IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 March 2024, accepted 7 April 2024, date of publication 12 April 2024, date of current version 23 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387742

RESEARCH ARTICLE

Prototype Comparison Convolutional Networks for One-Shot Segmentation

LINGBO LI[1], ZHICHUN LI[2], FUSEN GUO[3], (Graduate Student Member, IEEE), HAOYU YANG[4], JINGTIAN WEI[5], AND ZHENGYI YANG[5]

[1]Library of Information Center, Zhejiang Technical Institute of Economics, Hangzhou 310018, China
[2]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, SAR, China
[3]School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia
[4]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
[5]School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia

Corresponding author: Jingtian Wei (wjtohoh@163.com)

ABSTRACT In few-shot semantic segmentation (FSS), the key challenges are efficiently tuning the interaction between the support set and the query set and distinguishing between context, background, and interfering items. To address these challenges, we propose prototype comparison networks for one-shot segmentation (OPCN) to capture the details required for FSS. Specifically, we offer the Fusion Interaction Module (FIM) to improve the segmentation performance by capturing the correlation and semantic information between the support set and query set features. Subsequently, we propose the Feature Enhancement Module (FEM), which aims to enhance the information required in the support set and query set features while increasing the focus on critical details by reducing the weight of the background regions to provide a more targeted feature representation for subsequent query image segmentation. Then, we propose the Feature Refinement Module (FRM) to filter irrelevant background information in the features and specify the target location region. Finally, the Feature Matching Module (FM) generates the final segmentation mask for the query image. Extensive experiments on the PASCAL-5i and COCO-20i datasets show that our approach achieves excellent performance in the case of the one-shot setup.

INDEX TERMS Few-shot semantic segmentation, convolutional neural network, prototype network, feature matching.

I. INTRODUCTION

Few-shot learning (FSL) is a challenging problem in computer vision and machine learning, where the task is to classify or segment objects with limited labeled examples [1], [2], [3]. Traditional learning algorithms need help generalizing well to unseen classes in such scenarios due to the scarcity of training data. This limitation hinders the applicability of machine learning models in real-world settings where obtaining a large amount of labeled data for every class is often impractical [4], [5]. Few-shot learning has attracted considerable interest owing to its potential utility across diverse domains such as robotics, natural language processing, and computer vision [6], [7], [8]. By enabling machines to quickly adapt and learn from a few examples, FSL opens up possibilities for more flexible and efficient learning systems.

Among the different tasks in few-shot learning, the challenge is more prominent in several semantic segmentation scenarios [9], [10], [11]. Because there is a restricted quantity of labeled instances available for each category, the model must be able to segment categories that have yet to be previously observed accurately. This requires the model to understand the spatial context, boundaries, and appearance of objects despite the limited amount of labeled data. Existing methods can be categorized into two main groups: meta-learning and metric-based learning [12], [13], [14], [15]. With meta-learning, the model can quickly adapt to new categories in several sample contexts, thus better handling unseen semantic categories. Metric learning-based

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

methods, however, focus on learning similarity measures between samples, enabling the model to better generalize to unknown categories in small sample scenarios. Although these methods have achieved significant success in handling learning from few-shot samples, there is still potential for further improving segmentation accuracy and generalization to unseen categories.

Currently, state-of-the-art FSS methods mainly use prototype-based strategies. These methods extract critical information from the supporting features by masked average pooling to form category-representative prototypes. These prototypes are abstract representations of target objects and fuse information with query features through interaction mechanisms such as cosine similarity and feature splicing. Accurate prediction of objects by the model is achievable in the query image in few-shot scenarios through the prototype learning and interaction process. However, these methods often rely on a limited number of prototypes to mine more information While sacrificing specific inherent details of objects in query images.

In response to the abovementioned challenges, we present a novel approach, the Prototype Comparison Network (OPCN), designed to address the complexities inherent in accurate Few-Shot Segmentation (FSS). First, we propose the Fusion Interactive Module (FIM), facilitating improved feature extraction by fostering joint attention between the support set and query set features. Subsequently, the Feature Enhancement Module (FEM) is proposed with the explicit objective of enhancing pertinent information within the support set and query set features. The model focuses more on critical data by reducing the weights assigned to background regions. This weight adjustment empowers the model to comprehend better and exploit support set information, yielding a more targeted feature representation for subsequent query image segmentation, thereby enhancing overall performance. Following this, the feature refinement module (FEM) is employed to filter out extraneous background details from the features and elucidate target location information. Lastly, the feature matching module (FM) is deployed to generate the final segmentation mask for the query image. Our main contributions are as follows:

- We propose the Fusion Interaction Module (FIM), which optimizes convolutional features by introducing Non-local blocks and can enhance the feature distribution on the target input based on the reference input.
- We propose both the Feature Enhancement Module (FEM) and the Feature Refinement Module (FRM) to augment the feature representation, filter out extraneous background information and clarify the target location information.
- Extensive experiments on PASCAL-5^i and COCO-20^i show that our proposed prototype comparison networks for one-shot segmentation achieve significant performance improvement over other methods.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Recently, semantic segmentation has experienced remarkable advancements. Deep Convolutional Neural Networks (CNNs) such as U-Net [16], SegNet [17], and DeepLab [18] have propelled the domain forward by harnessing robust feature extraction capabilities and incorporating architectural components like skip connections and atrous spatial pyramid pooling. The adoption of encoder-decoder architectures, exemplified by Fully Convolutional Networks (FCN) [19], has become widespread. These architectures employ an encoder to capture high-level features and a decoder for making dense pixel-wise predictions.

Attention mechanisms, represented by models like Non-local Neural Networks [20] and SAGAN (Self-Attention Generative Adversarial Networks) [21], have been introduced to model long-range dependencies and enhance contextual understanding in semantic segmentation tasks. This allows the networks to focus on relevant image regions and improve the overall segmentation accuracy.

Efforts have also been directed towards designing efficient networks, such as EfficientNet [22] and MobileNet [23], which balance accuracy and computational efficiency. This is especially vital in resource-constrained scenarios where computational resources are restricted. These efficient architectures contribute to achieving satisfactory performance without compromising computational efficiency [24], [25].

However, a notable challenge persists in semantic segmentation - the reliance on large amounts of accurately labeled data for training [26], [27], [28], [29]. Acquiring such data is often a time-consuming, costly, and labor-intensive process. The scarcity of labeled data presents a significant hurdle, especially when data collection or labeling is inherently challenging. Addressing this issue is crucial for the continued progress of semantic segmentation methods, particularly in real-world scenarios with limited labeled data availability.

