Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

Jie Liu¹*, Yanqi Bao²*, Guo-Sen Xie³,⁴†, Huan Xiong⁴, Jan-Jakob Sonke⁵, Efstratios Gavves¹
¹University of Amsterdam, Netherlands  ²Northeastern University, China  ³The Netherlands Cancer Institute, Netherlands
⁴Nanjing University of Science and Technology, China  ⁵Mohamed bin Zayed University of Artificial Intelligence, UAE

Abstract

The key challenge for few-shot semantic segmentation (FSS) is how to tailor a desirable interaction among support and query features and/or their prototypes, under the episodic training scenario. Most existing FSS methods implement such support/query interactions by solely leveraging plain operations – e.g., cosine similarity and feature concatenation – for segmenting the query objects. However, these interaction approaches usually cannot well capture the intrinsic object details in the query images that are widely encountered in FSS, e.g., if the query object to be segmented has holes and slots, inaccurate segmentation almost always happens. To this end, we propose a dynamic prototype convolution network (DPCN) to fully capture the aforementioned intrinsic details for accurate FSS. Specifically, in DPCN, a dynamic convolution module (DCM) is firstly proposed to generate dynamic kernels from support foreground, then information interaction is achieved by convolution operations over query features using these kernels. Moreover, we equip DPCN with a support activation module (SAM) and a feature filtering module (FFM) to generate pseudo mask and filter out background information for the query images, respectively. SAM and FFM together can mine enriched context information from the query features. Our DPCN is also flexible and efficient under the k-shot FSS setting. Extensive experiments on PASCAL-5¹ and COCO-20⁵ show that DPCN yields superior performances under both 1-shot and 5-shot settings.

1. Introduction

Semantic segmentation has achieved tremendous success due to the advancement of deep convolutional neural networks [9, 10, 22]. Nevertheless, most leading image semantic segmentation models rely on large amounts of training images with pixel-wise annotations, which require huge human efforts. Semi- and weakly-supervised segmentation methods [15, 24, 28] are proposed to alleviate such expensive annotation cost. However, both semi- and weakly-supervised methods have to face a significant performance drop when only a few annotated samples for a novel object are available. In such a case, few-shot semantic segmentation (FSS) [20] is introduced to allow for dense pixel-wise prediction on novel object categories given only a few annotated samples.

Typically, most FSS methods adopt an episode-based meta-learning strategy [30], and each episode is composed of the support set and the query set, which share the same object class. The support set contains a few support images with pixel-wise annotation. FSS models are expected to learn to predict the segmentation mask for images in the query set with the guidance of the support set. Learning is based on episodes available with annotations during training. At test time, the model is expected to segment a query
image with respect to a class of interest, again provided with corresponding support set, only this time the class of interest in the query and support set is novel and not previously seen.

Currently, the most leading FSS methods are the prototype-based ones [12, 25]. As in Fig. 1(a), prototype-based paradigm typically generates multiple foreground and/or background prototypes by utilizing mask average pooling and/or clustering over support features. These prototypes are supposed to contain representative information of the target object in the support images, thus their interactions with query features by cosine similarity, element-wise summation, and feature concatenation can produce necessary predictions for the objects in the query image. However, the predictions achieved by solely relying on such limited prototypes and plain operations are inevitably losing some intrinsic object details in the query image. For instance, as in Fig. 1(a), since the objects plants have holes and slots which are intrinsic details, the segmented objects cannot well cover these details, i.e., a defective over-segmentation is achieved in this case. In addition, in the presence of large object variations (e.g., appearance and scale) in FSS, it is usually difficult to comprehensively encode the adequate patterns of the target objects by solely considering the support information as in most previous prototype-based methods.

To address the above challenges, we propose a dynamic prototype convolution network (DPCN) to fully capture the intrinsic object details for accurate FSS. DPCN belongs to prototype-based methods yet with several elegant extensions and merits. Specifically, we first propose a dynamic convolution module (DCM) to achieve more adequate interaction between support and query features, thus leading to more accurate prediction for the query objects. As in Fig. 1(b), we leverage three dynamic kernels, i.e., a square kernel and two asymmetric kernels, generated from the support foreground features. Then three convolution operations are employed in parallel onto the query features using these dynamic kernels. This interaction strategy is simple yet important to comprehensively tackle large object variations (e.g., appearance and scales) and can capture the intrinsic object details. Intuitively, the square kernel is capable of capturing the main information of an object (e.g., main body of the plant in Fig. 1(b)); By contrast, asymmetric kernels (i.e., kernel with size \( d \times 1 \) or \( 1 \times d \)) aim to capture the subtle object details, e.g., leaves in Fig. 1(b). As such, DPCN equipped with DCM can better handle the intrinsic object details using an extremely simple way.

