid-level query attention masks generated from different numbers of the lowest background similarity prototypes. Map 0, Map 1, and Map 2 are the prior mid-level query attention masks generated from the 5, 10, and 15 lowest background similarity prototypes, respectively. Best viewed in color and zoom-in.

A Number of the Lowest Background Similarity Prototypes. We conduct ablation experiments to analyze the effect of the lowest background similarity prototype numbers. As Table 3 shows, we generate the best performances when 10 lowest background similarity prototypes are used. The result improves when the number increase from 5 to 10. However, it decreases after that, denoting that the excessive number of prototypes brings more background noises rather than more foreground information.

We visualize the results of the lowest background similarity prototype numbers in Figure 10. It shows that the 10 lowest background similarity prototypes generate the best prior mask. Specifically, the prior mask, composed of 10 prototypes, pays more attention to the head of the airplane. However, the prior mask composed of the 15 prototypes barely boosts the attention to the airplane

and brings more background noises. Therefore, PAMM can yield a more robust and accurate prior mid-level query attention mask based on the 10 lowest background similarity prototypes, it whence helps the model focus on the object with fewer background noises.

In the end, we use the 10 lowest background similarity prototypes in our model.

PAMM	FPSM	Fold0	Fold1	Fold2	Fold3	Mean
		62.7	71.6	63.0	54.9	63.1
✓		68.8	73.4	67.9	58.7	67.2
	✓	64.4	72.0	64.2	56.0	64.2
✓	✓	**69.4**	**73.8**	**68.8**	**60.0**	**68.0**

Table 4. Ablation studies on the key modules in our FPNet. PAMM, FPSM denote prior attention mask module and frequency prototype selection module, respectively.

Effect of PAMM and FPSM. To demonstrate the effectiveness of the proposed PAMM and FPSM, we conduct ablation experiments in Table 4. The first line is the baseline result, which uses the single prototype produced by the masked GAP. We first introduce FPGM to replace the masked GAP and generate multiple frequency prototypes. Then, we use frequency prototypes in PAMM and FPSM. Finally, we orderly evaluate PAMM and FPSM to demonstrate their effectiveness. The goal of PAMM is to enhance the model generalization ability and identify on query target more precisely by the prior attention mask. Moveover, it improves the segmentation result from 63.1 % to 67.2 % in Table 4. FPSM aims to extract global and local object information, squeeze them into the base and complement prototypes. As shown in Table 4, FPSM improves performance by 1.1 %. Combining all modules, we can obtain another 0.8 % performance gain and reach 68.0 %. This ablation experiment also demonstrates that FPGM extracts more comprehensive and quality object prototypes from global and local perspectives.

	Fold0	Fold1	Fold2	Fold3	Mean
Baseline	62.7	71.6	63.0	54.9	63.1
Baseline + m_q^h	65.9	72.5	66.5	**59.5**	66.1
Baseline + m_q^m	64.0	72.2	63.8	56.9	64.2
Baseline + $m_q^h + m_q^m$	67.1	72.9	67.2	59.4	66.7
Baseline + $m_q^h + m_q^m + m_b$	**68.8**	**73.4**	**67.9**	58.7	**67.2**

Table 5. Ablation studies on components of PAMM. m_q^h denotes the prior high-level query mask and m_q^m denotes the prior mid-level query mask. m_b denotes the base learner result of BAM [46].

Components in PAMM. The output of PAMM (prior attention mask M_A) is mainly composed of the prior high-level query mask M_q^H and the prior mid-level query mask M_q^M. So we evaluate the effectiveness of each component

in Table 5. The baseline model produces the single prototype by the masked GAP for pixel-wise dense semantic prediction in the decoder. When we only use M_q^H or M_q^M in PAMM, it achieves 3.0 % and 1.1 % mIoU improvement over the baseline result, respectively. It proves that M_q^H and M_q^M can capture more object information and retain high generalization. By introducing the combination of M_q^H and M_q^M to the baseline model, the performance improves by 3.6 %. This result shows that these two masks are complementary; M_q^H can suppress the noise in M_q^M, and M_q^M can boost the foreground object feature in M_q^H. Moreover, when the M_b is integrated into the M_A, another promotion of 0.5 % is obtained.

Strategy	5-shot				
	Fold0	Fold1	Fold2	Fold3	Mean
Masks Average Fusion	72.4	75.2	70.8	63.6	70.5
Masks Max Fusion	**72.7**	**76.2**	**71.5**	**64.2**	**71.2**

Table 6. Ablation studies on 5-shot fusion schemes

5-Shot Fusion Schemes. As introduced in Section 3.6, we apply two schemes for producing the prior attention mask in the 5-shot setting: Masks Average Fusion (MAF) and Masks Max Fusion (MMF). Table 6 shows the ablation study on 5-shot fusion schemes. MMF achieves a better result than MAF and improves mIoU by 0.7 %. Therefore, the MMF scheme can generate more precise prior attention masks for query mask prediction.