B. FEW-SHOT SEMANTIC SEGMENTATION

Few-shot semantic segmentation aims at segmenting invisible object classes in a query image using only a small number of annotated samples. Most existing FSS methods use a prototype-based approach, where a meta-learning architecture is used to meta-train the base class and then meta-test new disjoint classes [30], [31], [32]. The learned prototypes represent the average features of each class. In recent years, various prototype-based FSS methodologies have surfaced in the research domain, such as CANet [33]adaptively knows the importance of features by introducing a global attention mechanism and applying it to the interaction between the support set and the query set to improve the segmentation accuracy.SG-One [34] employs a masked average pooling strategy to obtain robust bootstrap features from the supported image and uses cosine similarity to establish the relationship between bootstrap features and query image features.PANet [9] uses a metric learning approach to

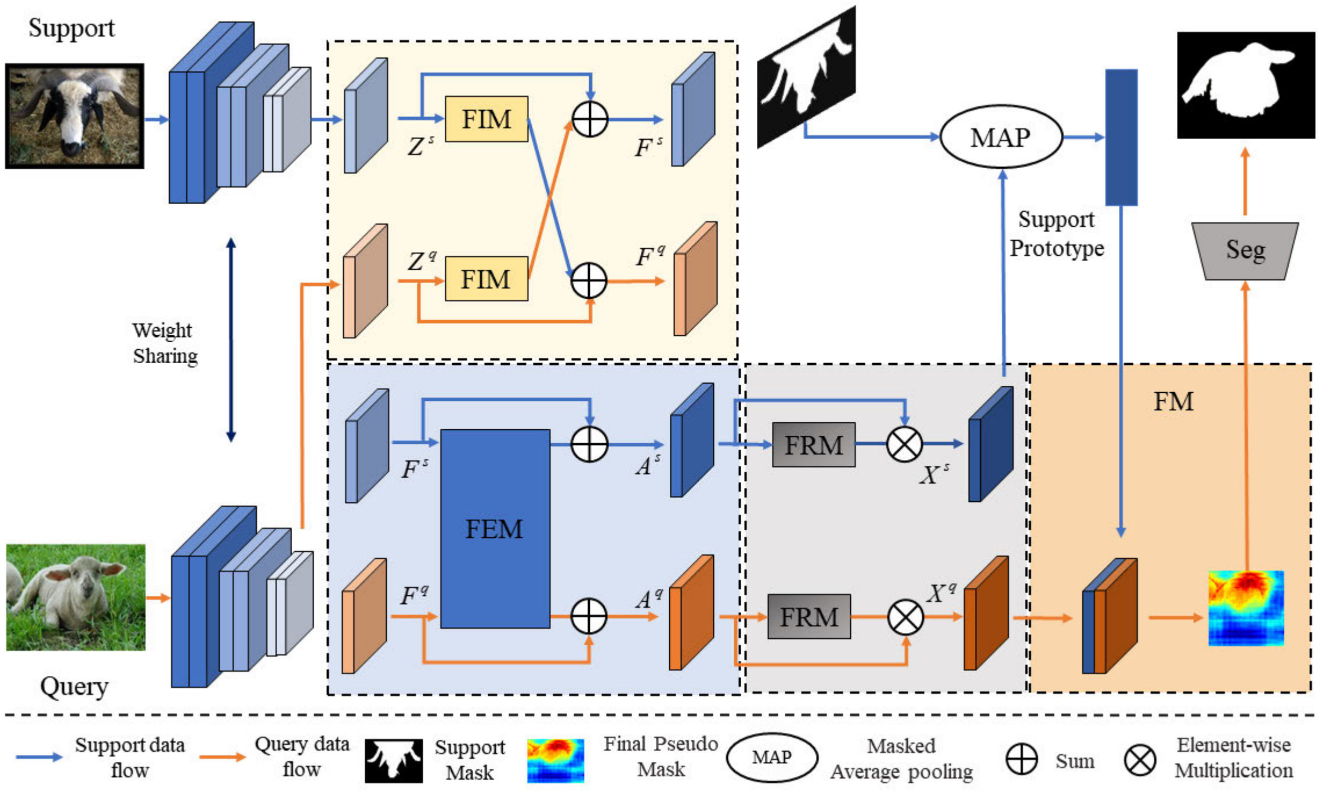


FIGURE 1. The overall architecture of our proposed Prototype Comparison Convolutional Networks for One-Shot Segmentation (OPCN). After extracting the features of the support and query images through a pre-trained backbone network, a fusion interaction module (FIM) is introduced to capture the association and semantic information between the features of the support and query sets and generate the co-attention features F^s and F^q. Meanwhile, the Feature Enhancement Module (FEM) is utilized to enhance the necessary information in the support and query set features and obtain the enhanced features A^s and A^q by residualizing them with F^s and F^q. The third part is to obtain the Feature Refinement Module (FRM), which filters irrelevant contextual information from the features and specifies the target location information to get the refinement features X^s and X^q. Finally, the feature matching module (FM) is utilized for final target prediction.

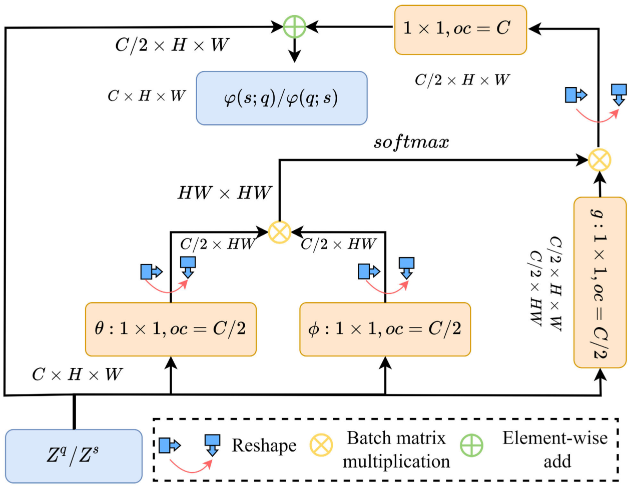


FIGURE 2. Detailed process and structure of non-local block.

learn category-specific prototypical representations from a small number of support images in the embedding space and segment the query image by matching each pixel to the learned prototype.PMMs [35], estimated through an expectation-maximization algorithm, amalgamate abundant channel and spatial semantics from a restricted set of supported images.PFENet [36] addresses the challenges of generalizability and spatial inconsistency in less-sample segmentation with a priori mask generation and feature enhancement modules that require no training. The Dynamic Prototype Convolutional Network (DPCN) [37] focuses on learning dynamic prototypes to adapt to new categories and improve segmentation performance.

However, the inherent limitation of prototypes in existing prototype learning methods leads to inevitable information loss. In contrast to prior methodologies, we employ the Fusion Interactive Module (FIM) to execute a co-attention mechanism between support and query features. This allows us to extract the maximum complementary target information from the support and query features. In addition, we use Feature Enhancement Module (FEM) and Feature Refinement Module (FEM) to clarify the query set target location and use Feature Matching Module (FM) to generate the final segmentation mask for the query image.

III. PROPOSED METHOD

A. PROBLEM DEFINITION

Few-shot semantic segmentation is an image segmentation task involving a limited set of annotated instances. Its problem definition involves training a model to learn from

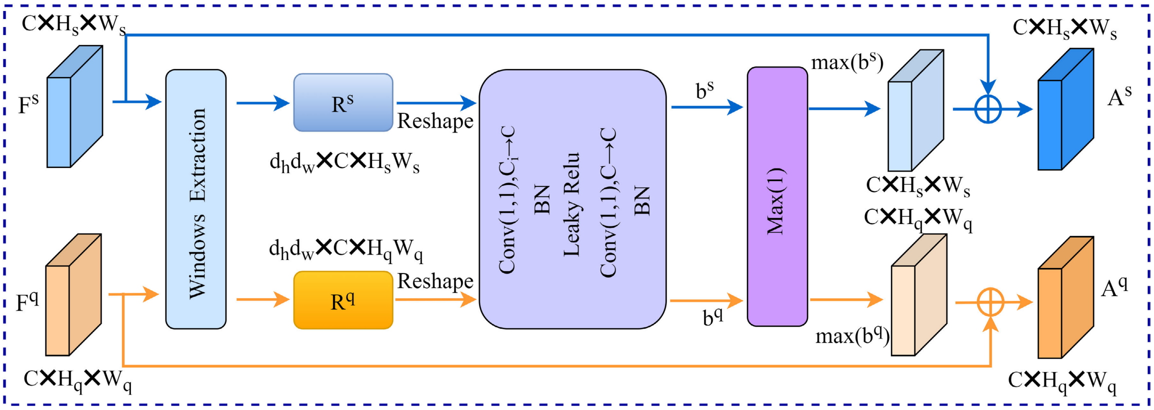


FIGURE 3. The specific structure diagram of Feature Enhancement Module (FEM).