Moreover, to comprehensively encode the adequate patterns of the target objects, we propose a support activation module (SAM) and a feature filtering module (FFM) to mine as much object-related context information from query image as possible. Specifically, SAM generates support activation maps and initial pseudo query mask using high-level support and query features. Then the support prototypes and pseudo query foreground features are fused to generate a refined pseudo mask for the query image in FFM. Compared with the original pseudo query mask, the refined one contains more object foreground context while filtering some noise information. Therefore, rich object-related context information from both support and query images are aggregated to the final feature, leading to better segmentation performance. Our main contributions are as follows:

- We propose a dynamic prototype convolution network (DPCN) to capture the intrinsic object details for accurate FSS. To the best of knowledge, we are the first one to do this in the FSS domain.
- We propose a novel dynamic convolution module (DCM) to achieve adequate support-query interactions. DCM can serve as a plug-and-play component to improve existing prototype learning methods.
- We propose a support activation module (SAM) and a feature filtering module (FFM) to mine complementary information of target objects from query images.

2. Related work

2.1. Semantic Segmentation

Semantic segmentation is a classical computer vision task which aims to give pixel-wise prediction for an input image. Recently, various networks [14] have been actively designed to further improve the semantic segmentation results. For capturing more contextual informations, dilated convolution [33], pyramid pooling [40], and deformable convolution [3], are proposed to enlarge the receptive filed. Meanwhile, some models leverage attention mechanisms [6, 27, 34, 37] to capture long-distance dependencies for semantic segmentation, which reach state-of-the-art performances. However, these semantic segmentation approaches still fail to preserve their initial performances when insufficient training data is provided.

2.2. Few-Shot Semantic Segmentation

Few-shot semantic segmentation (FSS) learns to segment target objects in query image given few pixel-wise annotated support images. Most existing FSS methods adopt two-branch architecture which implements meta-training on the base classes and then conducts meta-testing on the disjoint novel classes. OSLSM [20] is the first two-branch FSS model. Next, PL [4] introduces prototype learning paradigm, which generates prototypes from support images to guide the segmentation of query objects. Recently, many prototype-based FSS methods emerge in the research community, such as CANet [36], SG-One [39], PANet [26], PMMs [31], PFENet [25], and ASGNet [12]. The key idea
of these methods lies in generating or rearranging representative prototypes using different strategies, then the interaction between prototypes with query features can be formulated as a few-to-many matching problem. However, these prototype learning methods inevitably cause information loss due to limited prototypes. Therefore, graph-based methods have thrived recently as they try to preserve structural information with many-to-many matching mechanism. For instance, PGNet [36] applies attentive graph reasoning to propagate label information from support data to query data. SAGNN [30] constructs graph nodes using multi-scale features and performs k-step reasoning over nodes to capture cross-scale information. Most recently, HSNet [17] proposes to tackle the FSS task from the perspective of visual correspondence. It implements efficient 4D convolutions over multi-level feature correlation and achieves great success. Different from previous methods, we try to perform sufficient interaction between support and query features using dynamic convolution, and mine as much complementary target information from both support and query features.

2.3. Dynamic Convolution Networks

Dynamic convolution networks aim to generate diverse kernels and implement convolution over input feature with these kernels. Many previous works have explored the effectiveness of dynamic convolution in deep neural networks. DFN [11] proposes a dynamic filter network where filters are generated dynamically conditioned on input and achieves state-of-the-art performance on video and stereo prediction task. [2] aggregates multiple parallel convolution kernels dynamically based upon their attentions, and it boosts both image classification and keypoint detection accuracy. Dynamic convolution is also used in DMNet [8] to adaptively capture multi-scale contents for predicting pixel-level semantic labels. The core of these methods is constructing multiple kernels from input features. Most recently, dynamic convolution is introduced into the few-shot object detection task by [38], which generates various kernels from the object regions in support image and then implements convolution over query feature using these kernels, leading to a more representative query feature. In this paper, we propose to generate dynamic kernels from foreground support feature to interact with query feature by convolution. Instead of only using square kernels as in [38], we also introduce asymmetric kernels to capture subtle object details. Experiments in §4.3 demonstrate well the effectiveness of our method.

3. Method

3.1. Problem Setting

We adopt the standard FSS setting, i.e., following the episode-based meta-learning paradigm [23]. We start from classes \(C_{tr}\) and \(C_{ts}\) for the training set \(D_{tr}\) and the test set \(D_{ts}\), respectively. The key difference between FSS and general semantic segmentation task is that \(C_{tr}\) and \(C_{ts}\) in FSS are disjoint, \(C_{tr} \cap C_{ts} = \emptyset\). Both \(D_{tr}\) and \(D_{ts}\) consist of thousands of randomly sampled episodes, and each episode \((S, Q)\) includes a support set \(S\), and a query set \(Q\)
3.3 Support Activation Module

Inspired by PFENet [16, 25], recent FSS models [29, 30] usually leverage high-level features (e.g., conv5 of ResNet50) from the support and query image to generate the prior mask indicating the rough location of the target object in the query image. As this prior mask is usually obtained by element-to-element or square region-based matching between feature maps, a holistic context is not taken into account.