	1-shot					5-shot				
	Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
PFENet	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
PFENet + FPSM	63.2	**69.7**	56.2	55.8	61.2	65.0	70.8	56.5	57.0	62.3
PFENet + PAMM	63.3	**69.7**	56.6	**56.8**	61.6	64.5	70.0	56.8	**59.8**	62.8
PFENet + PAMM + FPSM	**64.7**	69.5	**57.2**	55.8	**61.8**	**66.7**	**73.1**	**57.2**	58.3	**63.8**

Table 7. Generalization ability of the proposed PAMM and FPSM for 1-shot and 5-shot segmentation on PASCAL-5^i

Generalization of PAMM and FPSM. FPGM can be considered a method to generate prototypes in the frequency domain compared with GAP and clustering. Moreover, it is plug-and-play. Because the inputs of PAMM and FPSM mainly come from FPGM, the PAMM and FPSM also can be plug-and-play for current prototype-based methods and further improve their performance.

	Setting	Backbone	1-shot				
			Fold0	Fold1	Fold2	Fold3	Mean
PFENet	1-shot	ResNet-101	36.8	41.8	38.7	36.7	38.5
PFENet + PAMM + FPSM		ResNet-50	**38.2**	**42.0**	**39.1**	**38.4**	**39.4**
PFENet	5-shot	ResNet-101	40.4	46.8	43.2	40.5	42.7
PFENet + PAMM + FPSM		ResNet-50	**41.8**	**47.5**	**44.6**	**42.9**	**44.2**

Table 8. Generalization ability of the proposed PAMM and FPSM for 1-shot and 5-shot segmentation on COCO-20^i

To verify this, we apply PAMM and FPSM to PFENet [14]. Note that the result of the base learner in BAM [46] is not integrated into PAMM, and we follow the same experiment settings as the PFENet. Table 7 shows that PAMM and FPSM bring 1.5 % and 2.6 % improvements in 1-shot and 5-shot settings on PASCAL-5^i, respectively. In 1-shot and 5-shot settings on COCO-20^i, the PAMM and FPSM also achieve 0.9 % and 1.5 % mIoU improvement, respectively, as Table 8 shown.

5 CONCLUSION

This paper proposes a novel frequency prototype network (FPNet) for the few-shot semantic segmentation task. Instead of the global prototype generated by GAP or local prototypes generated by clustering, we try to extract both global and local prototypes in the support image by multi-frequency pooling from various DCT frequency components. Therefore, frequency prototypes can provide more quality and comprehensive object guidance information. Based on guidance information, we produce the prior attention mask to boost object features in the query image, suppress background noise, and enhance the model's generalization ability. Extensive experiments verify the superiority of our proposed FPNet. In the future, we will focus on mining more valuable information from an image in the frequency domain.

Acknowledgement

This work was supported by the National Nature Science Foundation of China under Grant No. 61872143.

REFERENCES

[1] Long, J.—Shelhamer, E.—Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[2] Zhao, H.—Shi, J.—Qi, X.—Wang, X.—Jia, J.: Pyramid Scene Parsing Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[3] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F. (Eds.): Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, Lecture Notes in Computer Science, Vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[4] He, K.—Zhang, X.—Ren, S.—Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[5] Simonyan, K.—Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio, Y., LeCun, Y. (Eds.): 3rd International Conference on Learning Representations (ICLR 2015). 2015, doi: 10.48550/arXiv.1409.1556.

[6] Sun, S.—Zhi, S.—Liao, Q.—Heikkilä, J.—Liu, L.: Unbiased Scene Graph Generation via Two-Stage Causal Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 10, pp. 12562–12580, doi: 10.1109/TPAMI.2023.3285009.

[7] Sun, S.—Zhi, S.—Heikkilä, J.—Liu, L.: Evidential Uncertainty and Diversity Guided Active Learning for Scene Graph Generation. The Eleventh International Conference on Learning Representations (ICLR 2023), 2023, https://openreview.net/forum?id=xI1ZTtVOtlz.

[8] Shaban, A.—Bansal, S.—Liu, Z.—Essa, I.—Boots, B.: One-Shot Learning for Semantic Segmentation. British Machine Vision Conference 2017 (BMVC 2017), BMVA Press, 2017, https://bmva-archive.org.uk/bmvc/2017/papers/paper167/paper167.pdf.