several annotated samples and generalize to unseen target categories. The few-shot semantic segmentation dataset consists of two main parts: the training set (D_{base}) and the test set (D_{novel}). The training set is used for the meta-training of the model, while the test set is used to evaluate the model's generalization

performance. The set of categories in the dataset is categorized into training categories (C_{base}) and test categories (C_{novel}). Where $C_{base} \cap C_{novel} = \emptyset$, i.e., there is no intersection between training and testing categories. For the k-shot setup, each plot is denoted as (S, Q), where S contains a support set marked as $\mathcal{S} = \left\{\left(I_s^i, M_s^i\right)\right\}_{i=1}^{k}$, where I_s^i denotes the supported image and M_s^i denotes the corresponding binary mask. The query set Q contains the query images and the related masks, denoted as Q = (Iq, Mq). During the meta-training phase, the model takes the support set S and the query set image Iq as input in a specific category c, generating a prediction mask y_q for the query image. The goal of the task is to make the prediction mask as close as possible to the mask Mq. During the testing phase, the model undergoes evaluation by randomly sampling the test set to assess its ability to generalize to categories not encountered during training.

B. OVERVIEW

Our Prototype Comparison Convolutional Networks for One-Shot Segmentation (OPCN) consists of four key modules: Fusion Interactive Module (FIM), Feature Enhancement Module (FEM), Feature Refinement Module (FRM), and Feature Matching Module (FM)(Illustrated in Fig 1). Specifically, when provided with support and query images, Is and Iq, we employ a backbone network with shared weights to extract fundamental features. By using the interaction module (FIM), the association and semantic information between the support and query set features are captured. Following this, the Feature Enhancement Module (FEM) is introduced, specifically designed to explicitly amplify crucial information within the features of both the support and query sets. After FEM comes to Feature Refinement Module (FRM), which aims at filtering irrelevant contextual information from features and specifying target location information. Lastly, the Feature matching Module (FM) is applied to produce the ultimate segmentation mask for the query image. Subsequently, we provide a detailed description of each of the modules above.

C. FUSION INTERACTIVE MODULE

Inspired by previous work [38], to extract the features standard to the support and query sets and efficiently tune the interaction between them, we enhance the feature distribution on the target input based on the reference input. As shown in Fig 2, Specifically when the output of the support set is $Z^s \in \mathbb{R}^{C\times H_s\times W_s}$. The output of the query set is $Z^q \in \mathbb{R}^{C\times H_q\times W_q}$, where C, is the channel dimension, and H_s, W_s, H_q, W_q are the height and width of support and query features, respectively. Taking Z^q as the reference input, the output of the non-local block [20] of Z^s is $\varphi(s; q) \in \mathbb{R}^{C\times H_s\times W_s}$. Similarly, using Z^s as the reference input, one can obtain $\varphi(q; s) \in \mathbb{R}^{C\times H_q\times W_q}$ for Z^q, and the interaction between support and query feature can be thought of as carrying out the co-attention

$$F^s = Z^s \oplus \psi(s; q) \tag{1}$$

$$F^q = Z^q \oplus \psi(q; s) \tag{2}$$

The two extended feature maps are Eq. 1 and Eq. 2, which $\oplus$ are element-wise sum. Since F^q contains not only the features of the query set image but also the support and query weighted features, more information about the support set images can be found on this layer of features, which makes it easier to capture the critical features of the target object.

D. FEATURE ENHANCEMENT MODULE

FSS models [39], [40] typically utilize a priori masks to represent the approximate location of the target object in the query image. However, since these a priori masks are usually obtained by pairing elements between feature maps or region-based matching, they often neglect the overall context. We introduce the FEM module to accentuate pertinent features while suppressing irrelevant ones (Illustrated in Fig 3). This enhancement improves the model's capacity to concentrate on informative regions within the image. Taking inspiration from DPCN [37], we use co-attention support features $F^s \in \mathbb{R}^{C\times H_s\times W_s}$ and co-attention query features $F^q \in \mathbb{R}^{C\times H_q\times W_q}$ as input. To achieve comprehensive matching, the generated regional features R^s and R^q are expressed as follows:

$$R^s = \mathcal{W}\left(F^s\right) \in \mathbb{R}^{d_h d_w\times C_h\times H_s W_s} \tag{3}$$

$$R^q = \mathcal{W}\left(F^q\right) \in \mathbb{R}^{d_h d_w\times C_h\times H_q W_q} \tag{4}$$

where W is the sliding window operation, $W(F^s)$ and $W(F^q)$ are the sliding fixed window operations used independently on the support and query features, d_h and d_w is the window height and width. In our experiments, we choose symmetric windows, $(d_h, d_w) \in (5, 5)$. However, the above approach only represents the approximate location of the target object lacking the corresponding features of the channel dimensions; for this reason, unlike before, instead of generating a region-matching map, we merge the dimensions of dh, dw to get a new channel C_i and recover the number of channels as C by two convolutional blocks and take the maximum value of the channel dimension feature and residualized it with the common feature F to get the final augmented feature A^s and A^q.

$$\mathrm{a}^s = \mathcal{F}_{1\times 1}\left(R^s\right) \in \mathbb{R}^{C_i\times H_s\times W_s} \tag{5}$$

$$\mathrm{b}^s = \mathcal{F}_{1\times 1}\left(a_s\right) \in \mathbb{R}^{C\times H_s\times W_s} \tag{6}$$

$$A^s = F^s + \max\left(b^s\right) \tag{7}$$

where $\mathcal{F}_{1\times 1}$ denotes a 1×1 convolutional layer, $C_i = C \times d_h \times d_w$.Similarly, we can get the final augmented feature as A^q, which localizes the target object more accurately.

E. FEATURE REFINEMENT MODULE

In the preceding stage, we refined the features of the query and the support set, yielding a preliminary estimation of the target object's location. Nevertheless, achieving precise segmentation requires finer pixel-level predictions. We propose a feature refinement module when the dataset is limited. The channel grouping approach is employed to heighten the model's sensitivity to information within the feature channels, thereby amplifying the spatial semantic distribution of features.

