Cross-Domain Few-Shot Semantic Segmentation via Doubly Matching Transformation

Jiayi Chen[1,2] , **Rong Quan**[1,2] , **Jie Qin**[1,2,*]
[1]Nanjing University of Aeronautics and Astronautics
[2]State Key Laboratory of Integrated Services Networks, Xidian University
chenjiayi68@nuaa.edu.cn, {rongquan0806, qinjiebuaa}@gmail.com

Abstract

Cross-Domain Few-shot Semantic Segmentation (CD-FSS) aims to train generalized models that can segment classes from different domains with a few labeled images. Previous works have proven the effectiveness of feature transformation in addressing CD-FSS. However, they completely rely on support images for feature transformation, and repeatedly utilizing a few support images for each class may easily lead to overfitting and overlooking intra-class appearance differences. In this paper, we propose a Doubly Matching Transformation-based Network (DMTNet) to solve the above issue. Instead of completely relying on support images, we propose Self-Matching Transformation (SMT) to construct query-specific transformation matrices based on query images themselves to transform domain-specific query features into domain-agnostic ones. Calculating query-specific transformation matrices can prevent overfitting, especially for the meta-testing stage where only one or several images are used as support images to segment hundreds or thousands of images. After obtaining domain-agnostic features, we exploit a Dual Hypercorrelation Construction (DHC) module to explore the hypercorrelations between the query image with the foreground and background of the support image, based on which foreground and background prediction maps are generated and supervised, respectively, to enhance the segmentation result. In addition, we propose a Test-time Self-Finetuning (TSF) strategy to more accurately self-tune the query prediction in unseen domains. Extensive experiments on four popular datasets show that DMTNet achieves superior performance over state-of-the-art approaches. Code is available at https://github.com/ChenJiayi68/DMTNet.

1 Introduction

Relying on large-scale labeled datasets [Ros *et al.*, 2016; Richter *et al.*, 2016; Silberman *et al.*, 2012], semantic segmentation [Long *et al.*, 2015; Zhao *et al.*, 2016; Chen *et al.*, 2017; Xie *et al.*, 2021] has achieved rapid development in recent years. However, it is difficult to collect such a large amount of training data which requires massive time and expensive annotation costs in practical scenarios. Few-shot Semantic Segmentation (FSS) [Shaban *et al.*, 2017] has been proposed to reduce the heavy dependence of traditional semantic segmentation models on a large number of labeled images. FSS aims to achieve accurate segmentation of a query image only using a few annotated support images. Existing FSS methods [Rakelly *et al.*, 2018; Zhang *et al.*, 2018; Wang *et al.*, 2019; Liu *et al.*, 2020a; Okazawa, 2022] usually adopt meta-learning [Vinyals *et al.*, 2016; Snell *et al.*, 2017], which consists of two stages: meta-training and meta-testing. In the meta-training stage, an FSS model is trained on many meta-tasks using base classes. The trained model can then perform accurate segmentation on novel classes in the meta-testing stage.

However, in practical applications, there always exists a large domain gap between source and target datasets due to different label spaces and feature distributions, causing inferior generalization of FSS models to unseen domains. To solve the problem of significant performance degradation of FSS models under cross-domain scenarios, Cross-Domain Few-shot Semantic Segmentation (CD-FSS) [Lei *et al.*, 2022] is proposed to simultaneously solve the problems of few shot and domain gaps. The primary CD-FSS method, PATNet [Lei *et al.*, 2022], eliminates the domain gap by transforming domain-specific features into domain-agnostic ones and conducting segmentation in the domain-agnostic feature space. It combines the prototype set of support images with some learned anchor layers to construct transformation matrices, based on which domain-specific features are transformed into domain-agnostic ones. However, during the meta-testing stage, generating domain-agnostic features for hundreds or thousands of query images just based on the transformation matrices obtained from one or several support images can easily lead to overfitting. In addition, we find that most existing CD-FSS methods [Min *et al.*, 2021; Lei *et al.*, 2022] only concentrate on foreground object regions and ignore background regions during the segmentation process. They directly filter out the background of support images and only construct dense correlations with foreground objects. Considering that objects belonging to the same class

* Corresponding Author

mostly lie in similar environments, segmentation based on the similarities between not only the foreground objects but also the backgrounds is very likely to result in better performance.