To counter this, with the support activation module we generate multiple activation maps of the target object in the query image using holistic region-to-region matching. Specifically, as in Fig. 3, SAM takes as input the high-level support feature \( x^h_s \in \mathbb{R}^{C_s \times H_s \times W_s} \), the corresponding binary mask \( M_s \in \mathbb{R}^{H_s \times W_s} \), as well as the high-level query feature \( x^h_q \in \mathbb{R}^{C_q \times H_q \times W_q} \), where \( C_q \) is the channel dimension, \( H_s, W_s, H_q, W_q \) are the height and width of support and query feature, respectively.

To perform holistic matching, we first need to generate region features \( R_s \) and \( R_q \) with a fixed window operation \( W \) sliding on the support and query features, respectively.

\[
R_s = W(x^h_s \odot M_s) \in \mathbb{R}^{d_h \cdot d_w \times C_h \times H_s \times W_s},
\]

\[
R_q = W(x^h_q) \in \mathbb{R}^{d_h \cdot d_w \times C_h \times H_q \times W_q},
\]

(1)

where \( \odot \) stands for the Hadamard product and \( d_h, d_w \) are the window height and width. In our experiments we opt for symmetrical and asymmetrical windows, i.e., \((d_h, d_w) \in \{(5, 1), (3, 3), (1, 5)\}\) that are comprehensive and holistic regions, to account for possible object geometry variances. Having the region features, we proceed with matching by computing their cosine similarity, which generates the regional matching map \( \mathrm{Corr} \in \mathbb{R}^{d_h \cdot d_w \times C_h \times H_s \times W_s \times H_q \times W_q} \). Notably, we utilize both square window (3,3) and asymmetrical windows (i.e., (5,1) and (1,5)), where square window can introduce more contextual information on regular part of target objects like the main body of object \( \textit{plants} \), asymmetrical windows can incorporate contextual details of slender part (e.g., leaves of \( \textit{plants} \)).

We generate the final activation map \( M_{\text{act}} \in \mathbb{R}^{H_q \times W_q} \) by taking the mean value among all regions and the maximal value among all support features followed by normalization operation. As we have three windows, we have three activation maps, \( \{M^{(1)}_{\text{act}}\}_{1=1}^3 \). In the end, we obtain the initial pseudo-mask \( M^{(0)}_{\text{pse}} \in \mathbb{R}^{H_q \times W_q} \), which indicates the rough location of target objects, by a mean operation.

3.4 Feature Filtering Module

As in Fig. 2, the feature filtering module is constructed on mid-level support and query features, i.e., \( x_s \in \mathbb{R}^{C \times H \times W} \) and \( x_q \in \mathbb{R}^{C \times H \times W} \) where \( C, H, W \) are channel, height, and width, respectively. Given \( x_s, x_q \), and the initial pseudo mask \( M^{(0)}_{\text{pse}} \), the feature filtering module refines the pseudo mask, which is used to filter out irrelevant background information in the query image. We first apply masked average pooling on the features from the support set to get prototype vector \( p \in \mathbb{R}^{C \times 1 \times 1} \):

\[
p = \text{average}(x_s \odot R(M_s)),
\]

(2)

where \( R \) reshapes support mask \( M_s \) to be the same shape as \( x_s \). Then, we expand the support prototype vector \( p \) to match the dimensions of the feature maps, \( x_p \in \mathbb{R}^{C \times H \times W} \),

\[
\text{Mean}(0) & \text{Max}(1)
\]

\[
\text{Norm}
\]

\[
\text{For a specific class } c. \text{ For the } k\text{-shot setting, the support set that contains } k \text{ image-mask pairs can be formulated as } S = \{(I^s_i, M^s_i)\}^{k}_{i=1}, \text{ where } I^s_i \text{ represents ith support image and } M^s_i \text{ indicates corresponding binary mask. Similarly, we define the query set as } Q = \{(I^q_i, M^q_i)\}, \text{ where } I^q_i \text{ is the query image and its binary mask } M^q_i \text{ is only available in the model training phase. In the meta-training stage, the FSS model takes as input } S \text{ and } I^q_i \text{ from a specific class } c \text{ and generates a predicted mask } M^q_i \text{ for the query image. Then the model can be trained with the supervision of a binary cross-entropy loss between } M^q_i \text{ and } M^q_i, \text{ where } M^q_i \text{ is only available } C

\[
\text{Corr}
\]

\[
\text{Reshape}
\]

\[
\text{Norm}
\]

\[
\text{Mean}(0) & \text{Max}(1)
\]

\[
\text{For a specific class } c. \text{ For the } k\text{-shot setting, the support set that contains } k \text{ image-mask pairs can be formulated as } S = \{(I^s_i, M^s_i)\}^{k}_{i=1}, \text{ where } I^s_i \text{ represents ith support image and } M^s_i \text{ indicates corresponding binary mask. Similarly, we define the query set as } Q = \{(I^q_i, M^q_i)\}, \text{ where } I^q_i \text{ is the query image and its binary mask } M^q_i \text{ is only available in the model training phase. In the meta-training stage, the FSS model takes as input } S \text{ and } I^q_i \text{ from a specific class } c \text{ and generates a predicted mask } M^q_i \text{ for the query image. Then the model can be trained with the supervision of a binary cross-entropy loss between } M^q_i \text{ and } M^q_i, \text{ where } M^q_i \text{ is only available } C