As illustrated in Fig 4, the initial step involves grouping feature channels. For the input feature vector matrix, denoted as $A \in \mathbb{R}^{C\times H\times W}$, the input vectors are organized into G groups (G set to 4). This grouping operation yields the group eigenvector matrix $A^g \in \mathbb{R}^{C'\times H\times W}$, where $C' = C/4$, and A^g represents the eigenvector matrix of the g-th group. The group feature vector matrix A^g undergoes a global average pooling along the H and W dimensions. This process results in semantic vectors $I_g^h \in \mathbb{R}^{C'\times 1\times W}$ along the H dimension and $I_g^w \in \mathbb{R}^{C'\times H\times 1}$ along the W dimension. To introduce spatially localized contextual information and enhance the model's perceptual and expressive capabilities. Unlike before [41], we performed the convolution operation on the H and W dimensions with the semantic vectors, respectively. The semantic vectors I_g^h, I_g^w after the convolution operation are defined by the following equation:

$$I_{(g,c)}^h(h) = \mathcal{F}_{3\times 3}(\frac{1}{W}\sum_{0\leq i<W\prec} A_{(g,c)}(h, I)) \tag{8}$$

$$I_{(g,c)}^w(w) = \mathcal{F}_{3\times 3}(\frac{1}{H}\sum_{0\leq j<H\prec})A_{(g,c)}(j, w)) \tag{9}$$

where $\mathcal{F}_{3\times 3}$ represents a 3×3 convolutional layer,g signifies the g-th group of the feature vector matrix, and c denotes the feature matrix of the c-th channel in the set of vector matrices. Furthermore, (h, i) denotes the i-th point in the h-th row of the feature matrix, and (j, w) denotes the j-th point in the w-th row of the feature matrix. Then, the semantic vectors are multiplied with the group feature vector matrix to orient the spatial association strength matrix E^g between the semantic features. The following equation defines the association strength matrix E^g.

$$E^g = A^g \otimes I_g^h \otimes I_g^w \tag{10}$$

After obtaining the association strength matrix E^g, it is normalized to the spatial dimension. The sigmoid function is then applied to derive the final attention mapping map $\hat{E}^g$. The formulation for $\hat{E}^g$ is as follows.

$$\hat{E}^g = f\left(Norm\left(E^g\right)\right) \tag{11}$$

where $f()$ is the sigmoid function, and Norm denotes regularization. Finally, the matrix multiplication of the enhanced features A using the attention mapping map $\hat{E}^g$ results in the final refined features X^s and X^q:

$$X^s = \mathrm{reshape}\left(A^s \otimes \hat{E}^g\right) \tag{12}$$

$$X^q = \mathrm{reshape}\left(A^q \otimes \hat{E}^g\right) \tag{13}$$

X^s and X^q represent the final refined feature that discards irrelevant background and specifies the target's location.

F. FEATURE MATCHING MODULE

Inspired by previous work [44], we extend the application of dense feature matching to activate the target object region on X^q. To elaborate, the support prototype $P^s \in \mathbb{R}^{1\times 1\times C}$ is initially obtained through mask average pooling (MAP) on the support feature map X^s. This prototype is then expanded to $\hat{\boldsymbol{P}}^s \in \mathbb{R}^{H\times W\times C}$ and connected to the refined query feature X^q. Additionally, we incorporate an a priori confidence map $C^q \in \mathbb{R}^{H\times W\times 1}$ by determining the maximum similarity score at the pixel level, following the approach outlined in [44]. Subsequently, we derive the activated query feature $X_{act}^q \in \mathbb{R}^{H\times W\times C}$ through the initial target object prediction $y^q \in \mathbb{R}^{H\times W\times 1}$.

$$\boldsymbol{X}_{act}^q = \mathcal{F}_{1\times 1}\left(\boldsymbol{X}^q \oplus \hat{\boldsymbol{P}}^s \oplus \boldsymbol{C}^q\right) \tag{14}$$

$$y^q = \mathcal{F}_{3\times 3}\left(\boldsymbol{X}_{act}^q\right) \tag{15}$$

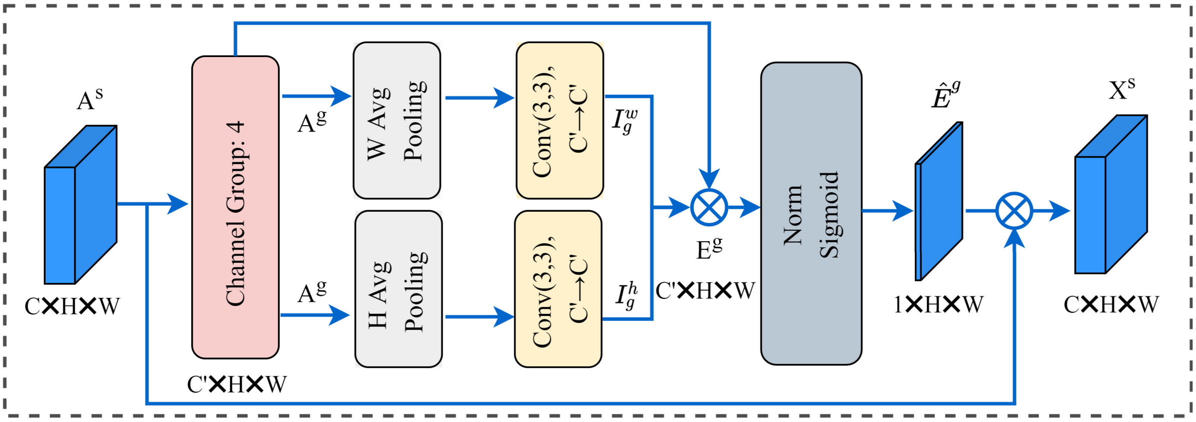


FIGURE 4. The specific structure diagram of Feature Refinement Module (FRM).

where $y^q \in \mathbb{R}^{H\times W\times 1}$ represents the ultimate segmentation outcome for the target object as produced by our entire model.

G. TRAINING LOSS

We binarize the final target prediction y^q as a prediction mask for the query image [37]. Subsequently, we employ training with Binary Cross-Entropy loss (BCE), utilizing the BCE loss between the prediction mask and the actual mask of the query image as our primary loss function.

$$\mathcal{L}_{\text{seg}}^{s\rightarrow q} = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} BCE\left(y^q(i,j), M_q(i,j)\right) \tag{16}$$

Parallel to Equation (16), we obtain the predicted support mask by calculating the Binary Cross-Entropy (BCE) loss between P^s and M_S, and an additional loss is obtained:

$$\mathcal{L}_{\text{seg}}^{q\rightarrow s} = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} BCE\left(P^s(i,j), M_s(i,j)\right) \tag{17}$$

Predicting the query and support masks share identical structures and parameters. To summarize, the ultimate loss is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{seg}}^{s\rightarrow q} + \lambda \mathcal{L}_{\text{seg}}^{q\rightarrow s} \tag{18}$$

where λ represents the weight that balances the contribution of each branch and is consistently set to 1.0 in all experimental settings.

IV. EXPERIMENTAL RESULTS

A. DATASET

In our research, we employ two widely recognized benchmarks for evaluation purposes: PASCAL-5^i [42] and COCO-20^i [52]. The PASCAL-5^i benchmark is derived from the combination of VOC2012 [55] and SBD [56] annotations. It focuses on 20 object classes divided into four folds for a robust 4-fold cross-validation. Five classes are designated for testing during this evaluation process, while the remaining 15 classes are utilized for training. On the other hand, the COCO-20^i benchmark is more challenging and originates from MSCOCO [57]. It encompasses a broader range of 80 object classes. Like the PASCAL benchmark, we divide the COCO-20^i classes into four folds, each consisting of 20 classes. In this case, 20 classes are reserved for testing purposes, while 60 classes are employed for training our models.

B. EXPERIMENTAL DETAILS

For our few-shot segmentation experiments, we employed VGG16 [58] and ResNet-50 [59] as the backbone networks, following established experimental setups. These backbone networks were pre-trained on the ImageNet classification task, and their weights remained fixed during training. We implemented our network using PyTorch and performed the experiments on NVIDIA RTX 3090 GPUs. During training, we cropped all images to 473 × 473, randomizing the crop size.