Based on the above considerations, we propose a novel Doubly Matching Transformation-based Network (DMTNet) for cross-domain few-shot semantic segmentation. DMTNet first exploits a Self-Matching Transformation (SMT) module to construct a unique transformation matrix for each image based on its own prototype. Then, domain-specific features of each query and support image are transformed into domain-agnostic ones self-adaptively, which can avoid overfitting during the meta-testing stage. After obtaining domain-agnostic features, we propose a Dual Hypercorrelation Construction (DHC) module to learn the hypercorrelation between the query image with both the foreground and background of the support image, and generate foreground and background predictions, correspondingly. Supervising both object-wise and background-wise segmentation can enhance the training performance considering the possible similarities between the backgrounds of similar objects. In addition, in the meta-testing stage, we design a Test-time Self-Finetuning (TSF) strategy to further improve query predictions in unseen domains by self-tuning a handful of parameters in the network with support images.

To evaluate the performance of DMTNet, we conduct extensive experiments and ablation studies on four benchmark datasets, including ISIC2018 [Codella *et al.*, 2019], Chest X-ray [Candemir *et al.*, 2014], Deepglobe [Demir *et al.*, 2018], and FSS-1000 [Wei *et al.*, 2019]. Experimental results show that DMTNet achieves remarkable improvement, surpassing state-of-the-art approaches on all four datasets.

In summary, our main contributions are as follows:

- We propose a novel Doubly Matching Transformation-based Network (DMTNet) for CD-FSS. Instead of completely relying on support images, we propose a Self-Matching Transformation (SMT) module to transform domain-specific query features into domain-agnostic ones in a self-adaptive manner. Compared with transformation completely based on one or several support images, DMTNet can avoid overfitting to a large extent.
- We propose a Dual Hypercorrelation Construction (DHC) module to learn the hypercorrelations between the query image and both the foreground and background of the support image. For the first time, we execute segmentation based on both foreground object similarities and background similarities.
- We design a Test-time Self-Finetuning (TSF) strategy for meta-testing. Only self-tuning a handful of parameters in the network can significantly improve query predictions in target domains while maintaining minimal complexity.
- Extensive experimental results show that DMTNet achieves state-of-the-art performance on four CD-FSS benchmarks, *i.e.*, ISIC2018, Chest X-ray, Deepglobe, and FSS-1000.

2 Related Work

Few-shot Semantic Segmentation. FSS aims to generate a pixel-wise prediction of the novel class with only a few labeled support images. Existing FSS methods can be divided into two categories: metric-based methods and relation-based ones. Metric-based methods [Wang *et al.*, 2019; Liu *et al.*, 2020b] represent support images as several class prototypes by utilizing masked average pooling, use non-parametric measurement tools such as cosine similarity to measure the similarity between these prototypes and query features, and segment the query image based on the similarities. However, metric-based methods lose much spatial information when compressing the global feature maps into a prototype vector [Li *et al.*, 2021; Zhang *et al.*, 2021]. Since the prototype has a limited ability to express a target category, some researchers propose relation-based methods [Zhang *et al.*, 2019b; Yang *et al.*, 2020; Tian *et al.*, 2020; Min *et al.*, 2021] to construct dense correspondences between support-query pairs by calculating their similarities. However, the segmentation performance of these FSS methods will decrease rapidly when facing a large domain gap between the source and the target domain and cannot generalize well to the target domain.

Cross-domain Semantic Segmentation. Cross-domain semantic segmentation can be categorized into domain adaptive semantic segmentation (DASS) and domain generalized semantic segmentation (DGSS). DASS trains the model by jointly using source domain data and some labeled or unlabeled target domain data so that the model can quickly generalize well to the target domain. Recent works can be grouped into adversarial training and self-training approaches. The former [Hoffman *et al.*, 2017; Long *et al.*, 2017] aims to align the distributions of the source domain and the target domain via generative adversarial networks or Fourier transforms. The latter [Zou *et al.*, 2018a; Zou *et al.*, 2018b] is trained with pseudo-labels for the target domain. DGSS tries to bridge the domain gap through two main approaches, including Normalization and Whitening (NW), and Domain Randomization (DR). NW [Pan *et al.*, 2019; Peng *et al.*, 2022] normalizes the mean and standard deviation of the source data and whitens the covariance of the source data. DR [Peng *et al.*, 2021; Huang *et al.*, 2021] transforms source images into randomly stylized images and trains the network with them together. Unlike the setting of cross-domain semantic segmentation, CD-FSS not only has no access to target domain data during training, but also has disjoint label space between the source and target domains.