\[
\text{Corr}
\]

\[
\text{Reshape}
\]

\[
\text{Norm}
\]

\[
\text{Mean}(0) & \text{Max}(1)
\]

\[
\text{For a specific class } c. \text{ For the } k\text{-shot setting, the support set that contains } k \text{ image-mask pairs can be formulated as } S = \{(I^s_i, M^s_i)\}^{k}_{i=1}, \text{ where } I^s_i \text{ represents ith support image and } M^s_i \text{ indicates corresponding binary mask. Similarly, we define the query set as } Q = \{(I^q_i, M^q_i)\}, \text{ where } I^q_i \text{ is the query image and its binary mask } M^q_i \text{ is only available in the model training phase. In the meta-training stage, the FSS model takes as input } S \text{ and } I^q_i \text{ from a specific class } c \text{ and generates a predicted mask } M^q_i \text{ for the query image. Then the model can be trained with the supervision of a binary cross-entropy loss between } M^q_i \text{ and } M^q_i, \text{ where } M^q_i \text{ is only available } C

\[
\text{Corr}
\]

\[
\text{Reshape}
\]

\[
\text{Norm}
\]

\[
\text{Mean}(0) & \text{Max}(1)
\]

\[
\text{For a specific class } c. \text{ For the } k\text{-shot setting, the support set that contains } k \text{ image-mask pairs can be formulated as } S = \{(I^s_i, M^s_i)\}^{k}_{i=1}, \text{ where } I^s_i \text{ represents ith support image and } M^s_i \text{ indicates corresponding binary mask. Similarly, we define the query set as } Q = \{(I^q_i, M^q_i)\}, \text{ where } I^q_i \text{ is the query image and its binary mask } M^q_i \text{ is only available in the model training phase. In the meta-training stage, the FSS model takes as input } S \text{ and } I^q_i \text{ from a specific class } c \text{ and generates a predicted mask } M^q_i \text{ for the query image. Then the model can be trained with the supervision of a binary cross-entropy loss between } M^q_i \text{ and } M^q_i, \text{ where } M^q_i \text{ is only available } C

\[
\text{Corr}
\]

\[
\text{Reshape}
\]

\[
\text{Norm}
\]

\[
\text{Mean}(0) & \text{Max}(1)
\]

\[
\text{For a specific class } c. \text{ For the } k\text{-shot setting, the support set that contains } k \text{ image-mask pairs can be formulated as } S = \{(I^s_i, M^s_i)\}^{k}_{i=1}, \text{ where } I^s_i \text{ represents ith support image and } M^s_i \text{ indicates corresponding binary mask. Similarly, we define the query set as } Q = \{(I^q_i, M^q_i)\}, \text{ where } I^q_i \text{ is the query image and its binary mask } M^q_i \text{ is only available in the model training phase. In the meta-training stage, the FSS model takes as input } S \text{ and } I^q_i \text{ from a specific class } c \text{ and generates a predicted mask } M^q_i \text{ for the query image. Then the model can be trained with the supervision of a binary cross-entropy loss between } M^q_i \text{ and } M^q_i, \text{ where } M^q_i \text{ is only available } C

\[
\text{Corr}
\]

\[
\text{Reshape}
\]

\[
\text{Norm}
\]

\[
\text{Mean}(0) & \text{Max}(1)
\]
and combine the target object information from both the support and query features. We refine the pseudo mask with the help of a smaller network $F$ composed of a 2D convolution layer followed by a sigmoid function,

$$M^r_{\text{pse}} = F((x_q \otimes R(M^0_{\text{pse}})) \oplus x_p) \in \mathbb{R}^{H \times W},$$

where $\oplus$ stands for the element-wise sum. Compared with $M^0_{\text{pse}}$, $M^r_{\text{pse}}$ gives more accurate estimation of the object location in the query image. Lastly, we obtain the final filtered query feature that discards irrelevant background by combining the feature $x_q$ with the prior mask:

$$\tilde{x}_q = (x_q \otimes M^r_{\text{pse}}) \oplus x_q \in \mathbb{R}^{C \times H \times W}.$$ (4)

### 3.5. Dynamic Convolution Module

In the previous step we obtain a foreground feature from the query, which is minimally affected by irrelevant background. Still, the operations so far have been so that to provide a rough estimate of the location of the target object. For accurate segmentation, however, much finer pixel-level predictions are required. In the absence of significant data to train our filters on, we introduce dynamic convolutions. We illustrate DCM in Fig. 2, and Fig. 4 depicts the details of the kernel generator.

