Pixel Matching Network for Cross-Domain Few-Shot Segmentation

Hao Chen[1] Yonghan Dong[2] Zheming Lu[1] Yunlong Yu[1*] Jungong Han[3]
[1]Zhejiang University [2]Huawei Technologies Ltd. [3]University of Sheffield
{chen_hao_zju, yuyunlong, zheminglu}@zju.edu.cn dongyonghan@huawei.com jungonghan77@gmail.com

Abstract

Few-Shot Segmentation (FSS) aims to segment the novel class images with a few annotated samples. In the past, numerous studies have concentrated on cross-category tasks, where the training and testing sets are derived from the same dataset, while these methods face significant difficulties in domain-shift scenarios. To better tackle the cross-domain tasks, we propose a pixel matching network (PMNet) to extract the domain-agnostic pixel-level affinity matching with a frozen backbone and capture both the pixel-to-pixel and pixel-to-patch relations in each support-query pair with the bidirectional 3D convolutions. Different from the existing methods that remove the support background, we design a hysteretic spatial filtering module (HSFM) to filter the background-related query features and retain the foreground-related query features with the assistance of the support background, which is beneficial for eliminating interference objects in the query background. We comprehensively evaluate our PMNet on ten benchmarks under cross-category, cross-dataset, and cross-domain FSS tasks. Experimental results demonstrate that PMNet performs very competitively under different settings with only 0.68M parameters, especially under cross-domain FSS tasks, showing its effectiveness and efficiency. Code will be released at: https://github.com/chenhao-zju/PMNet

1. Introduction

With the rapid development of computer vision [13, 55], semantic segmentation [33, 36], as one of the most important vision fields, has made remarkable progress. The great success of semantic segmentation benefits from a large amount of human-annotated datasets. However, the pixel-level annotations are hard to obtain due to the time consumption and labor. To alleviate this problem, Few-Shot Segmentation (FSS) [3,4,11,19,39,47,53] aims at segmenting samples with a few annotated support samples and has been attracting a lot of attention.

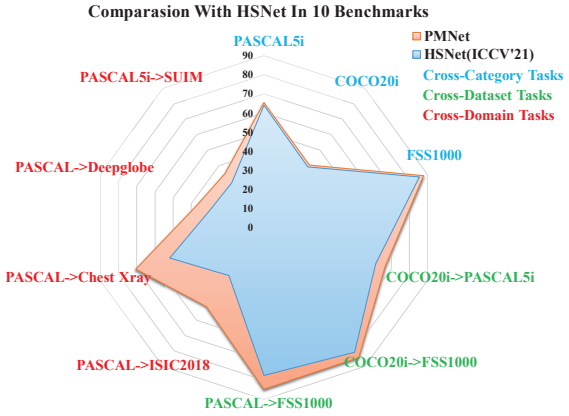

Figure 1. Comprehensive comparison results between HSNet [35] and PMNet on 10 benchmarks. All experiments are tested with Resnet50 backbone under the 1-Shot setting. The experiment results of HSNet are collected from [23, 35, 45].

FSS remains a very challenging task due to the scarcity of support samples and the diverse intra-class flavors. The crux of FSS is to exploit the affinities among the objects in the support-query pairs. Currently, the existing approaches roughly follow two groups, *i.e.*, class-wise and pixel-wise, based on the representations of support samples. The class-wise methods [18, 24, 28, 34, 52] perform the support masks on the feature maps of the support samples to obtain their foreground prototype vectors and utilize them to guide the segmentation of the query images. As the compressed prototype vectors only contain the most manifest information while losing the spatial structure information that is essential for the dense FSS task, the class-wise methods fail in conducting fine-grained matches with target objects in the query image. To remedy the spatial information loss, the pixel-wise methods [35, 38] represent the support-query pairs with pixel-wise features and obtain the support-query affinity by performing the dense many-to-many correspondence.

Though pixel-wise methods achieve superior performances, however, segmenting the query samples from the support-query correlation matrix once obtaining it would lead to inferior relation matching due to the following rea-

*The corresponding author.

sons. First, the pixel-to-pixel correlation hardly handles the case where the objects in support-query pairs are in different sizes. Second, the foreground objects and background objects may give high relevance but the parts of the foreground objects may get low relevance due to the inter-class similarity and intra-class diversity, which will mislead the model in incomplete query objects and discovering interference objects in the background. Third, most existing approaches remove the support backgrounds with the support mask in advance and only consider the object foregrounds in the dense correlations, which will omit some important information for the query segmentation.

To address the above issues, we develop a pixel matching network (PMNet) for FSS to enhance the correspondence matching from the naive support-query affinity by fully exploiting the dense pixel-to-patch relations with feasible sizes between the foreground and background. Different from HSNet [35] which removes the support background with the support mask before extracting the support-query correlations, PMNet obtains the support-query affinity from the whole feature maps of support-query pairs, which fully considers the support background. Denser support-query affinity is, more foreground-related features will be retained and more background-related features will be filtered in the follow-up hysteretic spatial filtering module, thus leading to a better prediction of query mask. In contrast to DCAMA [38] that weights the support-query correlation with the support mask once obtaining it, we propose HSFM to further enhance the support-query correlation with bidirectional 3D convolutions before filtering it. Specifically, the bidirectional 3D convolutions exploit the support-query correlations in both pixel-to-pixel and pixel-to-patch with flexible size, reaching more dense correlations and more fine-grained matching between the support-query pairs, which contributes to segmenting the query target objects with the huge different size from the support objects. After the support-query affinity enhancement, HSFM filters the query background-related features and retains the foreground-related features with the assistance of the support mask.

Besides, the existing dense pixel-level approaches (*e.g.*, DCAMA [38]) add a linear head on the top of each block of the backbone to reduce the noises before obtaining the affinity matrix. Though effective it is, the linear head introduces a large number of parameters, resulting in a heavy segmentation head. To lighten the model, we remove the linear head and mitigate the noises after obtaining the affinity matrix. With more convolutional layers produced in the affinity matrix, PMNet fuses multi-scale and multi-receptive-field feature maps with fewer parameters, which provides the learnable space for reducing the noises.