Cross-domain Few-shot Semantic Segmentation. Different from FSS, there are domain gaps between the source dataset and the target dataset, *i.e.*, both data feature distributions and label spaces in the meta-testing stage are different from the meta-training stage. Recently, some works have been proposed to address this task. For example, PixDA [Tavera *et al.*, 2021] proposes a pixel-by-pixel adversarial training strategy that uses a novel pixel-wise loss and discriminator to bridge the domain shift. To improve the generalization of the segmentation model, RTD [Wang *et al.*, 2022] designs a novel meta-memory module that transfers the intra-domain

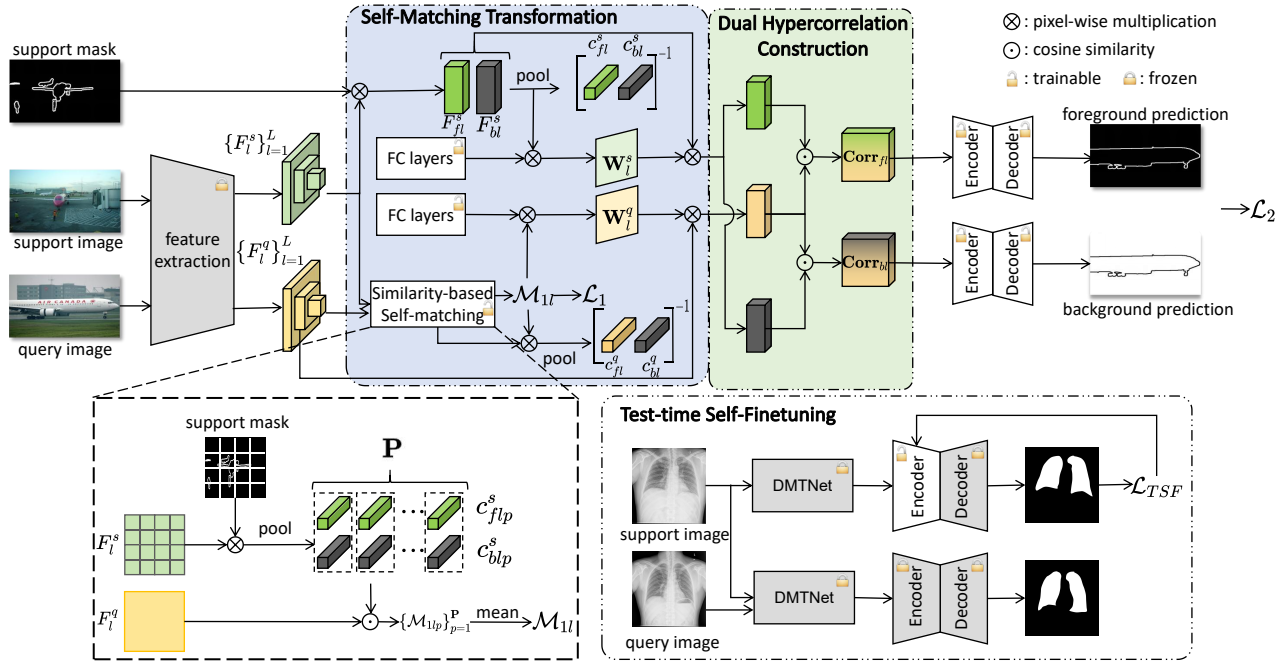

Figure 1: Overall architecture of the proposed DMTNet. After obtaining the pyramid features of support and query images, Self-Matching Transformation module (SMT) learns each image a self-adaptive transformation matrix, to transform its domain-specific features into domain-agnostic ones. Then, the Dual Hypercorrelation Construction (DHC) module is introduced to construct dense correlations between the query image with both the foreground and background of the support image. In the meta-testing stage, the Test-time Self-Finetuning (TSF) strategy fine-tunes a few parameters of the encoder to further improve the segmentation performance.

style information from the source domain images into target domain images. [Lu *et al.*, 2022] introduces a transductive fine-tuning method, which addresses the domain gap by using support labels to implicitly supervise query segmentation. PATNet [Lei *et al.*, 2022] establishes a new evaluation benchmark for CD-FSS and converts the domain-specific features to domain-agnostic ones to enhance generalization ability. Our proposed method tackles several key issues, including overfitting by only utilizing support information for feature transformation, intra-class appearance variances, and under-utilized information, which are overlooked by previous works.