Dynamic convolutions rely on meta-learning to infer what are the optimal kernel parameters given a subset of features, agnostic to the unknown underlying class semantics. Specifically, we input the mid-level support feature $x_s$ and the corresponding mask $M_s$ to a kernel generator, which generates dynamic kernels, i.e., one group square kernel and two groups of asymmetrical kernels. Then, we carry out three convolution operations over the filtered query feature $\tilde{x}_q$ using dynamic kernels. Firstly, we extract foreground vectors $P_{fg}$ from support feature:

$$P_{fg} = F_c(x_s \otimes M_s) \in \mathbb{R}^{N_{fg} \times C},$$

where $F_c$ is the foreground extraction function without any learnable parameters, $N_{fg}$ represents the number of foreground vectors. Next, two consecutive 1D pooling operations with kernel size $S$ and $S^2$ are leveraged to obtain two groups of prototypes $p_s \in \mathbb{R}^{S \times C}$ and $p_{s2} \in \mathbb{R}^{S^2 \times C}$:

$$p_s = \text{pool}_s(P_{fg}), p_{s2} = \text{pool}_{s2}(p_s).$$ (6)

As discussed above, we achieve dynamic convolution over query feature using a square kernel and two asymmetric kernels. As such, we use three parallel convolutional neural networks whose outputs are the generated kernel weights:

$$ker_v = F_{\text{conv1}}(p_s) \in \mathbb{R}^{S \times 1 \times C},$$
$$ker_h = F_{\text{conv2}}(p_s) \in \mathbb{R}^{1 \times S \times C},$$
$$ker_s = F_{\text{conv3}}(p_{s2}) \in \mathbb{R}^{S^2 \times S \times C},$$

where $ker_v$, $ker_h$, $ker_s$ are the vertical, horizontal, and square kernel weights, respectively. $F_{\text{conv1}}$, $F_{\text{conv2}}$, and $F_{\text{conv3}}$ represent corresponding convolution sub networks, which are achieved by two consecutive 1D convolution layers. We emphasize that the above parameter generating networks do not share parameters. Given the vertical kernel $ker_v$, the query feature $\tilde{x}_q$ can be enhanced as $\tilde{x}^v_q \in \mathbb{R}^{C \times H \times W}$:

$$\tilde{x}^v_q = F_{dc}(\tilde{x}_q|ker_v),$$ (8)

where $F_{dc}$ denotes the dynamic convolution operation, and $ker_v$ works as the kernel weight. Similarly, we can obtain other enhanced query features $\tilde{x}^h_q \in \mathbb{R}^{C \times H \times W}$ and $\tilde{x}^s_q \in \mathbb{R}^{C \times H \times W}$ with horizontal kernel $ker_h$ and square kernel $ker_s$, respectively. With the sufficient interaction between query feature and dynamic support kernels, the object context in the generated query feature are enhanced.

Then, the enhanced query features $\tilde{x}^v_q, \tilde{x}^h_q, \tilde{x}^s_q$, support foreground feature $x_p$, support activation maps $\{M^i_{\text{act}}\}_{i=1}^3$, and refined pseudo mask $M^r_{\text{pse}}$ are all reshaped to the same spatial size and concatenated to a representative feature $x_{out} \in \mathbb{R}^{(4C+4) \times H \times W}$:

$$x_{out} = F_{\text{cat}}(\tilde{x}^v_q, \tilde{x}^h_q, \tilde{x}^s_q, x_p, \{M^i_{\text{act}}\}_{i=1}^3, M^r_{\text{pse}}),$$ (9)

where $F_{\text{cat}}$ is the concatenation operation in channel dimension. Finally, $x_{out}$ is fed into a decoder to generate segmentation mask $M_q$ for query image $I_q$:

$$M_q = F_{\text{cls}}(F_{\text{ASPP}}(F_{\text{conv}}(x_{out}))),$$ (10)

where $F_{\text{conv}}$, $F_{\text{ASPP}}$, and $F_{\text{cls}}$ are three consecutive modules that constitute the decoder.

### 3.6. Extension to k-shot setting

So far we have focused on the one-shot setting, summarized in Fig. 2. For the $k$-shot setting, where more than one support images are available, most existing methods choose attention-based fusion or feature averaging. However, such a simple strategy does not make full use of the support information. By contrast, we can easily extend the dynamic convolutions to the $k$-shot setting and achieve substantial performance improvement. Specifically, given each support image-mask pair, we extract foreground vectors with Eq.
(5). By collecting all foreground vectors together, we get the overall support foreground vectors $P_{fg}$ from $k$ shots:

$$P_{fg} = (P_{fg}^1, P_{fg}^2, \ldots, P_{fg}^k) \in \mathbb{R}^{N_{fg} \times C},$$

where the number of foreground vectors is $N_{fg} = \sum_{i=1}^{k} N_{fg}^i$. By doing so, the kernel generator in DCM can generate more robust dynamic kernels, thus leading to more adequate interaction and accurate query mask estimation.