The main contributions of this work include:

- We propose a lightweight FSS framework, pixel matching network (PMNet), with 0.68M parameters, to boost the support-query correspondence matching by fully exploiting the pixel matching of both the background and foreground in each support-query pair.

- We develop HSFM to further enhance the support-query affinity with bidirectional 3D convolutions and filter it with the support mask, which exploits both the pixel-to-pixel and pixel-to-patch relationships.

- As far as we know, this is the first work to comprehensively evaluate FSS under cross-category, cross-dataset, and cross-domain settings. The experiments on ten benchmarks show the superior generalization ability of our method. As shown in Fig. 1, PMNet comprehensively outperforms the SOTA competitor HSNet [35] under different settings.

2. Related work

2.1. Fow-Shot Semantic Segmentation

The existing FSS works [8, 12, 21] could roughly follow two groups, *i.e.*, class-wise [37, 40, 49, 54] and pixel-wise [5, 20] correspondence, based on the interaction between the support-query pairs. The class-wise approaches represent the objects in the support samples as a prototype through masked average pooling and segment the query samples based on the similarities between the support prototype and the pixel feature embeddings of the query sample. The class-wise approaches differ in the way of obtaining the support prototypes. For example, ASGNet [24] extracts multiple prototypes via clustering and adaptively allocates these prototypes to the most related query pixels, DPCN [28] introduces a dynamical convolutional module to extract prototypes containing support object details and extracts query foreground through feature fusion, and NTRENet [30] extracts the background prototype to eliminate the similar query region in a prototype-to-pixel way. Though efficient, the class-wise approaches compress the support feature maps into a prototype vector and lose their spatial information which is essential for the segmentation tasks.

To remedy the spatial information loss, the pixel-wise methods obtain the support-query correlation by densely calculating the similarities between the pixel feature embeddings of support-query pairs. Though more operations than the class-wise competitors, the pixel-wise approaches obtain superior performances and gain much more attention in recent years. These approaches boost the performances by enhancing the support-query correlations. For example, as one of the earliest works to compute pixel-to-pixel correlation, DAN [43] enhances the support-query correlation with a graph attention mechanism. HSNet [35]

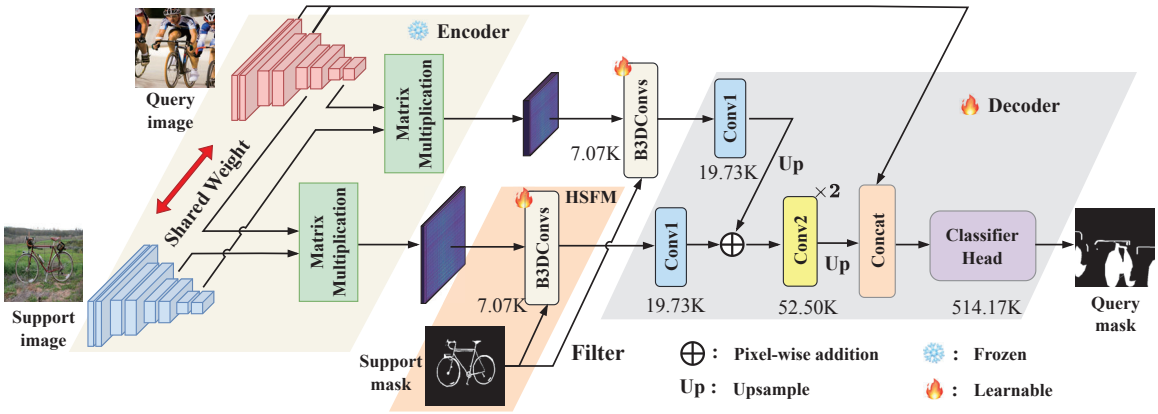

Figure 2. Illustration of the proposed pixel matching network (PMNet). PMNet extracts the multi-scale features from the frozen backbone and reaches the support-query affinity matrix between query features and support features from each layer of the last two backbone blocks. Then, HSFM filters the query background and retains the query foreground with the support mask in two different scale paths (only one path is drawn). In the HSFM, the query-support affinity matrix is further enhanced with the proposed bidirectional 3D convolutions (B3DConvs). Finally, the filtered query-support affinity matrix from different blocks combined with the low-level query features are used to predict the query mask with a decoder. The number next to each module denotes its parameter.

enhances support-query correspondence with 4D convolution operations. Following HSNet [35], VAT [14] replaces the 4D convolution operations with a 4D swin transformer [32] to further enhance the support-query correspondence. DCAMA [38] enhances the support-query correlations by fully aggregating the pixel-to-pixel relations of all layers of each backbone block. Our PMNet is related to HSNet [35] and DCAMA [38] and enhances the support-query correlation by first densely calculating the pixel-to-pixel relations and then further processing it with bidirectional 3D convolutions. Different from HSNet and DCAMA that filter the support background [35] or weight the affinity matrix with the support mask once obtaining it [38], PMNet introduces a hysteretic spatial filtering module to fully consider the pixel-to-pixel and pixel-to-patch relationships of background and foreground between the support-query pairs before filtering it.

Many works have explored the effects of support backgrounds. BAM [22] focuses on the interference brought by the support background and designs a branch to learn the base category distribution. HM [29] compares the difference between the backbone features with or without the background of the input image and complements each other. Different from these methods, PMNet focuses on the affinity matching of background in a more fine-grained way, which contributes to filtering similar objects in query samples.