3 Method

3.1 Problem Setting

The problem setting of CD-FSS can be formulated as follows. There is a source domain $(\mathcal{X}_s, \mathcal{Y}_s)$ and a target domain $(\mathcal{X}_t, \mathcal{Y}_t)$. $\mathcal{X}_s$ and $\mathcal{X}_t$ represent the input data distributions while $\mathcal{Y}_s$ and $\mathcal{Y}_t$ represent the label spaces. In CD-FSS, $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$, *i.e.*, the input data distribution of the source domain is different from that of the target domain, and the label spaces of the domains do not intersect. Based on the data distributions and the label spaces, we construct the training set $\mathcal{D}_{train}$ and testing set $\mathcal{D}_{test}$. In term of *N*-way *K*-shot segmentation, both $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ consist of a large number of episodes. Each episode contains a support set $\mathcal{S} = \{(\mathcal{I}_i^s, \mathcal{M}_i^s)\}_{i=1}^{N\times K}$ and a query set $\mathcal{Q} = \{(\mathcal{I}_i^q, \mathcal{M}_i^q)\}_{i=1}^{Q}$, where $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB image and $\mathcal{M} \in \mathbb{R}^{H\times W}$ is the binary mask. So $\mathcal{D}_{train} = \left\{\mathcal{I}_{\mathcal{S}/\mathcal{Q}}, \mathcal{M}_{\mathcal{S}/\mathcal{Q}}\right\}_{source}$ and $\mathcal{D}_{test} = \left\{\mathcal{I}_{\mathcal{S}/\mathcal{Q}}, \mathcal{M}_{\mathcal{S}/\mathcal{Q}}\right\}_{target}$. In meta-training stage, the model is trained on $\mathcal{D}_{train}$, without exposure to the target domain. After completing episodes training, the model segmentation performance is evaluated using $\mathcal{D}_{test}$ in the meta-testing stage.

3.2 Overview of DMTNet

The overall architecture of DMTNet is illustrated in Figure 1. DMTNet consists of two major functional modules to bridge the domain gap: the Self-Matching Transformation (SMT) module and the Dual Hypercorrelation Construction (DHC) module.

During the meta-training stage, the support set and query set are first fed to a shared convolutional neural network to extract multi-level pyramid features. Then, we exploit SMT to learn each support and query image a transformation matrix, and transform the domain-specific features into domain-agnostic ones. Subsequently, DHC constructs the support foreground-query and support background-query hypercorrelations based on the domain-agnostic features, respectively. Finally, two available modules, 4D convolutional pyramid encoder and 2D convolutional pyramid decoder [Min *et al.*, 2021] are adopted to obtain the predicted query mask based on the dual hypercorrelations.

During the meta-testing stage, two steps are involved in obtaining the final prediction mask. We design a novel Test-time Self-Finetuning (TSF) strategy in the first step, where only fine-tuning a few parameters of the network can significantly

refine the coarse mask in the second step and encourage the model to adapt quickly to the target domain.

3.3 Self-Matching Transformation

In cross-domain scenarios, domain style information in features hinders the model from accurately segmenting foreground objects. Therefore, ensuring the invariance of foreground objects while generalizing the domain style information can improve the generalization of representations. One useful strategy is to perform a feature transformation to transform the domain-relevant features into domain-irrelevant features. However, existing methods [Lei *et al.*, 2022] transform the support and query image features only based on the prototypes of the support image, which may pose potential issues for meta-testing. As the support set during the meta-testing stage has only a few images for each class, repeatedly utilizing the same support images may cause overfitting. In addition, according to [Fan *et al.*, 2022], there may exist a huge appearance difference between the support and query images, even if they belong to the same class. In this case, the transformation matrix derived only from the support image may not be useful for the query images. This motivates us to reduce the dependence on the support features and mine information from the query image itself during feature transformation. So we propose a Self-Matching Transformation (SMT) module. As shown in Figure 1, SMT consists of two stages. In the first stage, a Similarity-based Self-matching module is used to generate a rough segmentation mask, based on which the prototype of the query image is obtained. In the second stage, some learnable anchor layers are used to transform the support and query images' domain-specific features into domain-agnostic ones.

Similarity-based Self-matching. Inspired by [Fan *et al.*, 2022], we first generate coarse segmentation masks for the query images via similarity-based self-matching between the query features with the foreground and background prototypes of the support image. Since compressing a global feature map into a prototype vector will lose much detailed information, we propose to divide the support features into several local features and generate a more fine-grained predicted query mask by measuring the similarity between support local prototypes and query global features.