### 3.7. Training Loss

Our dynamic prototype convolution network is trained in an end-to-end manner with the binary cross-entropy loss (BCE). Given predicted mask $\hat{M}_q$ and ground-truth mask $M_q$ for query image $I_q$, we formulate the BCE loss between $\hat{M}_q$ and $M_q$ as our main loss:

$$L_{seg}^{q\rightarrow q} = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} BCE(\hat{M}_q(i,j), M_q(i,j)).$$

Inspired by [26], we implement another branch to estimate the support mask using query image $I_q$ and its corresponding predicted mask $\hat{M}_q$. Similar with Eq. (10), we get the predicted support mask $\hat{M}_s$. Then we get another loss by calculating BCE loss between $\hat{M}_s$ and $M_s$:

$$L_{seg}^{q\rightarrow s} = \frac{1}{hs \times ws} \sum_{i=1}^{hs} \sum_{j=1}^{ws} BCE(\hat{M}_s(i,j), M_s(i,j)).$$

where $h_s$ and $w_s$ are the height and width of ground-truth mask $M_s$ for support image $I_s$. Note that both query and support mask prediction process share the same structure and parameters. In summary, the final loss is:

$$L = L_{seg}^{q\rightarrow q} + \lambda L_{seg}^{q\rightarrow s},$$

where $\lambda$ is weight to balance the contribution of each branch and set to 1.0 in all experiments.

### 4. Experiments

#### 4.1. Experimental Settings

**Datasets.** We follow [25] and adopt PASCAL-5i [20] and COCO-20i [18] benchmarks for evaluation. PASCAL-5i comes from PASCAL VOC2012 [5] and additional SBD [7] annotations. It contains 20 object classes split into 4 folds, which are used for 4-fold cross validation. For each fold, 5 classes are used for testing and the remaining 15 classes for training. COCO-20i is a more challenging benchmark, which is created from MSCOCO [13] and contains 80 object classes. Similarly, we split classes in COCO-20i into 4 folds with 20 classes per fold, for each fold, we utilize 20 classes for testing and the remaining 60 classes for training.

**Metrics and Evaluation.** Following [17, 25, 30], mean intersection over union (mIoU) and foreground-background IoU (FB-IoU) are adopted as our metrics for evaluation. While FB-IoU neglects object classes and directly averages foreground and background IoU, mIoU averages IoU values of all classes in a fold. During evaluation, 1,000 episodes are sampled from the test set for metrics calculation, and multi-scale testing is used like most existing FSS methods.

**Implementation Details.** We use ResNet-50 [9] and VGG-16 [22] pre-trained on ImageNet as our backbone networks. The backbone weights are fixed except for layer4, which is required to learn more robust activation maps on PASCAL-5i, meanwhile, all these weights are fixed for COCO-20i to pursue a faster training time. The model is trained with the SGD optimizer on PASCAL-5i for 200 epochs and COCO-20i for 50 epochs. The learning rate is initialized as 0.005 with batch size 8 (0.002 with batch size 32) for PASCAL-5i (COCO-20i). Data augmentation strategies in [25] are adopted in the training stage, and all images are cropped to $473 \times 473$ patches for two benchmarks. In addition, the window sizes in SAM are set to $\{5 \times 1, 3 \times 3, 1 \times 5\}$ for PASCAL-5i; since COCO-20i contains much larger amounts of training images with plentiful types (already having holistic contexts compared with images in PASCAL-5i), we only utilize the initial prior mask for speeding up our model training. The kernel sizes in DCM are set as $5 \times 1, 5 \times 5$, and $1 \times 5$ for two datasets. We implement our model with PyTorch1.7 and conduct all the experiments with Nvidia Tesla V100 GPUs.

#### 4.2. Comparisons with State-of-the-Arts

**PASCAL-5i.** We report the mIoU and FB-IoU under both 1-shot and 5-shot settings in Table 1. It can be seen that (i) DPCN achieves state-of-the-art performance under both 1-shot and 5-shot settings. Especially for the 1-shot setting, we surpass HSNet [17], which holds the previous state-of-the-art results, by 2.0% and 2.7% with VGG16 and ResNet50 as backbone networks, respectively. In addition, DPCN also presents comparable performance with HSNet under 5-shot setting while using less mid-level features. (ii) DPCN outperforms its baseline method with a large margin (e.g., mIoU 66.7% versus 61.4% with ResNet50 backbone for 1-shot setting), which is implemented with same architecture except for proposed components (i.e., SAM, FFM, and DCM). The results further demonstrate that DPCN can effectively mine complementary information from both support and query features to facilitate query image segmentation.

**COCO-20i.** COCO-20i is a more challenging benchmark which usually contains multiple objects and exhibits great variance. Table 2 presents the performance comparison of mIoU and FB-IoU on COCO-20i dataset. As can be seen, using VGG16 and ResNet50 as backbone, our model
Table 1. Comparison with state-of-the-arts on COCO-20 dataset under both 1-shot and 5-shot settings. mIoU of each fold, and averaged mIoU & FB-IoU of all folds are reported. Baseline results are achieved by removing three modules (i.e., SAM, FFM, and DCM) in DPCN.