2.2. Cross-Domain Few-Shot Segmentation

Cross-Domain Few-Shot Segmentation is a specific FSS scenario that performs the trained model on the novel classes from a novel domain. In contrast to ordinary FSS, the cross-domain FSS is more challenging due to the shortage of the prior distribution of the testing set. There are a few attempts to address cross-domain FSS. RTD [45] designs a two-stage method to transfer the feature-enhancement knowledge to target samples. PATNet [23] measures the cross-domain tasks difficulty and proposes a pyramid-based module to transform the domain-specific features into domain-agnostic ones. In this work, we extensively evaluate PMNet in cross-domain tasks and show that PMNet has a great transfer ability between different domains.

3. Methods

3.1. Problem Definition

Given a base set consisting of some images and their masks, FSS aims to train a model to segment some samples from the novel classes, under the condition of one or a few support samples associated with their masks. In this work, we adopt the popular meta-learning paradigm to train the model with the episodes sampled from the base set, where each episode contains a support set $\mathcal{S} = \{I_k^s,\ M_k^s\}_{k=1}^K$ and a query set $\mathcal{Q} = \{I^q, M^q\}$, I^s and I^q represents the input support and query images, M^s and M^q denotes the corresponding masks, and K denotes the number of support images. M^q is required to be predicted during inference.

3.2. Framework

The FSS task is to match the objects from the same category in the support-query pairs. In this paper, we propose a novel framework to address FSS by densely exploiting pixel-level support-query pairs. As illustrated in Fig. 2,

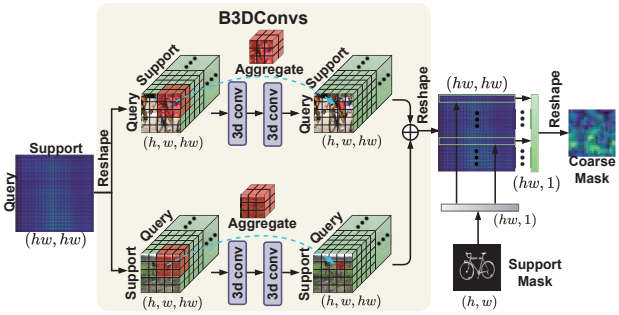

Figure 3. Illustration of the proposed Hysteretic Spatial Filtering Module (HSFM). Bidirectional 3D Convolutions (B3DConvs) aim to enhance the support-query affinity. Then, the enhanced support-query affinity is filtered by suppressing the query background and retaining the query foreground with the assistance of the support mask.

the framework could be divided into four steps. (1) The support-query pair is taken as the input to a feature encoder (*e.g.*, ResNet50) to obtain their feature maps from different blocks. (2) The feature maps from the last two blocks are used to reach the support-query affinities. Same as the pixel-wise methods [35,38], we compute the support-query correlations with the fully pixel-to-pixel inner product of feature maps from each layer in blocks and concatenate them separately. (3) The affinity matrix is filtered through the hysteretic spatial filtering module (HSFM) with the corresponding support mask, which contributes to obtaining the coarse mask predictions of the query sample by filtering the query background and retaining the query foreground. In the HSFM, the support-query affinities are enhanced with the bidirectional 3D convolutions that could further exploit the pixel-to-patch relationships, contributing to segmenting the query objects with different sizes from the support objects. (4) The coarse mask predictions are refined in a decoder network to predict the query mask with the low-level query feature embeddings from the first two backbone blocks.

Steps (3) and (4) are key designs in our framework, which will be introduced in detail.

3.3. Hysteretic Spatial Filtering Module

Given the feature maps $F_s \in \mathbb{R}^{c\times h\times w}$ and $F_q \in \mathbb{R}^{c\times h\times w}$ of both support and query samples, where c, h, and w denote the channel, height, and width, respectively, the support-query affinity matrix $S \in \mathbb{R}^{hw\times hw}$ could be obtained with

$$S = R(F_s)^T \times R(F_q), \tag{1}$$

where R denotes the reshaping operation that reshapes the feature maps of support and query samples to $\mathbb{R}^{c\times hw}$. To further enhance the affinity matrix, we introduce the bidirectional 3D convolutions (B3DConvs) on it to exploit the pixel-to-patch correlations between support-query pairs. As shown in Fig. 3, we first reshape S to $\mathbb{R}^{h\times w\times hw}$ and input it to a 3D convolution network with different kernels. Each network contains two 3D convolution blocks, consisting of a 3D convolutional layer, a BatchNorm layer, and a ReLU operation. Thus, each pixel of the support sample could interact with the query patches whose size is the same as the kernels (*e.g.*, 3×3 and 5×5). Similarly, we could explore the interactions of each query pixel and the support patches by transposing and reshaping S and then inputting it into the 3D convolution network. In this way, we could obtain the pixel-to-patch relationships in a *bidirectional* way, *i.e.*, support-to-query, and query-to-support. This process is formulated as:

$$S_{enh} = R_2(H(R_1(S))) + R_2(H(R_1(S^T)))^T, \tag{2}$$

where $S_{enh} \in \mathbb{R}^{hw\times hw}$ denotes the enhanced affinity matrix, H denotes a 3D convolution network, R_1 and R_2 are both shaping operations, T is the metric transpose.

Once obtained the enhanced support-query affinity matrix, we introduce the support mask to filter the query foreground. As shown in Fig. 3, the reshaped support mask $M_s \in \mathbb{R}^{hw\times 1}$ performs as a convolution operator on the affinity matrix along the query dimension to obtain the coarse query mask, which is formulated as:

$$M_{coarse} = Conv(input : S_{enh}, weight : M_s) \tag{3}$$

where $M_{coarse} \in \mathbb{R}^{hw\times 1}$ denotes the coarse mask of query sample, $Conv$ denotes the convolution operation.

In contrast to DCAMA [38] which weights affinity matrix with the support mask once obtaining it, PMNet introduces a hysteretic spatial filtering module, which takes the affinity matrix as two parts and produces 3D convolutions to them separately before filtering. Through a fine-grained operation, our method guarantees the affinity matching of the foreground and background, benefiting suppressing the related background and highlighting the related foreground in spatial filtering. Fig. 4 provides some filtered feature maps of different approaches and shows that PMNet retains more query foreground and filters more query background, which will benefit in the following mask prediction.