Specifically, for a 1-way 1-shot task, we first obtain the L-level pyramid features of the support and query images, *i.e.*, $\{F_l^s, F_l^q\}_{l=1}^{L}$, where $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. Then we calculate the support global foreground prototype $\boldsymbol{c}_{fl}^s \in \mathbb{R}^{C_l}$ at intermediate layer l with the support mask $\mathcal{M}^s \in \{0,1\}^{H \times W}$ via Masked Average Pooling (MAP), as follows:

$$\boldsymbol{c}_{fl}^s = \frac{\sum_{(x,y)} F_l^s(x,y)\delta_l\left[\mathcal{M}^s\right](x,y)}{\sum_{(x,y)} \delta_l\left[\mathcal{M}^s\right](x,y)}, \tag{1}$$

where (x, y) are spatial positions, and $\delta_l[\cdot]$ denotes the bilinear interpolation. For simplicity, we represent $\delta_l[\mathcal{M}^s]$ as $\mathcal{M}_l^s$. Similarly, the support global background prototype $\boldsymbol{c}_{bl}^s$ can be calculated in the same way. Then we divide the support feature map F_l^s into $\mathbf{P}$ support local feature maps, *i.e.*, $\left\{F_{lp}^s\right\}_{p=1}^{\mathbf{P}}$, $F_{lp}^s \in \mathbb{R}^{C_l \times \gamma H_l \times \gamma W_l}$. γ is the division ratio and we set it to 0.25. We also obtain the local support masks $\left\{\mathcal{M}_{lp}^s\right\}_{p=1}^{\mathbf{P}}$ following the same division. Then we calculate the p-th support local foreground prototype $\boldsymbol{c}_{flp}^s \in \mathbb{R}^{C_l}$ by MAP:

$$\boldsymbol{c}_{flp}^s = \frac{\sum_{(x,y)} F_{lp}^s(x,y)\mathcal{M}_{lp}^s(x,y)}{\sum_{(x,y)} \mathcal{M}_{lp}^s(x,y)}. \tag{2}$$

The p-th support local background prototype $\boldsymbol{c}_{blp}^s$ is similar to this. We then calculate the confidence matching correlation maps between query features and support local prototypes to obtain the naive query mask:

$$\mathcal{M}_{1l} = \frac{1}{\mathbf{P}} \sum_{p=1}^{\mathbf{P}} \left[\eta\left(\xi\left(\boldsymbol{c}_{flp}^s, F_l^q\right)\right), \eta\left(\xi\left(\boldsymbol{c}_{blp}^s, F_l^q\right)\right)\right], \tag{3}$$

where $\eta(\cdot)$ denotes the softmax function. $\xi(\cdot)$ denotes a similarity measure function. In this work, we use the cosine similarity. With $\mathcal{M}_{1l}$, the query foreground and background prototype $\boldsymbol{c}_{fl}^q$, $\boldsymbol{c}_{bl}^q$ can be calculated.

To make the predicted rough query mask as accurate as possible, and thus provide a most accurate transformation matrix for later adaptive feature transformation, we propose to use a binary cross-entropy (BCE) loss function as supervision here, formulated as:

$$\mathcal{L}_1 = \frac{1}{L} \sum_{l=1}^{L} \mathrm{BCE}\left(\mathcal{M}_{1l}, \delta_l\left[\mathcal{M}^q\right]\right). \tag{4}$$