| Methods | Backbone | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
|---------|----------|--------|--------|--------|--------|------|--------|--------|--------|--------|--------|------|--------|
| Baseline | VGG16 | 33.6 | 55.3 | 40.9 | 33.5 | 40.8 | 61.3 | 35.9 | 58.1 | 42.7 | 39.1 | 43.9 | 61.5 |
| DPCN | VGG16 | 36.2 | 75.6 | 44.9 | 32.4 | 41.1 | 60.1 | 37.5 | 70.0 | 44.4 | 41.3 | 44.1 | 60.2 |
| ResNet50 | VGG16 | 47.7 | 50.2 | 46.7 | 34.7 | 43.4 | 61.9 | 40.0 | 55.3 | 49.9 | 40.1 | 46.4 | 62.1 |
| HSNNet (ICCV'21) [17] | VGG16 | 56.9 | 68.2 | 54.2 | 35.4 | 58.0 | 72.0 | 59.0 | 69.1 | 54.8 | 52.9 | 59.0 | 72.3 |
| PFNet (TPAMI'20) [25] | VGG16 | 59.6 | 65.7 | 59.6 | 54.0 | 59.2 | 73.4 | 64.9 | 69.0 | 64.1 | 58.6 | 64.1 | 76.6 |
| Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
| 1-shot mIoU | 65.5 | 67.4 | 65.7 | 67.1 | 65.4 | 67.1 | 60.6 | 64.1 | 58.6 | 64.1 | 76.6 |
| PWB (ICCV'19) [15] | VGG16 | 18.4 | 16.7 | 19.6 | 25.4 | 20.0 | 23.8 | 20.3 | 19.2 | 21.7 | 23.4 | 22.6 | 23.8 |
| PFNet (TPAMI'20) [25] | VGG16 | 33.4 | 36.0 | 34.1 | 32.8 | 34.1 | 60.0 | 35.9 | 40.7 | 38.1 | 36.1 | 37.7 | 61.6 |
| SAGNN (CVPR'21) [30] | VGG16 | 35.0 | 40.5 | 37.6 | 36.0 | 37.3 | 61.2 | 37.2 | 45.2 | 40.4 | 40.0 | 40.7 | 63.1 |
| Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
| 1-shot mIoU | 63.2 | 68.0 | 58.0 | 58.9 | 58.9 | 61.7 | 59.2 | 68.1 | 60.0 | 59.0 | 60.4 | 74.5 |
| Baseline | VGG16 | 58.9 | 69.1 | 63.2 | 55.7 | 61.7 | 73.7 | 63.4 | 70.7 | 68.1 | 59.0 | 65.3 | 77.2 |
| DPCN | VGG16 | 61.1 | 69.8 | 58.4 | 56.3 | 61.4 | 71.5 | 63.7 | 70.9 | 58.7 | 57.4 | 62.7 | 73.7 |
| ResNet50 | VGG16 | 65.7 | 71.6 | 69.1 | 60.6 | 66.7 | 78.0 | 70.0 | 73.2 | 70.9 | 65.5 | 69.9 | 80.7 |

Table 2. Comparison with state-of-the-arts on COCO-20 dataset under both 1-shot and 5-shot settings. mIoU of each fold, and averaged mIoU & FB-IoU of all folds are reported. Baseline results are achieved by removing three modules (i.e., SAM, FFM, and DCM) in DPCN.

| Methods | Backbone | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
|---------|----------|--------|--------|--------|--------|------|--------|--------|--------|--------|--------|------|--------|
| Baseline | VGG16 | 38.4 | 68.0 | 58.0 | 58.9 | 58.8 | 71.2 | 60.7 | 68.8 | 60.2 | 52.2 | 60.4 | 74.5 |
| DPCN | VGG16 | 38.9 | 69.1 | 63.2 | 55.7 | 61.7 | 73.7 | 63.4 | 70.7 | 68.1 | 59.0 | 65.3 | 77.2 |
| Baseline | VGG16 | 61.1 | 69.8 | 58.4 | 56.3 | 61.4 | 71.5 | 63.7 | 70.9 | 58.7 | 57.4 | 62.7 | 73.7 |
| DPCN | ResNet50 | 65.7 | 71.6 | 69.1 | 60.6 | 66.7 | 78.0 | 70.0 | 73.2 | 70.9 | 65.5 | 69.9 | 80.7 |

Table 3. Ablation studies of main model components.

DPCN significantly outperforms recent methods under both 1-shot and 5-shot settings. With the ResNet50 backbone, DPCN achieves 3.8% and 2.9% of mIoU improvement over HSNet [17] (previous SOTA) under 1-shot and 5-shot settings, respectively. In addition, DPCN gains significant improvement over the baseline models. For example, DPCN with VGG16 backbone achieves 5.6% and 9.0% mIoU improvement over the baseline model, which proves the superiority of our model in such challenging scenarios.

Qualitative Results. In Fig. 5, we report some quantitative results generated from our DPCN and baseline model on the PASCAL-5-i and COCO-20-i benchmarks. Compared with the baseline, we can see that DPCN exhibits better performance in capturing object details. For instance, more tiny details are preserved in the segmentation of plants and bike (first two columns in Fig. 5). Refer to supplementary materials for more qualitative results.