3.4. Decoder

The decoder takes the coarse query masks from different layers as the input to predict the final mask. Specifically, the decoder fuses the multi-scale and multi-receptive-field coarse masks with Conv1 and Conv2. The coarse masks of two blocks are enhanced with Conv1 separately and fused by pixel-wise addition. The fused feature is continuously enhanced with two Conv2 blocks. To supplement the structural information, the feature maps are concatenated with

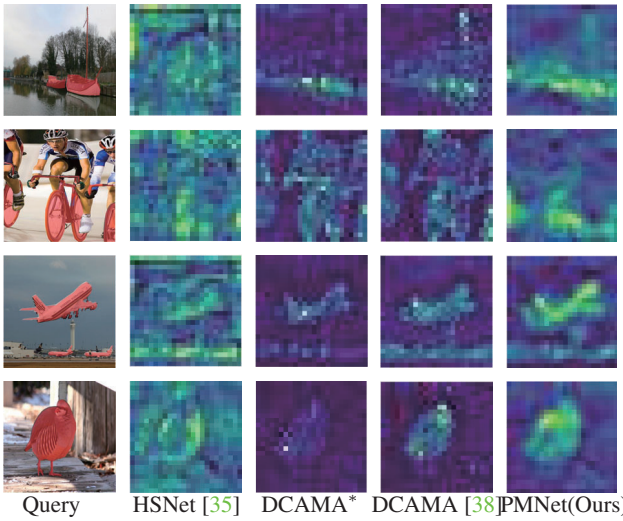

Figure 4. The visualization of the coarse mask in the second block of different models, trained under 1-shot task with Resnet50 backbone on PASCAL5^i dataset. $*$ denotes DCAMA with the last two backbone blocks.

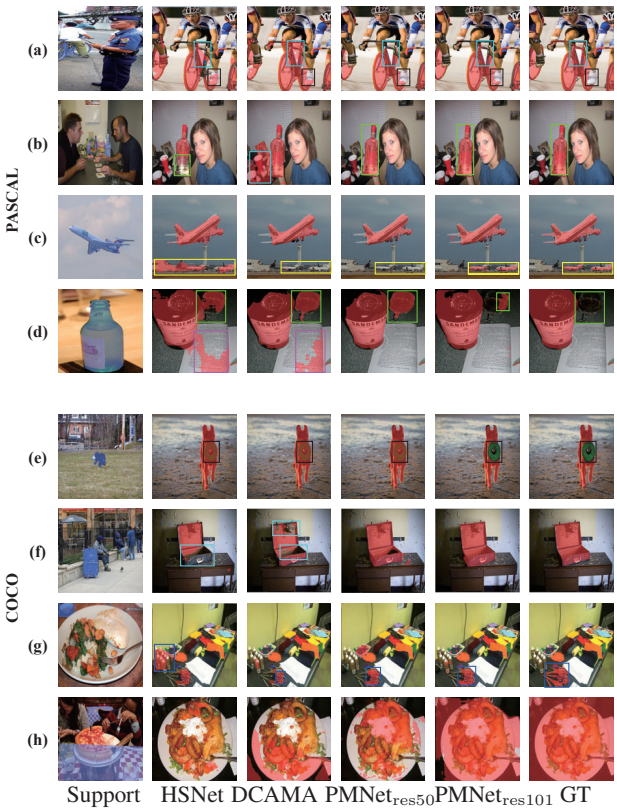

Figure 5. The visualization of prediction results with different methods in COCO and PASCAL datasets under various challenging scenarios. 'GT' denotes the query ground-truth mask. 'PMNet$_{res50}$' and 'PMNet$_{res101}$' denote PMNet with Resnet50 and Resnet101 backbone, respectively.

the low-level query features and segmented with the classifier head. The detailed Conv1, Conv2, and Classifier head will be given in Supplementary Material.

In [38, 51], a linear head is added for each of the last two modules to transform the feature embeddings into the other feature embedding to better fit the segmentation task from the pre-trained backbone. Though the linear head could assist the model in improving the cross-category FSS performance, it consists of a large number of learnable parameters. To lighten the model, we remove the linear heads as shown in Fig. 2 and remedy it with the post-processing in the decoder with fewer parameters. Besides, we observe that removing the linear heads could preserve more domain-invariant features, which will benefit the cross-domain FSS tasks, please refer to more details in Supplementary Material. During training, the binary cross-entropy loss is used to supervise the model. Under the 5-shot setting, multiple support features and masks are concatenated in the channel dimension in Hysteretic Spatial Filtering Module.

4. Experiments

4.1. Datasets.

PASCAL-5^i [9] is the expansion of PASCAL VOC 2012, which is divided into 4 subsets following [1]. Each subset contains 15 base classes for training and 5 novel classes for testing. **COCO-20^i [26]** contains 80 common classes in natural scenery. Same as PASCAL-5^i, COCO-20^i is divided into 4 subsets, each containing 60 base classes and 20 novel classes. **FSS-1000 [25]** is a natural image dataset, consisting of 1,000 classes and each class has 10 annotated samples. The official split has been used in our FSS experiments. The results are tested on the testing set, containing 240 classes and 2,400 samples. **Deepglobe [7]** is a satellite image dataset, containing 7 classes: urban land, agriculture, rangeland, forest, water, barren, and unknown. Following PATNet [23], each image is cut into 6 pieces. **ISIC2018 [6, 42]** is a dataset on dermatoscopic images, containing 2,596 skin cancer screening samples. In PATNet [23], ISIC2018 is divided into three classes, while the classification basis has not been published. In our work, the whole dataset is seen as one class, which is more challenging. **Chest X-ray [2, 17]** is an X-ray dataset for Tuberculosis, which includes 566 annotated x-ray images, collected from 58 cases with a manifestation of Tuberculosis and 80 normal cases. **SUIM [16]** is an underwater imagery dataset, containing over 1,500 pixel-annotated images for eight classes.