Adaptive Feature Transformation. Similar to [Seo *et al.*, 2020; Lei *et al.*, 2022], we use a linear transformation as the transformation mapper. The difference is that we construct specialized transformation matrices for support and query features, respectively, to ensure the invariance of the foreground objects during the adaptive transformation process. We construct the support prototype matrix and the query prototype matrix as $\mathbf{C}_l^s = \left[\frac{\mathbf{c}_{fl}^s}{\|\mathbf{c}_{fl}^s\|}, \frac{\mathbf{c}_{bl}^s}{\|\mathbf{c}_{bl}^s\|}\right]$, $\mathbf{C}_l^q = \left[\frac{\mathbf{c}_{fl}^q}{\|\mathbf{c}_{fl}^q\|}, \frac{\mathbf{c}_{bl}^q}{\|\mathbf{c}_{bl}^q\|}\right]$, respectively. Similarly, two trainable anchor weight matrices are defined as $\mathbf{A}_l^s = \left[\frac{\mathbf{a}_{fl}^s}{\|\mathbf{a}_{fl}^s\|}, \frac{\mathbf{a}_{bl}^s}{\|\mathbf{a}_{bl}^s\|}\right]$, $\mathbf{A}_l^q = \left[\frac{\mathbf{a}_{fl}^q}{\|\mathbf{a}_{fl}^q\|}, \frac{\mathbf{a}_{bl}^q}{\|\mathbf{a}_{bl}^q\|}\right]$, where $\mathbf{a}_{fl}^s, \mathbf{a}_{bl}^s, \mathbf{a}_{fl}^q, \mathbf{a}_{bl}^q \in \mathbb{R}^{C_l}$. Similar to PATNet [Lei *et al.*, 2022], we only set three anchor layers for support and query features, respectively. They correspond to the feature maps of three dimensions, *i.e.*, low, medium and high-level features. So we construct two transformation matrices $[\mathbf{W}_l^s, \mathbf{W}_l^q]$ by solving $\mathbf{W}_l^s \mathbf{C}_l^s = \mathbf{A}_l^s$ and $\mathbf{W}_l^q \mathbf{C}_l^q = \mathbf{A}_l^q$. Since $\mathbf{C}_l^s$ and $\mathbf{C}_l^q$ are non-square matrices, we calculate their generalized inverse by $\mathbf{C}_l^{s+} = \left\{\mathbf{C}_l^{sT}\mathbf{C}_l^s\right\}^{-1}\mathbf{C}_l^{sT}$, $\mathbf{C}_l^{q+} = \left\{\mathbf{C}_l^{qT}\mathbf{C}_l^q\right\}^{-1}\mathbf{C}_l^{qT}$. In particular, we propose to refine $\mathbf{C}_l^{q+}$ by integrating $\mathbf{C}_l^{s+}$ into $\mathbf{C}_l^{q+}$:

$$\mathbf{C}_l^{q+} = \beta\mathbf{C}_l^{q+} + (1-\beta)\,\mathbf{C}_l^{s+}, \tag{5}$$

where β mainly controls the integration ratio, which is set to 0.5. Finally, the support and query transformation matrices at layer l can be calculated as $\mathbf{W}_l^s = \mathbf{A}_l^s\mathbf{C}_l^{s+}$, $\mathbf{W}_l^q = \mathbf{A}_l^q\mathbf{C}_l^{q+}$ respectively, where $\mathbf{W}_l^s, \mathbf{W}_l^q \in \mathbb{R}^{C^l \times C^l}$.

3.4 Dual Hypercorrelation Construction

Considering that objects from the same category are very likely to lie in similar environments, the background information correlations between the query and support images can also be used in CD-FSS. However, the existing method [Min *et al.*, 2021] directly filters out the support background region using the support mask. Differently, we propose a Dual Hypercorrelation Construction module (DHC) to explore the dense correlations between the query features with the foreground and background features of the support images in domain-agnostic space.

Firstly, we construct the 4D correlation tensor $\mathbf{Corr}_{fl} \in \mathbb{R}^{H_l \times W_l \times H_l \times W_l}$ based on the support foreground features F^s_{fl} and query features F^q_l by cosine similarity:

$$\mathbf{Corr}_{fl} = \mathrm{ReLU}\left(\frac{\mathbf{W}^s_l F^s_{fl}(x_1, y_1) \cdot \mathbf{W}^q_l F^q_l(x_2, y_2)}{\left\|\mathbf{W}^s_l F^s_{fl}(x_1, y_1)\right\| \cdot \left\|\mathbf{W}^q_l F^q_l(x_2, y_2)\right\|}\right), \quad (6)$$

where (x_1, y_1) and (x_2, y_2) denotes 2D spatial positions o support and query feature maps, respectively.

Secondly, we construct $\mathbf{Corr}_{bl}$ based on the support back ground features F^s_{bl} and query features F^q_l:

$$\mathbf{Corr}_{bl} = \mathrm{ReLU}\left(\frac{\mathbf{W}^s_l F^s_{bl}(x_1, y_1) \cdot \mathbf{W}^q_l F^q_l(x_2, y_2)}{\left\|\mathbf{W}^s_l F^s_{bl}(x_1, y_1)\right\| \cdot \left\|\mathbf{W}^q_l F^q_l(x_2, y_2)\right\|}\right) \quad (7)$$

Then, the dense correlation maps are fed to two availabl modules, 4D convolutional pyramid encoder and 2D convo lutional pyramid decoder [Min *et al.*, 2021] to generate th predicted query foreground mask $\mathcal{M}_f$ and background mas $\mathcal{M}_b$. The training supervision on the two predicted masks is

$$\mathcal{L}_2 = \mathrm{BCE}\left(\mathcal{M}_f, \mathcal{M}^q\right) + \alpha_1 \cdot \mathrm{BCE}\left(\mathcal{M}_b, 1 - \mathcal{M}^q\right), \quad (8)$$

where α_1 is a tuning weight. Finally, we train the model in an end-to-end manner by jointly optimizing all the losses:

$$\mathcal{L} = \alpha_2 \cdot \mathcal{L}_1 + \mathcal{L}_2, \quad (9)$$

where α_2 is a balancing hyperparameter. α_1 and α_2 are set as 1.0 and 0.5, respectively.