[9] Rakelly, K.—Shelhamer, E.—Darrell, T.—Efros, A. A.—Levine, S.: Conditional Networks for Few-Shot Semantic Segmentation. 6th International Conference on Learning Representations (ICLR 2018), 2018, https://openreview.net/forum?id=SkMjFKJwG.

[10] Siam, M.—Oreshkin, B. N.—Jägersand, M.: AMP: Adaptive Masked Proxies for Few-Shot Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), 2019, pp. 5248–5257, doi: 10.1109/ICCV.2019.00535.

[11] Hu, T.—Yang, P.—Zhang, C.—Yu, G.—Mu, Y.—Snoek, C. G. M.: Attention-Based Multi-Context Guiding for Few-Shot Semantic Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, No. 1, pp. 8441–8448, doi: 10.1609/aaai.v33i01.33018441.

[12] Gairola, S.—Hemani, M.—Chopra, A.—Krishnamurthy, B.: SimPropNet: Improved Similarity Propagation for Few-Shot Image Segmentation. In: Bessiere, C.

(Ed.): Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020). 2020, pp. 573–579, doi: 10.24963/ijcai.2020/80.

[13] Dong, N.—Xing, E. P.: Few-Shot Semantic Segmentation with Prototype Learning. British Machine Vision Conference 2018 (BMVC 2018), BMVA Press, 2018, https://bmva-archive.org.uk/bmvc/2018/contents/papers/0255.pdf.

[14] Tian, Z.—Zhao, H.—Shu, M.—Yang, Z.—Li, R.—Jia, J.: Prior Guided Feature Enrichment Network for Few-Shot Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, 2022, No. 2, pp. 1050–1065, doi: 10.1109/TPAMI.2020.3013717.

[15] Wang, K.—Liew, J. H.—Zou, Y.—Zhou, D.—Feng, J.: PANet: Few-Shot Image Semantic Segmentation with Prototype Alignment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9196–9205, doi: 10.1109/ICCV.2019.00929.

[16] Zhang, C.—Lin, G.—Liu, F.—Yao, R.—Shen, C.: CANet: Class-Agnostic Segmentation Networks with Iterative Refinement and Attentive Few-Shot Learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5212–5221, doi: 10.1109/CVPR.2019.00536.

[17] Zhang, X.—Wei, Y.—Yang, Y.—Huang, T. S.: SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation. IEEE Transactions on Cybernetics, Vol. 50, 2020, No. 9, pp. 3855–3865, doi: 10.1109/TCYB.2020.2992433.

[18] Li, G.—Jampani, V.—Sevilla-Lara, L.—Sun, D.—Kim, J.—Kim, J.: Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8330–8339, doi: 10.1109/CVPR46437.2021.00823.

[19] Liu, Y.—Liu, N.—Cao, Q.—Yao, X.—Han, J.—Shao, L.: Learning Non-Target Knowledge for Few-Shot Semantic Segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11563–11572, doi: 10.1109/CVPR52688.2022.01128.

[20] Liu, W.—Zhang, C.—Lin, G.—Liu, F.: CRNet: Cross-Reference Networks for Few-Shot Segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4164–4172, doi: 10.1109/CVPR42600.2020.00422.

[21] Zhang, B.—Xiao, J.—Qin, T.: Self-Guided and Cross-Guided Learning for Few-Shot Segmentation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8308–8317, doi: 10.1109/CVPR46437.2021.00821.

[22] Liu, J.—Bao, Y.—Xie, G. S.—Xiong, H.—Sonke, J. J.—Gavves, E.: Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11543–11552, doi: 10.1109/CVPR52688.2022.01126.

[23] Ying, X.—Li, X.—Chuah, M. C.: Weakly-Supervised Object Representation Learning for Few-Shot Semantic Segmentation. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1496–1505, doi: 10.1109/WACV48630.2021.00154.

[24] Zhang, X.—Wei, Y.—Li, Z.—Yan, C.—Yang, Y.: Rich Embedding Fea-

tures for One-Shot Semantic Segmentation. IEEE Transactions on Neural Networks and Learning Systems, Vol. 33, 2022, No. 11, pp. 6484–6493, doi: 10.1109/TNNLS.2021.3081693.

[25] Zhang, Z.—Jiang, S.—Pan, X.: RGN-Net: A Global Contextual and Multiscale Information Association Network for Medical Image Segmentation. Computing and Informatics, Vol. 41, 2022, No. 5, pp. 1383–1400, doi: 10.31577/cai_2022_5_1383.