4.2. FSS Tasks.

On the basis of the different distributions of the training dataset and testing dataset, the FSS tasks are divided

Backbone	Methods	Type	1-shot					5-shot					Learnable
			Fold-0	Fold-1	Fold-2	Fold-3	mIoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	Params(M)
Pascal-5^i													
ResNet-50	NTRENet(CVPR'22) [30]	Prototype	65.4	72.3	59.4	59.8	64.2	66.2	72.8	61.7	62.2	65.7	18.6
	SSP(ECCV'22) [10]		60.5	67.8	**66.4**	51.0	61.4	67.5	72.3	**75.2**	62.1	69.3	8.7
	IPMT(NeurIPS'22) [31]		**72.8**	**73.7**	59.2	61.6	66.8	73.1	**74.7**	61.6	63.4	68.2	-
	CyCTR(NeurIPS'21) [51]	Pixelwise	65.7	71.0	59.5	59.7	64.0	69.3	73.5	63.8	63.5	67.5	15.9*
	HSNet(ICCV'21) [35]		64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	67.1	69.5	2.6
	AAFormer(ECCV'22) [46]		69.1	73.3	59.1	59.2	65.2	72.5	**74.7**	62.0	61.3	67.6	-
	VAT(ECCV'22) [14]		67.6	72.0	62.3	60.1	65.5	72.4	73.6	68.6	65.7	70.1	3.2
	DCAMA(ECCV'22) [38]		67.5	72.3	59.6	59.0	64.6	70.5	73.9	63.7	65.8	68.5	14.2*
	FECANet(TMM'23) [27]		69.2	72.3	62.4	**65.7**	**67.4**	72.9	74.0	65.2	**67.8**	70.0	3.5
	ABCNet(CVPR'23) [15]		68.8	73.4	62.3	59.5	66.0	71.7	74.2	65.4	67.0	69.6	-
	PMNet (Ours)		67.3	72.0	62.4	59.9	65.4	**73.6**	74.6	69.9	67.2	**71.3**	**0.68**
ResNet-101	NTRENet(CVPR'22) [30]	Prototype	65.5	71.8	59.1	58.3	63.7	67.9	73.2	60.1	66.8	67.0	18.6
	SSP(ECCV'22) [10]		63.2	70.4	**68.5**	56.3	64.6	70.5	76.4	**79.0**	66.4	73.1	27.7
	IPMT(NeurIPS'22) [31]		**71.6**	73.5	58.0	61.2	66.1	**75.3**	**76.9**	59.6	65.1	69.2	-
	HSNet(ICCV'21) [35]	Pixelwise	67.3	72.3	62.0	63.1	66.2	71.8	74.4	67.0	68.3	70.4	2.6
	AAFormer(ECCV'22) [46]		69.9	**73.6**	57.9	59.7	65.3	75.0	75.1	59.0	63.2	68.1	-
	VAT(ECCV'22) [14]		70.0	72.5	64.8	**64.2**	67.9	75.0	75.2	68.4	69.5	72.0	3.3
	DCAMA(ECCV'22) [38]		65.4	71.4	63.2	58.3	64.6	70.7	73.7	66.8	61.9	68.3	14.2*
	ABCNet(CVPR'23) [15]		65.3	72.9	65.0	59.3	65.6	71.4	75.0	68.2	63.1	69.4	-
	PMNet (Ours)		71.3	72.4	66.9	61.9	**68.1**	74.9	75.5	75.3	**69.8**	**73.9**	**0.68**
COCO-20^i													
ResNet-50	NTRENet(CVPR'22) [30]	Prototype	36.8	42.6	39.9	37.9	39.3	38.2	44.1	40.4	38.4	40.3	18.6
	SSP(ECCV'22) [10]		35.5	39.6	37.9	36.7	37.4	40.6	47.0	45.1	43.9	44.1	8.7
	IPMT(NeurIPS'22) [31]		41.4	45.1	45.6	40.0	43.0	43.5	49.7	48.7	47.9	47.5	-
	HSNet(ICCV'21) [35]	Pixelwise	36.3	43.1	38.7	38.7	39.2	43.3	51.3	48.2	45.0	46.9	2.6
	AAFormer(ECCV'22) [46]		39.8	44.6	40.6	41.4	41.6	42.9	50.1	45.5	49.2	46.9	-
	VAT(ECCV'22) [14]		39.0	43.8	42.6	39.7	41.3	44.1	51.1	50.2	46.1	47.9	3.2
	DCAMA(ECCV'22) [38]		41.9	45.1	44.4	**41.7**	43.3	45.9	50.5	50.7	46.0	48.3	14.2*
	FECANet(TMM'23) [27]		38.5	44.6	42.6	40.7	41.6	44.6	51.5	48.4	45.8	47.6	3.5
	ABCNet(CVPR'23) [15]		**42.3**	**46.2**	**46.0**	42.0	**44.1**	45.5	**51.7**	**52.6**	46.4	49.1	-
	PMNet (Ours)		39.8	41.0	40.1	40.7	40.4	**50.1**	51.0	50.4	**49.6**	**50.3**	**0.68**
ResNet-101	NTRENet(CVPR'22) [30]	Prototype	38.3	40.4	39.5	38.1	39.1	42.3	44.4	44.2	41.7	43.2	18.6
	SSP(ECCV'22) [10]		39.1	45.1	42.7	41.2	42.0	47.4	54.5	50.4	49.6	50.2	27.7
	IPMT(NeurIPS'22) [31]		40.5	45.7	44.8	39.3	42.6	45.1	50.3	49.3	46.8	47.9	-
	HSNet(ICCV'21) [35]	Pixelwise	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5	2.6
	DCAMA(ECCV'22) [38]		41.5	**46.2**	**45.2**	41.3	43.5	48.0	**58.0**	**54.3**	47.1	51.9	14.2*
	PMNet (Ours)		**44.7**	44.3	44.0	**41.8**	**43.7**	**52.6**	53.3	53.5	**52.8**	**53.1**	**0.68**