3.5 Test-time Self-Finetuning

In the meta-testing stage, we design a Test-time Self-Finetuning (TSF) strategy to refine query predictions in unseen domains. PATNet[Lei *et al.*, 2022] proposes to finetune the anchor layers by reducing the distribution distance between support foreground prototypes and the query foreground prototypes which are obtained by the predicted query mask. However, they assume images belonging to the same class have similar appearances, which is not always true in few-shot scenarios. We contend that there may exist significant appearance variances even within the same class. Therefore, aligning the foreground prototypes of the support and query images may cause overfitting to support images and distortion of query foreground features. To this end, we propose TSF to self-tune the network by trying to predict the ground-truth masks of the support images. By finetuning the network on support images, our model can learn style information of the target domain, leading to the generation of more accurate masks for the query images.

As shown in the Figure 1, TSF has two steps. In the first step, the model outputs the predicted support masks $\overline{\mathcal{M}^s}$ and updates the network with the loss:

$$\mathcal{L}_{TSF} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{BCE}\left(\overline{\mathcal{M}^s_k}, \mathcal{M}^s_k\right). \quad (10)$$

In the second step, we freeze the entire network and execute the final prediction for the query image.

Similar to PATNet, we do not finetune the whole network. Instead, we only finetune a few parameters of the encoder, which is validated by the quantitative experiments. The details can be found in Section 4.4.

4 Experiments

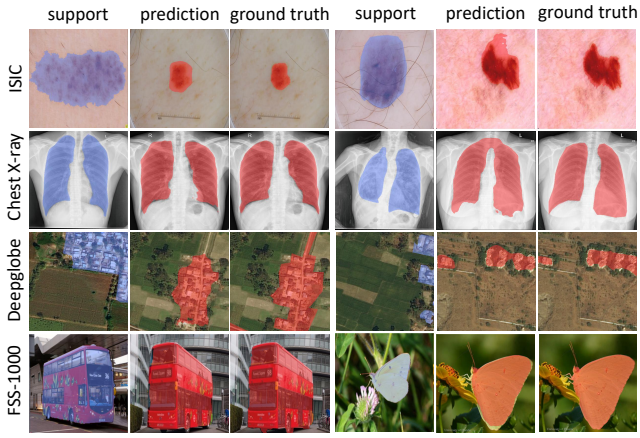

Figure 2: Qualitative results on the ISIC, Chest X-ray, Deepglobe, and FSS-1000 datasets under the 1-shot setting. The blue parts represent support masks and the red parts represent query masks and query predictions.

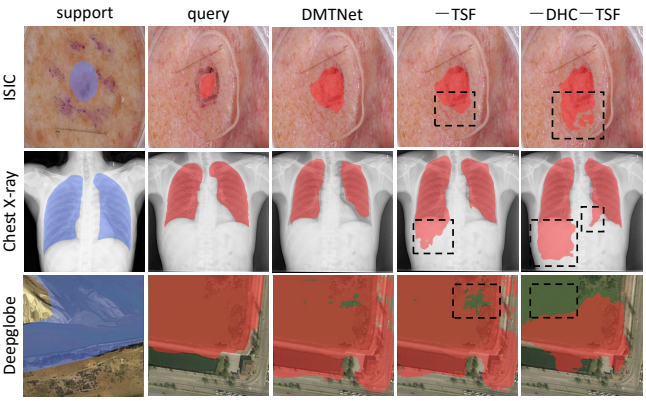

Figure 3: Qualitative results w.r.t. SMT, DHC, and TSF. The first two columns show the ground truth of the support and query images. The third column shows the predicted masks of DMTNet. The fourth column shows the prediction masks without the TSF. The last column shows the prediction masks without DHC and TSF modules.