The optimization process employed Stochastic Gradient Descent (SGD) as the optimizer, with an initial learning rate of 0.05, a batch size of 8, weight decay set to 0.0001, and momentum set to 0.9. The training was conducted on the few-shot datasets PASCAL-5^i and COCO-20^i for 200 and 50 epochs, respectively. The learning rate underwent attenuation using a polynomial annealing strategy, with the power set to 0.1. During the evaluation phase, we followed the methodology outlined in previous work, randomly selecting 1000 pairs of support queries for the PASCAL-5^i dataset and 4000 pairs for the COCO-20^i dataset for evaluation.

C. EVALUATION METRICS

Few-shot segmentation refers to the image segmentation task with limited data. Two standard evaluation criteria are mIoU (Mean Intersection over Union) and FB-IoU (Foreground-Background Intersection over Union). mIoU: is the standard metric for semantic segmentation. It calculates the average of the intersection ratio between the predicted

TABLE 1. Comparison with state-of-the-arts on PASCAL-5^i dataset under 1-shot.

Methods	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
OSLSM [42]	VGG16	33.6	55.3	40.9	33.5	40.8	61.3
co-FCN [13]	VGG16	36.7	50.6	44.9	32.4	41.1	60.1
AMP-2 [43]	VGG16	41.9	50.2	46.7	34.7	43.4	61.9
RPMM [35]	VGG16	47.1	65.8	50.6	48.5	53.0	-
PFENet [36]	VGG16	56.9	68.2	54.4	52.4	58.0	72.3
MMNet [39]	VGG16	57.1	67.2	56.6	52.3	58.3	-
NTRENet [44]	VGG16	57.7	67.6	57.1	53.7	59.0	73.1
HSNet [45]	VGG16	59.6	65.7	59.6	54.0	59.7	73.4
DPCN [37]	VGG16	58.9	69.1	63.2	55.7	61.7	73.7
CANet [33]	ResNet50	52.5	65.9	51.3	51.9	55.4	66.2
RPMM [35]	ResNet50	55.2	66.9	52.6	50.7	56.3	-
ASGNet [46]	ResNet50	58.8	67.9	56.8	53.7	59.3	69.2
ReRPI [47]	ResNet50	59.8	68.3	62.1	48.5	59.7	-
PFENet [36]	ResNet50	61.7	69.5	55.4	56.3	60.8	73.3
SCL [48]	ResNet50	63.0	70.0	56.5	57.7	61.8	71.9
MMNet [39]	ResNet50	62.7	70.2	57.3	57.0	61.8	-
SAGNN [40]	ResNet50	64.7	69.6	57.0	57.3	62.1	73.2
MLC [49]	ResNet50	59.2	71.2	65.6	52.5	62.1	-
NTRENet [44]	ResNet50	65.4	72.3	59.4	59.8	64.2	77.0
SiGCN [50]	ResNet50	65.1	70.1	65.2	60.8	65.3	77.5
DPCN [37]	ResNet50	65.7	71.6	**69.1**	60.6	66.7	78.0
QCLNet [51]	ResNet50	**67.9**	72.4	64.3	**63.4**	67.0	-
OPCN(ours)	VGG16	59.4	69.3	62.2	56.4	61.8	73.1
OPCN(ours)	ResNet50	66.2	**72.5**	68.4	61.3	**67.1**	**78.2**

segmentation results and the accurate segmentation mask. For each category, the intersection area of its predicted area and the actual area are calculated, divided by the union area, and then averaged across all categories. The specific formula is as follows:

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{19}$$

N represents the number of samples; P_i represents the predicted segmentation area of the ith sample; G_i represents the actual segmentation area of the ith sample; $|\cdot|$ represents the region's area. FB-IoU refers to calculating the IoU of the foreground and background separately to evaluate the segmentation performance between foreground objects and background. The formula is as follows:

$$\text{FB-IoU} = \frac{|P_{\mathrm{fg}} \cap G_{\mathrm{fg}}|}{|P_{\mathrm{fg}} \cup G_{\mathrm{fg}}|} + \frac{|P_{\mathrm{bg}} \cap G_{\mathrm{bg}}|}{|P_{\mathrm{bg}} \cup G_{\mathrm{bg}}|} \tag{20}$$

where P_{fg} and P_{bg} respectively represent the foreground and background areas of predicted segmentation; G_{fg} and G_{bg} represents the real segmented foreground and background areas respectively.

D. COMPARISON WITH OTHER METHODS

1) PASCAL-5^i

The performance of our method was rigorously evaluated against several state-of-the-art models in a 1-shot setting on the PASCAL-5^i dataset. The comprehensive results

TABLE 2. Comparison with state-of-the-arts on COCO-20^i dataset under 1-shot.

Methods	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
FWB [52]	VGG16	18.4	16.7	19.6	25.4	20.0	-
PFENet [36]	VGG16	33.4	36.0	34.1	32.8	34.1	60.0
SAGNN [40]	VGG16	35.0	40.5	37.6	36.0	37.3	61.2
DPCN [37]	VGG16	38.5	43.7	38.2	37.7	39.5	62.5
PPNet [53]	ResNet50	28.1	30.8	29.5	27.7	29.0	-
RPMM [35]	ResNet50	29.5	36.8	28.9	27.0	30.6	-
MLC [49]	ResNet50	**46.8**	35.3	26.2	27.1	33.9	-
RePRI [46]	ResNet50	31.2	38.1	33.3	33.0	34.0	-
MMNet [39]	ResNet50	34.9	41.0	37.2	37.0	37.5	-
HSNet [43]	ResNet50	36.3	43.1	38.7	38.7	39.2	68.2
NTRENet [44]	ResNet50	36.8	42.6	39.9	37.9	39.3	**68.5**
SiGCN [50]	ResNet50	38.7	46.3	43.1	37.5	41.4	62.7
IMR-HSNet [54]	ResNet50	39.5	43.8	42.4	**44.1**	42.4	-
DPCN [37]	ResNet50	42.0	47.0	43.2	39.7	43.0	63.2
OPCN(ours)	VGG16	39.2	44.1	37.8	37.2	39.6	62.7
OPCN(ours)	ResNet50	43.2	**47.5**	**43.4**	39.6	**43.4**	64.5

TABLE 3. Ablation experiments for main modules on PASCAL-5^i and COCO-20^i dataset.

id	FIM	FEM	FRM	PASCAL-5^i						COCO-20^i					
				Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
(a)	×	×	×	60.8	69.5	58.2	56.7	61.3	71.2	39.6	42.2	39.8	35.4	39.3	59.8
(b)	×	✓	×	62.1	69.9	62.7	57.3	63.0	73.4	40.2	42.8	40.6	36.1	40.0	60.5
(c)	×	×	✓	62.5	70.1	63.1	57.8	63.4	73.9	40.3	43.2	40.5	36.3	40.1	60.7
(d)	✓	×	×	63.1	70.4	63.8	58.2	63.9	74.6	41.2	43.9	40.8	36.5	40.6	61.3
(e)	×	✓	✓	63.4	70.6	65.2	58.4	64.4	75.1	41.5	44.8	41.0	36.9	41.1	61.7
(f)	✓	×	✓	65.7	71.8	67.9	60.5	66.5	76.6	42.6	46.7	42.3	38.2	42.5	63.1
(g)	✓	✓	×	64.2	70.9	66.4	59.6	65.2	75.8	41.9	45.3	41.8	37.2	41.6	62.1
(h)	✓	✓	✓	**66.2**	**72.5**	**68.4**	**61.3**	**67.1**	**78.2**	**43.2**	**47.5**	**43.4**	**39.6**	**43.4**	**64.5**

are summarized in Table 1. Utilizing both VGG16 and ResNet50 as backbone networks, our method consistently demonstrates superior performance across various cross-validation folds. For the VGG16 backbone, our approach exhibits a notable advantage across all folds (Fold-0 to Fold-3), showcasing substantial improvements compared to other methods. The mIoU and FB-IoU metrics attain impressive average values of 61.8% and 73.1%, respectively. When using ResNet50 as the backbone network, our method performs exceptionally well across all folds, particularly achieving an outstanding 72.5% IoU on Fold-1, surpassing competing models. Specific metrics indicate that our method surpasses the latest SiGCN by 1.8% in mIoU and exceeds 0.7% in FB-IoU. Compared to the newest QCLNet, our method demonstrates competitive performance and slightly outperforms mIoU. Table 1 provides a comprehensive summary of

TABLE 4. Effect of different window values on a PASCAL-5^i dataset.