[26] Zhang, X.—Li, Q.—Quan, Z.—Yang, W.: Pyramid Geometric Consistency Learning for Semantic Segmentation. Pattern Recognition, Vol. 133, 2023, Art. No. 109020, doi: 10.1016/j.patcog.2022.109020.

[27] Chen, L. C.—Papandreou, G.—Kokkinos, I.—Murphy, K.—Yuille, A. L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, 2018, No. 4, pp. 834–848, doi: 10.1109/TPAMI.2017.2699184.

[28] Chen, L. C.—Papandreou, G.—Schroff, F.—Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR, 2017, doi: 10.48550/arXiv.1706.05587.

[29] Huang, Z.—Wang, C.—Wang, X.—Liu, W.—Wang, J.: Semantic Image Segmentation by Scale-Adaptive Networks. IEEE Transactions on Image Processing, Vol. 29, 2020, pp. 2066–2077, doi: 10.1109/TIP.2019.2941644.

[30] Zhang, Z.—Jiang, S.—Pan, X.: CTransNet: Convolutional Neural Network Combined with Transformer for Medical Image Segmentation. Computing and Informatics, Vol. 42, 2023, No. 2, pp. 392–410, doi: 10.31577/cai_2023_2_392.

[31] Chen, L. C.—Zhu, Y.—Papandreou, G.—Schroff, F.—Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11211, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[32] Badrinarayanan, V.—Kendall, A.—Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2017, No. 12, pp. 2481–2495, doi: 10.1109/TPAMI.2016.2644615.

[33] Yang, B.—Liu, C.—Li, B.—Jiao, J.—Ye, Q.: Prototype Mixture Models for Few-Shot Semantic Segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12353, 2020, pp. 763–778, doi: 10.1007/978-3-030-58598-3_45.

[34] Yang, L.—Zhuo, W.—Qi, L.—Shi, Y.—Gao, Y.: Mining Latent Classes for Few-Shot Segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8701–8710, doi: 10.1109/ICCV48922.2021.00860.

[35] Gueguen, L.—Sergeev, A.—Liu, R.—Yosinski, J.: Faster Neural Networks Straight from JPEG. 6th International Conference on Learning Representations (ICLR 2018), 2018, https://openreview.net/forum?id=S1ry6Y1vG.

[36] Xu, K.—Qin, M.—Sun, F.—Wang, Y.—Chen, Y. K.—Ren, F.: Learning in the Frequency Domain. 2020 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2020, pp. 1737–1746, doi: 10.1109/CVPR42600.2020.00181.

[37] Tang, C.—Wang, X.—Bai, Y.—Wu, Z.—Zhang, J.—Huang, Y.: Learning Spatial-Frequency Transformer for Visual Object Tracking. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 33, 2023, No. 9, pp. 5102–5116, doi: 10.1109/TCSVT.2023.3249468.

[38] Zhang, Y.—Li, Q.—Qi, M.—Liu, D.—Kong, J.—Wang, J.: Multi-Scale Frequency Separation Network for Image Deblurring. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 33, 2023, No. 10, pp. 5525–5537, doi: 10.1109/TCSVT.2023.3259393.

[39] Wang, H.—Wu, X.—Huang, Z.—Xing, E. P.: High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8681–8691, doi: 10.1109/CVPR42600.2020.00871.

[40] Geirhos, R.—Rubisch, P.—Michaelis, C.—Bethge, M.—Wichmann, F. A.—Brendel, W.: ImageNet-Trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. 7th International Conference on Learning Representations (ICLR 2019), 2019, https://openreview.net/forum?id=Bygh9j09KX.

[41] Qin, Z.—Zhang, P.—Wu, F.—Li, X.: FcaNet: Frequency Channel Attention Networks. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 763–772, doi: 10.1109/ICCV48922.2021.00082.

[42] Azad, R.—Fayjie, A. R.—Kauffmann, C.—Ayed, I. B.—Pedersoli, M.—Dolz, J.: On the Texture Bias for Few-Shot CNN Segmentation. IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2673–2682, doi: 10.1109/WACV48630.2021.00272.

[43] Luo, C.—Lin, Q.—Xie, W.—Wu, B.—Xie, J.—Shen, L.: Frequency-Driven Imperceptible Adversarial Attack on Semantic Similarity. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15294–15303, doi: 10.1109/CVPR52688.2022.01488.

[44] Zhang, C.—Lin, G.—Liu, F.—Guo, J.—Wu, Q.—Yao, R.: Pyramid Graph Networks with Connection Attentions for Region-Based One-Shot Semantic Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9586–9594, doi: 10.1109/ICCV.2019.00968.