Table 1. FSS performances (%) on PASCAL-5^i and COCO-20^i with different backbones (ResNet50 and ResNet101). '$*$' denotes the results obtained by ourselves with the released codes. The results of all the competitors are from the published literature. The best and the second best results are marked in **bold** and underline, respectively.

into three types, cross-category, cross-dataset, and cross-domain. **(1) Cross-category setting** considers a scenario where both the base categories and test categories are sampled from the same dataset. **(2) Cross-dataset setting** evaluates the model trained with one dataset on the other dataset without fine-tuning. Note that the trained dataset and the evaluated dataset follow either the same distribution or different distributions. In this work, the cross-dataset setting refers to both the trained and evaluated datasets following the same distribution if not specified. **(3) Cross-domain setting** is a specific case of cross-dataset setting where the trained and evaluated datasets are from different domains. This is a more challenging setting as the model not only deals with the novel classes but also has to address the domain gap among different datasets.

4.3. Comparison with State-of-the-Art

Cross-category Task. Tab. 1 shows the comparison results on both PASCAL-5^i and COCO-20^i with two different backbones. With Resnet50 backbone, PMNet achieves 71.3% in PASCAL-5^i and 50.3% in COCO-20^i under the 5-shot setting, and outperforms the second-best competitors with 1.2% and 2.0%, respectively. With Resnet101 backbone, PMNet achieves the best performance under both 1-shot and 5-shot settings in PASCAL-5^i and COCO-20^i datasets, outperforming the second-best competitors with 0.2% and 0.8% in PASCAL5^i, 0.2% and 1.2% in COCO20^i, respectively. Besides, we observe that PMNet has the

Backbone	Methods	1-shot					5-shot					Learnable
		Fold-0	Fold-1	Fold-2	Fold-3	mIoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	Params(M)
		COCO20^i → Pascal5^i										
ResNet-50	PFENet(TPMAI) [41]	43.2	65.1	66.5	69.7	61.1	45.1	66.8	68.5	73.1	63.4	34.3
	RePRI(CVPR'21) [1]	52.2	64.3	64.8	71.6	63.2	56.5	68.2	70.0	76.2	67.7	-
	HSNet(ICCV'21) [35]	45.4	61.2	63.4	75.9	61.6	56.9	65.9	71.3	80.8	68.7	2.6
	VAT(ECCV'22) [14]	52.1	64.1	67.4	74.2	64.5	58.5	68.0	72.5	79.9	69.7	3.2
	HSNet-HM(ECCV'22) [29]	43.4	68.2	**69.4**	**79.9**	65.2	50.7	71.4	**73.4**	**83.1**	69.7	-
	VAT-HM(ECCV'22) [29]	48.3	64.9	67.5	79.8	65.1	55.6	68.1	72.4	82.8	69.7	-
	RTD(CVPR'22) [45]	57.4	62.2	68.0	74.8	65.6	65.7	69.2	70.8	75.0	70.1	-
	PMNet (Ours)	**68.8**	**70.0**	65.1	62.3	**66.6**	**73.9**	**74.5**	73.3	72.1	**73.4**	**0.68**
ResNet-101	HSNet(ICCV'21) [35]	47.0	65.2	67.1	77.1	64.1	57.2	69.5	72.0	82.4	70.3	2.6
	HSNet-HM(ECCV'22) [29]	46.7	68.6	**71.1**	**79.7**	66.5	53.7	70.7	75.2	**83.9**	70.9	-
	RTD(CVPR'22) [45]	59.4	64.3	70.8	72.0	66.6	67.2	72.7	72.0	78.9	72.7	-
	PMNet (Ours)	**71.0**	**72.3**	66.6	63.8	**68.4**	**75.2**	**76.3**	**77.0**	72.6	**75.3**	**0.68**

Table 2. FSS performances (%) on cross-dataset task, COCO-20^i → Pascal-5^i, with different backbones (ResNet50 and ResNet101). The results of all the competitors are from the published literature. The best and the second best results are marked in **bold** and underline, respectively.

Backbone	Methods	Deepglobe		ISIC2018		Chest X-ray		FSS-1000		average	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ResNet50	PANet(ICCV'19) [44]	36.55	**45.43**	25.29	33.99	57.75	69.31	69.15	71.68	47.19	55.1
	RPMMs(ECCV'20) [48]	12.99	13.47	18.02	20.04	30.11	30.82	65.12	67.06	31.56	32.85
	PFENet(TPAMI'20) [41]	16.88	18.01	23.5	23.83	27.22	27.57	70.87	70.52	34.62	34.98
	RePRI(CVPR'21) [1]	25.03	27.41	23.27	26.23	65.08	65.48	70.96	74.23	46.09	48.34
	HSNet(ICCV'21) [35]	29.65	35.08	31.2	35.1	51.88	54.36	77.53	80.99	47.57	51.38
	PATNet(ECCV'22) [44]	**37.89**	42.97	41.16	53.58	66.61	70.2	78.59	81.23	56.06	61.99
	PMNet (Ours)	37.10	41.60	**51.2**	**54.5**	**70.4**	**74.0**	**84.6**	**86.3**	**60.83**	**64.1**

Table 3. FSS performances (%) on four tasks with Resnet50 backbone under 1-shot and 5-shot settings. All models are trained with the whole PASCAL dataset. Following PATNet [23], **Deepglobe, ISIC2018, and Chest X-ray are cross-domain tasks, and FSS-1000 is a cross-dataset task.** The best results are marked in **bold**.