$\mathcal{W}$	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
$(3,3)$	65.4	71.6	67.5	60.2	66.2	77.3
$(5,5)$	66.2	**72.5**	**68.4**	61.3	**67.1**	**78.2**
$(7,7)$	**66.4**	71.8	67.9	60.8	66.7	77.6
$(9,9)$	65.8	71.6	68.2	**61.5**	66.8	77.8

these results, showcasing the outstanding performance of our method across different backbone networks and folds, further enhancing its effectiveness in challenging 1-shot scenarios.

2) COCO-20^i

The COCO-20^i dataset encompasses a diverse range of objects with significant variations, posing a more challenging segmentation task than PASCAL-5i. Despite these complexities, our method's performance in the 1-shot setting on the COCO-20^i dataset has been thoroughly evaluated against several state-of-the-art models, with the results presented in Table 2. When employing the VGG16 backbone network, our method exhibits competitive performance across all four cross-validation folds (Fold-0 to Fold-3). The mIoU and FB-IoU are particularly noteworthy, where our method significantly outperforms other cutting-edge models. This underscores the robustness and effectiveness of our approach in addressing the challenges posed by the diverse nature of the COCO-20^i dataset.Our method has achieved remarkable results under the scenario of employing ResNet50 as the backbone network. Specifically, in the 1-shot setting of the COCO-20^i dataset, we attained a performance of mIoU 43.4% and FB-IoU 64.5%. This performance represents a notable improvement compared to the latest SiGCN and IMR-HSNet, with enhancements of 1.7% and 0.7%, respectively, in terms of mIoU. Table 2 provides a comprehensive overview of the performance comparison, illustrating the superiority of our method in achieving accurate and robust segmentation results under the challenging 1-shot setting on the COCO-20^i dataset. These findings contribute to positioning our approach as a leading solution for segmentation tasks in datasets with diverse and complex object categories.

E. ABLATION STUDIES

To assess the efficacy of our proposed main modules–FIM (Feature Integration Module), FEM (Feature Enhancement Module), and FRM (Feature Refinement Module)–we conducted a comprehensive ablation study on the PASCAL-5^i and COCO-20^i dataset, using OPCN as our experimental framework. The ablation study involved removing these modules individually and in combinations to analyze their impact on model performance, and the results are summarized in Table 3. In terms of single module introduction, FIM achieved the most significant improvement. Compared to the baseline model, FIM increased the average mIoU by 2.6% and FB-IoU by 3.4% on PASCAL-5^i. On COCO-20^i, FIM increased the average mIoU by 1.3% and FB-IoU by 1.5%. In contrast, FEM showed slightly inferior performance when introduced alone. Introducing FIM and FRM modules in configuration (f) significantly improved the model's performance, with an increase of 5.2% in average mIoU and 5.4% in FB-IoU on PASCAL-5^i. On COCO-20^i, the average mIoU increased by 3.2%, and FB-IoU increased by 3.3%. This indicates that these modules have significantly enhanced feature integration and refinement, demonstrating clear superiority compared to the baseline without the main modules. In configuration (e), merging the FEM and FRM modules without using the FIM module resulted in a slight performance improvement. Similarly, in configuration (g) using both FIM and FEM modules, the model's performance was better than (e), achieving an average mIoU of 65.2% on PASCAL-5^i and 41.6% on COCO-20^i. However, the best model configuration was completed in (h) when all three main modules (FIM, FEM, and FRM) were used. The model excelled in all aspects of this scenario, with an average mIoU of 67.1% and FB-IoU of 78.2% on PASCAL-5^i, and an average mIoU of 43.4% and FB-IoU of 64.5% on COCO-20^i. Compared to the baseline, this represents a significant improvement of 5.8% and 7.0% in average mIoU and 4.1% and 4.7% in FB-IoU, respectively.

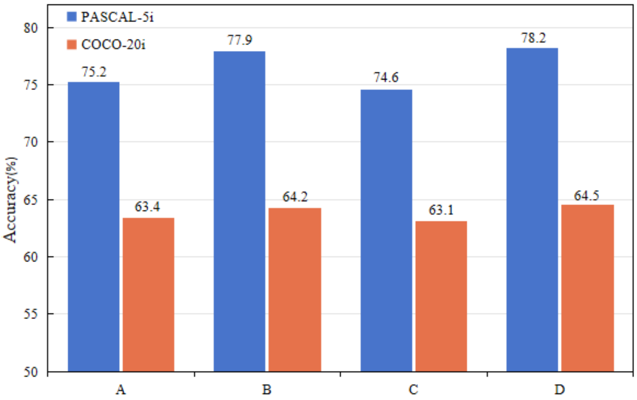


FIGURE 5. Impact of adding contingent substitution cross-attention on network performance.

To comprehensively assess the influence of sliding window size within the Feature Enhancement Module (FEM) on