[45] Deng, J.—Dong, W.—Socher, R.—Li, L. J.—Li, K.—Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[46] Lang, C.—Cheng, G.—Tu, B.—Han, J.: Learning What Not to Segment: A New Perspective on Few-Shot Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8047–8057, doi: 10.1109/CVPR52688.2022.00789.

[47] Nguyen, K.—Todorovic, S.: Feature Weighting and Boosting for Few-Shot Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 622–631, doi: 10.1109/ICCV.2019.00071.

[48] Ding, H.—Zhang, H.—Jiang, X.: Self-Regularized Prototypical Network for Few-Shot Semantic Segmentation. Pattern Recognition, Vol. 133, 2023, Art. No. 109018, doi: 10.1016/J.PATCOG.2022.109018.

[49] Everingham, M.—Van Gool, L.—Williams, C. K. I.—Winn, J.—Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, Vol. 88, 2010, No. 2, pp. 303–338, doi: 10.1007/S11263-009-0275-4.

[50] Hariharan, B.—Arbeláez, P.—Girshick, R.—Malik, J.: Simultaneous Detection and Segmentation. In: Fleet, D. J., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.): Computer Vision – ECCV 2014. Springer, Cham, Lecture Notes in Computer Science, Vol. 8695, 2014, pp. 297–312, doi: 10.1007/978-3-319-10584-0_20.

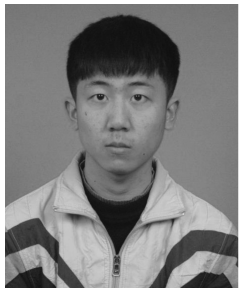
[51] Lin, T. Y.—Maire, M.—Belongie, S.—Hays, J.—Perona, P.—Ramanan, D.—Dollár, P.—Zitnick, C. L.: Microsoft COCO: Common Objects in Context. In: Fleet, D. J., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.): Computer Vision – ECCV 2014. Springer, Cham, Lecture Notes in Computer Science, Vol. 8693, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[52] Min, J.—Kang, D.—Cho, M.: Hypercorrelation Squeeze for Few-Shot Segmenation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6921–6932, doi: 10.1109/ICCV48922.2021.00686.

[53] Wu, Z.—Shi, X.—Lin, G.—Cai, J.: Learning Meta-Class Memory for Few-Shot Semantic Segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 497–506, doi: 10.1109/ICCV48922.2021.00056.

[54] Mao, B.—Zhang, X.—Wang, L.—Zhang, Q.—Xiang, S.—Pan, C.: Learning from the Target: Dual Prototype Network for Few Shot Semantic Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, No. 2, pp. 1953–1961, doi: 10.1609/aaai.v36i2.20090.

[55] Lang, C.—Tu, B.—Cheng, G.—Han, J.: Beyond the Prototype: Divide-and-Conquer Proxies for Few-Shot Segmentation. In: De Raedt, L. (Ed.): Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022). 2022, pp. 1024–1030, doi: 10.24963/ijcai.2022/143.



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

Huang Hui received her Ph.D. degree in the Department of Information and Image Processing from the University of Rennes, Rennes, France, in collaboration with the Institute Mines-Télécom, Télécom Bretagne, and LaTIM Laboratory, INSERM U1101, Brest, France, in 2011. She received her B.Sc. and her M.Sc. degree from the School of Biomedical Engineering, Southeast University, Nanjing, China, in 2004 and 2007. She is currently Associate Professor of College of Information, Mechanical and Electrical Engineering, Shanghai Normal University. She is also a member of the IEEE and 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 is currently Professor of the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, China. She received her Ph.D. degree from Shanghai Jiao Tong University, China. Her research interests include data mining, image segmentation, and action recognition.

Feiniu Yuan received his B.Eng. and his M.Eng. degrees in mechanical engineering from the Hefei University of Technology, Hefei, China, in 1998 and 2001, respectively, and his Ph.D. degree in pattern recognition and intelligence system from the University of Science and Technology of China (USTC), Hefei, in 2004. He is a senior member of IEEE and CCF. From 2004 to 2006, he worked as a post-doctor with 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 Professor, Ph.D. supervisor and a Vice Dean with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, China. His research interests include deep learning, image segmentation, pattern recognition and 3D modeling.

Hongqing Zhu is currently Professor at the East China University of Science and Technology, Shanghai. She received her Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000. From 2003 to 2005, she was a Post-Doctoral 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 IEEE and IEICE.

Peng Zhu received his Master Degree in the Department of Biomedical Engineering, Southeast University, Nanjing, China. He is the Chief Engineer of Shanghai Vixdetect Inspection Equipment Co., Ltd. His development interests include high resolution industrial X-Ray imaging and foreign body detection image equipment based on artificial intelligence.