Methods	Fold-0	Fold-1	Fold-2	Fold-3	mIoU
COCO-5^i → FSS1000					
ASGNet [24]	76.2	72.2	72.7	71.6	73.2
HSNet [35]	79.9	80.5	81.1	82.1	80.8
SCL [50]	81.6	78.3	77.5	74.4	78.0
RTD [45]	82.2	82.6	79.6	83.4	81.9
PMNet (Ours)	**85.3**	**83.7**	**83.2**	**84.9**	**84.3**

Table 4. FSS performances (%) on cross-dataset task, COCO-20^i→FSS1000, with Resnet50 backbone under 1-shot setting. The best results are marked in **bold**.

Methods	Fold-0	Fold-1	Fold-2	Fold-3	mIoU
Pascal-5^i → SUIM					
ASGNet [24]	32.4	30.9	28.9	35.2	31.9
HSNet [35]	30.7	30.0	27.3	27.0	28.8
SCL [50]	31.3	31.2	32.2	32.5	31.8
RTD [45]	35.2	33.4	34.3	**36.0**	34.7
PMNet (Ours)	**37.1**	**35.1**	**35.2**	31.7	**34.8**

Table 5. FSS performances (%) on cross-domain task, PASCAL-5^i→SUIM, with Resnet50 backbone under 1-shot setting. The best results are marked in **bold**.

fewest parameters. Please refer to the supplementary material for the results of the FSS1000 dataset.

To visualize the segmentation results, some qualitative results of the 1-shot segmentation task are provided in Fig. 5. From the results, we observe that our method with Resnet50 segments well in (a) multiple targets, (b) complicated background, (f) full segmentation, and (g) object details, but performs not so well in (c) small objects, (d) objects with interference, (e) object is covered, and (h) huge object. These challenging cases cause performance to decline. In contrast, our method with Resnet101 segments well under all challenging scenarios. Owing to B3DConvs connecting the support and query object regions with feasible sizes, PMNet could segment query objects with different sizes to support objects.

Cross-dataset tasks. We conduct experiments on 3

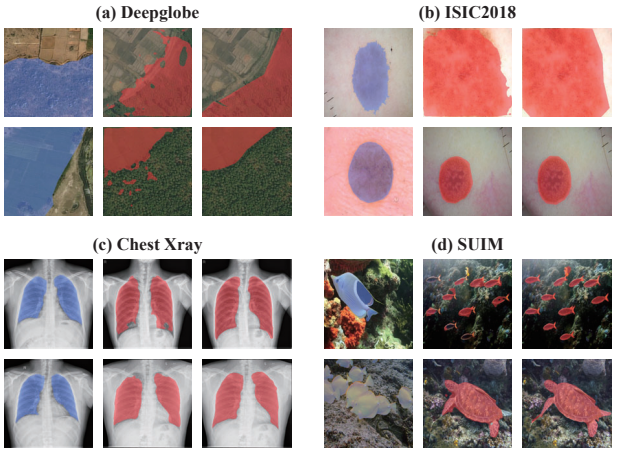

Figure 6. The visualization of prediction results in 4 cross-domain tasks. 'GT' denotes the query ground truth.

	Fold-0	Fold-1	Fold-2	Fold-3	Mean
w/o B3DConvs	67.1	71.2	58.9	58.8	64.0
w/o Background	59.0	65.0	54.3	51.2	57.4
PMNet	67.3	72.0	62.4	59.9	65.4

Table 6. Impact (%) of B3DConvs and the Support Background under 1-shot setting.

cross-dataset tasks. As shown in Tab. 2, PMNet achieves the best performance in all settings of COCO-20^i→PASCAL-5^i task, surpassing the second-best competitors with 1.0% and 3.3% under 1-shot and 5-shot settings with Resnet50 backbone, 1.8% and 2.6% under 1-shot and 5-shot setting with Resnet101 backbone, respectively. For the COCO-20^i→FSS1000 task in Tab. 4, PMNet achieves 84.3% as the best performance and surpasses the second-best competitors by 2.4%. Following PATNet [23], the results of the PASCAL→FSS1000 task are shown in the sixth column of Tab. 3. With Resnet50 backbone, PMNet achieves 84.6% and 86.3%, outperforming the second-best competitors by 6.0% and 5.0% in 1-shot and 5-shot settings, respectively. The significant advantages show the effectiveness of PMNet for the cross-dataset FSS tasks.

Cross-domain tasks. We conduct experiments on cross-domain tasks with 4 datasets, which are collected from different scenarios, not intersecting with the trained dataset PASCAL. As shown in Tab. 3, PMNet achieves the best performance under 1-shot and 5-shot settings on both ISIC2018 and Chest Xray datasets, surpassing the second-best competitors by 10.0% and 0.9% in ISIC2018, 3.8% and 4.8% in Chest Xray, respectively. On the Deepglobe dataset, PMNet achieves the second-best performance with 37.1% and 41.6% under 1-shot and 5-shot settings, respectively. Following RTD [45], the results of the PASCAL-5^i→SUIM task are shown in Sec. 4. Our PMNet performs slightly better than the second-best competitor with Resnet50 backbone under the 1-shot setting. The results on both Tab. 3 and Sec. 4 demonstrate the effectiveness of the proposed approach on cross-domain FSS tasks. As shown in Fig. 6, the segmentation results of four cross-domain tasks show that our method performs well in satellite, dermatoscopic, X-ray, and underwater scenarios.

4.4. Ablation Study

In this subsection, we design a series of ablation studies to evaluate the effects of different modules. All the results are obtained with the ResNet50 backbone under the 1-shot task on the PASCAL dataset.

B3DConvs. Tab. 6 illustrates the impact of B3DConvs on the mIoU under 1-shot setting. From the experiment, we observe that B3DConvs brings a 1.4% improvement to mIoU and a 3.5% improvement in fold-2, which shows the effectiveness of the B3DConvs module, indicating that exploiting the support-query affinity would be beneficial for FSS performance improvement.