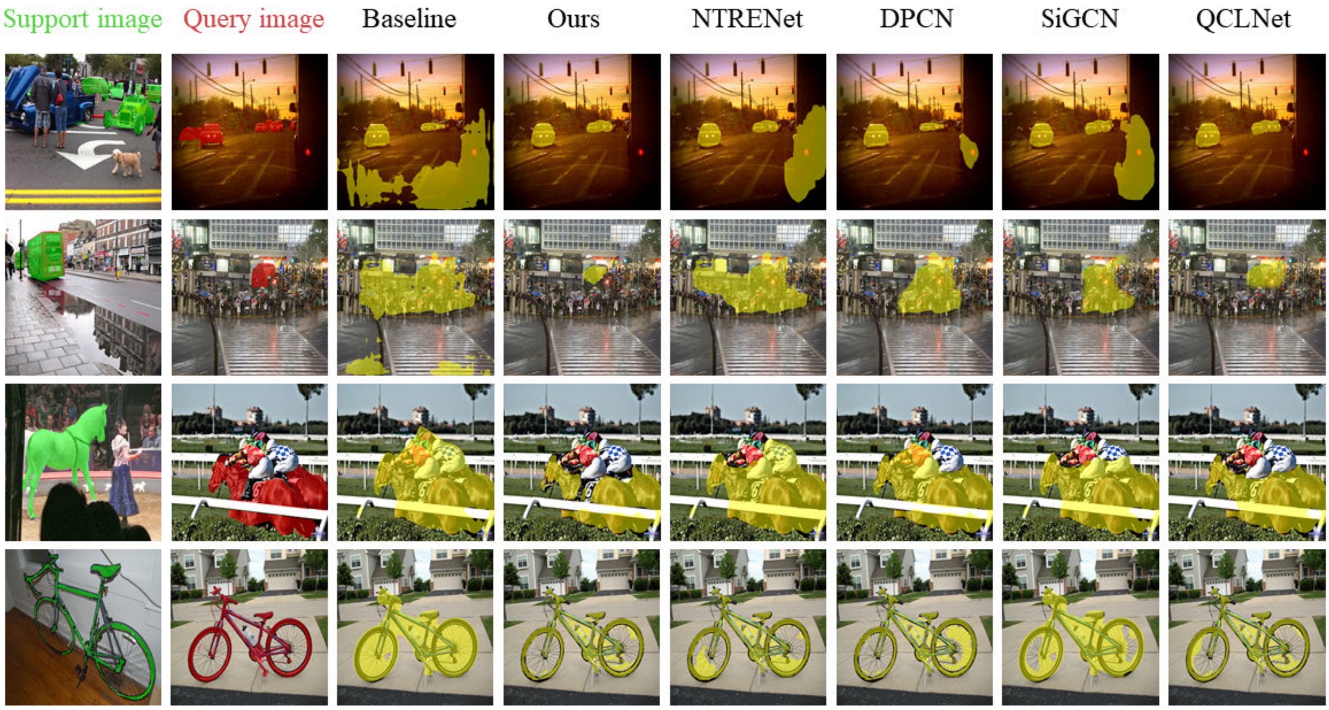


FIGURE 6. Results of our approach OPCN and other model on PASCAL-5^i and COCO-20^i datasets.

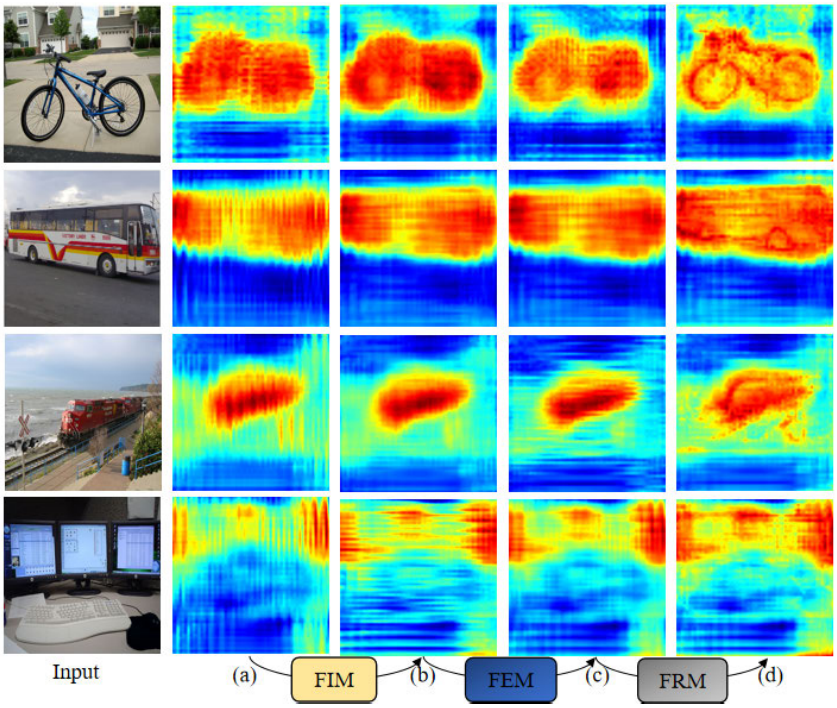


FIGURE 7. Visualization of feature map weights for different layers in OPCN (a)Backbone network(ResNet-50) Output(b)Feature map after FIM (c)Feature map after FEM (d)Feature map after FRM.

network performance, we conducted experiments using varying window sizes, specifically (3, 3), (5, 5), (7, 7), and (9, 9), on the PASCAL-5^i dataset. The results are summarized in Table 4. The network's performance exhibits variations across different window sizes, demonstrating the model's sensitivity to the spatial context captured by the sliding window. Notably, the best performance is achieved on Fold-0 and Fold-3 when using window sizes of $(7, 7)$ and $(9, 9)$. While there is some variability in performance across folds, the overall optimal configuration for OPCN is found with a window size of $(5, 5)$. This configuration yields the highest mean mIoU of 67.1% and FB-IoU of 78.2%.

Additionally, we investigated the impact of Cross-Attention on our model by designing four sets of contrasting experiments. Given that 'Cross-Attention' operates as a co-attention mechanism, we introduced it before FIM (A), after FIM (B), replaced FIM (C), and kept the original method unchanged (D). The results are illustrated in Fig 5; Group D achieved the best results on both PASCAL-5i and COCO-20i datasets. From Groups A, B, and C, we observed varying degrees of performance decline when cross-attention was introduced either before or after FIM or when replacing the FIM module. This occurrence may be attributed to the inherently better compatibility of cross-attention with Transformer architectures. Additionally, the FIM module proved more effective in extracting co-attention for One-Shot Segmentation.

F. QUALITATIVE RESULTS

In Fig 6, we present the segmentation results of our proposed OPCN method compared to the other model on the PASCAL-5^i and COCO-20^i datasets. OPCN demonstrates superior performance in accurately segmenting target objects within complex environments. Specifically, in the second row of Figure 6, our proposed OPCN method accurately segments cars in a crowd. In contrast, other methods incorrectly include irrelevant background elements, such as crowds and traffic lights, while segmenting cars. In the third row of Figure 6, OPCN can accurately segment horses and people, whereas the latest QCLNet incorrectly includes people's feet while segmenting horses. In addition,in the first and fourth rows of Figure 6, the OPCN can preserve finer segmentation details, especially for the more complex information of bicycles.

In addition, we conducted a visualization of the feature map weights within OPCN across various modules. As depicted in the second column of Fig 7, the feature map weights derived from the backbone network appear relatively dispersed. Following enhancement by the Fusion Interactive Module (FIM) and Feature Enhancement Module (FEM), the feature map weights progressively intensify, becoming more concentrated. Ultimately, through the Feature Refinement Module (FRM) filtering process, the feature map weights become effectively localized within the target area.

V. CONCLUSION

We propose a One-Shot Segmentation Prototype Comparison Network (OPCN) with four main components (FIM, FEM, FRM, and FM) to address the challenges in the FSS task. We use the FIM module to capture standard features between support and query set features to achieve the entire interaction between support and query features. In addition, we specify the target location details in the features through FEM and FRM and finally generate the final segmentation mask through FM. Through extensive experimentation on the PASCAL-5^i and COCO-20^i datasets, our OPCN demonstrates outstanding performance in the single-shot setting, effectively addressing the challenges inherent in Few-Shot Segmentation tasks. However, when facing scenes with K-shot data characteristics, our method is the same as the currently commonly adopted method: average the extracted prototypes. However, this method assumes that the distribution of each sample is different, which may not produce optimal results because images from other scenes cannot provide targeted guidance, so the model will underperform in the k-shot setting. Our future efforts will focus on generalizing our approach to any k-shot (where $k \geq 1$) few-shot segmentation task.

ACKNOWLEDGMENT

(Lingbo Li and Zhichun Li are co-first authors.)

REFERENCES

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[2] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[3] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning," in *Proc. BMVC*, 2018, vol. 3, no. 4, pp. 1–13.

[4] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015.