4.3. Ablations

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

Components Analysis. DPCN contains three major components, i.e., support activation module (SAM), feature filtering module (FFM), and dynamic convolution module (DCM). Table 3 presents our validation on the effectiveness of each component. DCM, which is the most important component, achieves 71.1% mIoU improvement over the baseline model, which proves the superiority of our model in such challenging scenarios.
Figure 5. Qualitative results of our method DPCN and baseline model on PASCAL-5i and COCO-20i benchmarks. Zoom in for details.

Table 5. Ablation studies on different kernel variants of DCM.

| Kernel variants | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
|-----------------|--------|--------|--------|--------|------|--------|
| w/o DCM         | 63.6   | 69.7   | 65.0   | 59.6   | 64.5 | 75.2   |
| 5 × 5           | 64.7   | 71.2   | 65.3   | 58.7   | 65.0 | 76.2   |
| 1 × 5           | 64.9   | 71.0   | 65.2   | 59.7   | 65.2 | 75.4   |
| 1 × 5, 5 × 1    | 63.2   | 71.4   | 64.8   | 59.2   | 64.7 | 75.4   |

| 1 × 5, 5 × 5 × 1| 65.7   | 71.6   | 69.1   | 60.6   | 66.7 | 78.0   |

Table 6. Generalization ability of the proposed DCM.

| Methods         | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
|-----------------|--------|--------|--------|--------|------|--------|
| CANet           | 53.5   | 65.9   | 51.3   | 51.9   | 55.4 | 66.2   |
| CANet+DCM       | 64.7   | 66.8   | 51.8   | 51.9   | 58.8 | 69.3   |
| PFENet          | 61.7   | 69.5   | 55.4   | 56.3   | 60.8 | 73.3   |
| PFENet+DCM      | 62.2   | 69.6   | 59.2   | 58.0   | 62.3 | 73.5   |

Table 7. Fusion strategies for 5-shot setting.

| Methods         | Fold-0 | Fold-1 | Fold-2 | Fold-3 | Mean | FB-IoU |
|-----------------|--------|--------|--------|--------|------|--------|
| 1-shot baseline | 65.7   | 71.6   | 69.1   | 60.6   | 66.7 | 78.0   |
| Vote            | 66.2   | 72.3   | 71.1   | 62.6   | 68.1 | 79.1   |
| Mask-Avg        | 67.2   | 72.4   | 71.4   | 62.7   | 68.4 | 79.7   |
| Mask-OR         | 66.0   | 72.1   | 70.5   | 61.7   | 67.6 | 79.1   |
| Ours            | 69.2   | 73.2   | 70.9   | 65.5   | 69.9 | 81.1   |

Table 5. Ablation studies on different kernel variants of DCM.

Choosing both square kernel and asymmetric kernels as dynamic kernels.

Generalization of DCM. DCM can be utilized as a plug-and-play module to further improve current prototype-based methods. To verify this, we apply DCM to CANet [36] and PFENet [25]. As shown in Table 6, DCM brings significant improvements on both CANet and PFENet.

5-shot Fusion Strategies. We compare our 5-shot fusion strategy (discussed in §3.6) with voting strategy [17], average [20] and OR [36] strategies on masks in Table 7. We can see that our 5-shot fusion strategy achieves 3.2% mIoU improvement and outperforms other fusion strategies.

5. Conclusion

We propose a dynamic prototype convolution network (DPCN) with three major components (i.e., SAM, FFM, and DCM) to address the challenging FSS task. To better mine information from query image, we propose SAM and FFM to generate pseudo query mask and filter background information, respectively. Moreover, a plug-and-play module DCM is designed to implement sufficient interaction between support and query features. Extensive experiments demonstrate that DPCN achieves state-of-the-art results.
References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Plantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13979–13988, 2021. 

[2] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11030–11039, 2020. 

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017. 

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

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010. 

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3146–3154, 2019. 

[7] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In 2011 International Conference on Computer Vision, pages 991–998. IEEE, 2011. 

[8] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multiscale filters for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3562–3572, 2019. 

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017. 

[11] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. Advances in neural information processing systems, 29:667–675, 2016. 

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

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. 

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 

[15] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In European Conference on Computer Vision, pages 784–800. Springer, 2020. 

[16] Xiaolui Luo, Zhuotao Tian, Taiping Zhang, Bei Yu, Yuan Yan Tang, and Jiaya Jia. Pfnet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. arXiv preprint arXiv:2109.13788, 2021. 

[17] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 622–631, 2019. 

[18] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5249–5258, 2019. 

[19] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018. 

[20] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4315–4324, 2017. 

[21] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5249–5258, 2019. 

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 

[23] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. arXiv preprint arXiv:1703.05175, 2017. 

[24] Guolei Sun, Wenguang Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In European conference on computer vision, pages 347–365. Springer, 2020. 

[25] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior mask. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9197–9206, 2019. 

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Proposal attention for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018. 

[27] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Hang Su. Self-supervised equivariant attention mechanisms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11543–11551, 2020.
nism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[29] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 517–526, 2021.

[30] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5475–5484, 2021.

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

[32] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. *arXiv preprint arXiv:2103.15402*, 2021.

[33] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[34] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

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

[36] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[38] Lu Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2021.

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

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.