Support Background. Tab. 6 also illustrates the impact of the background affinity matching on the mIoU. 'w/o Background' denotes filtering the background feature from support feature maps with the support mask before the affinity calculation. From the results, removing the background before the affinity would lead to a huge dropping, 8.0%, in mIoU, which indicates that the support background benefits in the segmentation of the query sample.

5. Conclusion

In this work, we have proposed a lightweight pixel matching network (PMNet) for FSS by fully exploiting the relationships between the foreground and background in each support-query pair from both pixel-to-pixel and pixel-to-patch ways, which benefits suppressing the background and highlighting the foreground in query features with a hysteretic spatial filtering module (HSFM). From the results, we conclude that the support background could contribute significantly to the query segmentation by assisting in both filtering the query background and exploring the pixel-to-patch correlations in each support-query pair. Extensive experimental results show that PMNet performs very competitively on ten benchmarks under cross-category, cross-dataset, and cross-domain FSS tasks with an extremely small number of parameters.

Acknowledgment

This work is supported in part by the Key R&D Program of Zhejiang Province, China (2023C01043, 2021C01119), NSFC (62002320, U19B2043).

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *CVPR*, pages 13979–13988, 2021. 5, 7

[2] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *TMI*, 33(2):577–590, 2013. 5

[3] Hao Chen, Zhe-ming Lu, and YangMing Zheng. Multi-similarity enhancement network for few-shot segmentation. *IEEE Access*, 2023. 1

[4] Hao Chen, Yunlong Yu, Yonghan Dong, Zheming Lu, Yingming Li, and Zhongfei Zhang. Multi-context interaction network for few-shot segmentation. *arXiv preprint arXiv:2303.06304*, 2023. 1

[5] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *NeurIPS*, 34:9011–9023, 2021. 2

[6] Noel C. F. Codella, Veronica M Rotemberg, Philipp Tschandl, M. E. Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Armando Marchetti, Harald Kittler, and Allan C. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *ArXiv*, abs/1902.03368, 2019. 5

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPR Workshops*, June 2018. 5

[8] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *WACV*, pages 2488–2497, January 2023. 2

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

[10] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, 2022. 6

[11] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *CVPR*, pages 1185–1194, 2021. 1

[12] Byeolyi Han and Tae-Hyun Oh. Learning few-shot segmentation from bounding box annotations. In *WACV*, pages 3750–3759, January 2023. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[14] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *ECCV*, pages 108–126, 2022. 3, 6, 7

[15] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322, June 2023. 6

[16] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IROS*, pages 1769–1776, 2020. 5

[17] S Jaeger, A Karargyris, S Candemir, L Folio, J Siegelman, F Callaghan, Zhiyun Xue, K Palaniappan, R. K Singh, and S and Antani. Automatic tuberculosis screening using chest radiographs. *TMI*, 33(2):233–245, 2014. 5

[18] Siyu Jiao, Gengwei Zhang, Shant Navasardyan, Ling Chen, Yao Zhao, Yunchao Wei, and Humphrey Shi. Mask matching transformer for few-shot segmentation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022. 1

[19] Joakim Johnander, Johan Edstedt, Michael Felsberg, Fahad Shahbaz Khan, and Martin Danelljan. Dense gaussian processes for few-shot segmentation. In *ECCV*, pages 217–234, 2022. 1

[20] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *CVPR*, pages 9979–9990, 2022. 2

[21] Alper Kayabaşı, Gülin Tüfekci, and İlkay Ulusoy. Elimination of non-novel segments at multi-scale for few-shot segmentation. In *WACV*, pages 2559–2567, January 2023. 2

[22] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, pages 8057–8067, 2022. 3

[23] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *ECCV*, pages 73–90, Cham, 2022. Springer Nature Switzerland. 1, 3, 5, 7, 8

[24] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, pages 8334–8343, 2021. 1, 2, 7

[25] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, pages 2869–2878, 2020. 5

[26] 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 *ECCV*, pages 740–755, 2014. 5

[27] Huafeng Liu, Pai Peng, Tao Chen, Qiong Wang, Yazhou Yao, and Xian-Sheng Hua. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *TMM*, 2023. 6

[28] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *CVPR*, pages 11553–11562, 2022. 1, 2

[29] Weide Liu, Chi Zhang, Henghui Ding, Tzu-Yi Hung, and Guosheng Lin. Few-shot segmentation with optimal transport matching and message flow. In *ECCV*, 2022. 3, 7

[30] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *CVPR*, pages 11573–11582, 2022. 2, 6

[31] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022. 6

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[34] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, pages 8741–8750, 2021. 1

[35] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, pages 6941–6952, 2021. 1, 2, 3, 4, 5, 6, 7

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1

[37] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017. 2

[38] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *ECCV*, pages 151–168, 2022. 1, 2, 3, 4, 5, 6

[39] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. In *NeurIPS*, 2022. 1

[40] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, 2022. 2

[41] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*, 2020. 7

[42] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), aug 2018. 5

[43] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, pages 730–746, 2020. 2

[44] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pages 9197–9206, 2019. 7

[45] Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *CVPR*, pages 7065–7074, June 2022. 1, 3, 7, 8

[46] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *ECCV*, pages 36–52, 2022. 6

[47] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*, pages 5471–5480, 2021. 1

[48] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, pages 763–778, 2020. 7

[49] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *ICCV*, pages 8721–8730, October 2021. 2

[50] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, pages 8312–8321, 2021. 7

[51] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *NeurIPS*, 34:21984–21996, 2021. 5, 6

[52] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022. 1

[53] Shan Zhang, Tianyi Wu, Sitong Wu, and Guodong Guo. Catrans: Context and affinity transformer for few-shot segmentation. In *IJCAI*, 2022. 1

[54] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *TCYB*, 50(9):3855–3865, 2020. 2

[55] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023. 1