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[6] Q. Ye, B. Yuchen Lin, and X. Ren, "CrossFit: A few-shot learning challenge for cross-task generalization in NLP," 2021, *arXiv:2104.08835*.

[7] Y. Ge, Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker, "Few-shot learning for medical text: A review of advances, trends, and opportunities," *J. Biomed. Informat.*, vol. 144, Aug. 2023, Art. no. 104458.

[8] Y. Guo, R. Du, Y. Dong, T. Hospedales, Y.-Z. Song, and Z. Ma, "Task-aware adaptive learning for cross-domain few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1590–1599.

[9] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9197–9206.

[10] G.-S. Xie, H. Xiong, J. Liu, Y. Yao, and L. Shao, "Few-shot semantic segmentation with cyclic memory network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7293–7302.

[11] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," Tech. Rep., 2018.

[12] P. Tian, Z. Wu, L. Qi, L. Wang, Y. Shi, and Y. Gao, "Differentiable meta-learning model for few-shot semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12087–12094.

[13] W. Wang, L. Duan, Y. Wang, Q. En, J. Fan, and Z. Zhang, "Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7065–7074.

[14] Y.-H. Lee, F.-E. Yang, and Y. F. Wang, "A pixel-level meta-learner for weakly supervised few-shot semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1607–1617.

[15] M. Zhang, M. Shi, and L. Li, "MFNet: Multiclass few-shot segmentation network with pixel-wise metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8586–8598, Dec. 2022.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, Munich, Germany. Springer, 2015, pp. 234–241.

[17] 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.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[19] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

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

[21] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[24] A. Fateh, M. Fateh, and V. Abolghasemi, "Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning," *Inf. Sci.*, vol. 581, pp. 479–494, Dec. 2021.

[25] A. Fateh, M. Fateh, and V. Abolghasemi, "Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection," *Eng. Rep.*, Dec. 2023, Art. no. e12832.

[26] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[27] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.

[28] T. Zhou, W. Wang, E. Konukoglu, and L. Van Goo, "Rethinking semantic segmentation: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2572–2583.

[29] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 1140–1156.

[30] M. Zhang, Y. Zhou, B. Liu, J. Zhao, R. Yao, Z. Shao, and H. Zhu, "Weakly supervised few-shot semantic segmentation via pseudo mask enhancement and meta learning," *IEEE Trans. Multimedia*, vol. 25, pp. 7980–7991, 2022.

[31] K. Huang, F. Wang, Y. Xi, and Y. Gao, "Prototypical kernel learning and open-set foreground perception for generalized few-shot semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19256–19265.

[32] Y. Liu, N. Liu, X. Yao, and J. Han, "Intermediate prototype mining transformer for few-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 38020–38031.

[33] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5212–5221.

[34] 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.

[35] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Computer Vision—ECCV*. Glasgow, U.K. Springer, 2020, pp. 763–778.

[36] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022.

[37] J. Liu, Y. Bao, G.-S. Xie, H. Xiong, J.-J. Sonke, and E. Gavves, "Dynamic prototype convolution network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11553–11562.

[38] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[39] Z. Wu, X. Shi, G. Lin, and J. Cai, "Learning meta-class memory for few-shot semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 517–526.

[40] G.-S. Xie, J. Liu, H. Xiong, and L. Shao, "Scale-aware graph neural network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5475–5484.

[41] Y. Su, P. Yan, J. Lin, C. Wen, and Y. Fan, "Few-shot defect recognition for the multi-domain industry via attention embedding and fine-grained feature enhancement," *Knowl.-Based Syst.*, vol. 284, Jan. 2024, Art. no. 111265.

[42] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," 2017, *arXiv:1709.03410*.

[43] M. Siam, B. Oreshkin, and M. Jagersand, "AMP: Adaptive masked proxies for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5248–5257.

[44] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11573–11582.

[45] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 6941–6952.

[46] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8334–8343.

[47] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13979–13988.

[48] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8312–8321.

[49] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "Mining latent classes for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8701–8710.

[50] J. Liu, Y. Bao, W. Yin, H. Wang, Y. Gao, J.-J. Sonke, and E. Gavves, "Few-shot semantic segmentation with support-induced graph convolutional network," 2023, *arXiv:2301.03194*.

[51] Z. Zheng, G. Huang, X. Yuan, C.-M. Pun, H. Liu, and W.-K. Ling, "Quaternion-valued correlation learning for few-shot semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2102–2115, May 2023.

[52] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 622–631.

[53] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Computer Vision—ECCV*, Glasgow, U.K. Springer, 2020, pp. 142–158.

[54] H. Wang, L. Liu, W. Zhang, J. Zhang, Z. Gan, Y. Wang, C. Wang, and H. Wang, "Iterative few-shot semantic segmentation from image label text," 2023, *arXiv:2303.05646*.

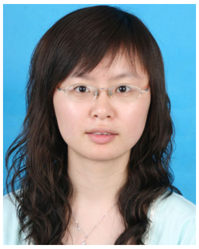
[55] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[56] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in*Computer Vision—ECCV*, Zurich, Switzerland. Springer, 2014, pp. 740–755.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.



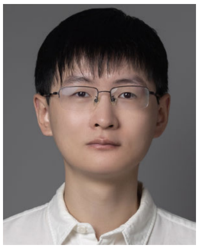
LINGBO LI received the M.S. degree in software engineering from Zhejiang University, in 2007. Since 2007, she has been with the Information Center, Zhejiang Technical Institute of Economics. She has been committed to the school's information construction for many years. Her research interests include information system development, computer vision, and machine learning. She passed the Computer Technology and Software Professional Qualification (Level) Examination and obtained the Senior Engineer Qualification Certificate, in 2023.

ZHICHUN LI received the B.S. degree in computer science and technology from Hong Kong Baptist University, in 2022. He is currently pursuing the M.Phil. degree with the Department of Health Technology and Informatics, The Hong Kong Polytechnic University. His research interests include deep learning applications, image-guided radiation therapy, and medical image processing.

FUSEN GUO (Graduate Student Member, IEEE) received the Bachelor of Information Technology degree in business information systems from Monash University, Australia, in 2022, and the degree (Hons.) in information technology from Deakin University, Australia, in 2023. He is currently pursuing the Ph.D. degree in computer science with the Swinburne University of Technology, Australia. His research interests include power grids, cyber-physical systems, cyber-attacks, and explainable artificial intelligence.

HAOYU YANG received the B.S. degree from the School of Gifted Young, University of Science and Technology of China, in 2023. He is currently pursuing the M.S. degree with the College of Computing, Georgia Institute of Technology. He has cooperated with researcher Feilong Ma and Associate Professor Weiwei Zhuang. His research interests include digital image processing, computer vision, and autonomous driving.

JINGTIAN WEI received the Bachelor of Engineering degree (Hons.) in computer engineering from the University of New South Wales (UNSW), in 2023, where he is currently pursuing the M.Phil. degree with the Data and Knowledge Research Group (DKR), School of Computer Science and Engineering, under the supervision of Dr. Zhengyi Yang and Prof. Wenjie Zhang. His research interest includes graph mining, with a current focus on algorithms for hypergraphs.



ZHENGYI YANG received the Ph.D. degree in computer science from the School of Computer Science and Engineering, University of New South Wales (UNSW). He is currently an Associate Lecturer and a Ph.D. Supervisor with the School of Computer Science and Engineering, UNSW. His expertise lies in developing efficient and scalable algorithms and systems for managing and processing large-scale data, including graph, relational, spatial–temporal, and high-dimensional data.