Methods	Backbone	ISIC		Chest X-ray		Deepglobe		FSS-1000		Average	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Transfer Learning Methods											
Ft-last-1_{FCN}	VGG-16	15.17	19.75	33.63	48.08	29.80	32.25	32.51	53.62	27.78	38.43
Ft-last-2_{FCN}	VGG-16	17.52	21.65	36.35	53.85	32.90	35.34	32.15	57.44	29.82	42.07
Ft-last-3_{FCN}	VGG-16	17.91	25.58	45.61	56.05	32.91	35.54	33.32	60.86	32.34	44.51
1NN_{FCN}	VGG-16	15.68	23.66	46.26	52.70	32.42	38.63	41.51	46.64	33.97	40.41
Linear_{FCN}	VGG-16	15.51	30.65	37.69	50.07	33.56	38.75	41.09	49.16	31.96	42.16
Ft-last-$1_{Deeplab}$	ResNet-50	11.08	16.57	30.43	35.54	28.11	28.65	25.14	35.86	23.69	29.41
Ft-last-$2_{Deeplab}$	ResNet-50	10.22	17.56	31.16	51.57	24.09	36.74	20.68	42.50	21.29	37.10
1N-$N_{Deeplab}$	ResNet-50	21.44	26.04	47.76	57.93	32.28	35.96	45.81	55.95	36.82	43.97
$\text{Linear}_{Deeplab}$	ResNet-50	19.42	30.04	43.52	60.29	32.95	39.69	40.50	58.36	34.10	47.10
Few-shot Semantic Segmentation Methods											
AMP [Siam *et al.*, 2019]	VGG-16	28.42	30.41	51.23	53.04	37.61	40.61	57.18	59.24	43.61	45.83
PGNet [Zhang *et al.*, 2019a]	ResNet-50	21.86	21.25	33.95	27.96	10.73	12.36	62.42	62.74	32.24	31.08
PANet [Wang *et al.*, 2019]	ResNet-50	25.29	33.99	57.75	69.31	36.55	45.43	69.15	71.68	47.19	55.10
CaNet [Zhang *et al.*, 2019b]	ResNet-50	25.16	28.22	28.35	28.62	22.32	23.07	70.67	72.03	36.63	37.99
RPMMs [Yang *et al.*, 2020]	ResNet-50	18.02	20.04	30.11	30.82	12.99	13.47	65.12	67.06	31.56	32.85
PFENet [Tian *et al.*, 2020]	ResNet-50	23.50	23.83	27.22	27.57	16.88	18.01	70.87	70.52	34.62	34.98
RePRI [Boudiaf *et al.*, 2020]	ResNet-50	23.27	26.23	65.08	65.48	25.03	27.41	70.96	74.23	46.09	48.34
HSNet [Min *et al.*, 2021]	ResNet-50	31.20	35.10	51.88	54.36	29.65	35.08	77.53	80.99	47.57	51.38
Cross-domain Few-shot Semantic Segmentation Methods											
PATNet [Lei *et al.*, 2022]	VGG-16	33.07	45.83	57.83	60.55	28.74	34.83	71.60	76.17	47.81	54.35
PATNet [Lei *et al.*, 2022]	ResNet-50	41.16	**53.58**	66.61	70.20	<u>37.89</u>	42.97	78.59	81.23	<u>56.06</u>	<u>61.99</u>
RestNet [Huang *et al.*, 2023]	ResNet-50	<u>42.25</u>	51.10	70.43	73.69	22.70	29.99	<u>81.53</u>	<u>84.89</u>	54.23	59.92
DAM [Chen *et al.*, 2023]	ResNet-50	-	-	70.4	74.0	37.1	41.6	**84.6**	**86.3**	-	-
DMTNet	VGG-16	34.26	40.66	<u>73.02</u>	<u>75.84</u>	34.85	<u>47.77</u>	74.32	77.11	54.11	60.35
DMTNet	ResNet-50	**43.55**	<u>52.30</u>	**73.74**	**77.30**	**40.14**	**51.17**	81.52	83.28	**59.74**	**66.01**

Table 1: Performance of transfer learning, FSS, and CD-FSS methods in Mean-IoU under (1-way) 1-shot and (1-way) 5-shot settings. The best and second-best results are in bold and underlined, respectively.

	SMT	DHC	TSF	1-shot	△
DMTNet	√	√	√	**59.74**	0.0
-TSF	√	√		57.23	↓ 2.51
-DHC	√			56.36	↓ 3.35
-STM				47.57	↓ 12.17

Table 2: Ablation results of our proposed modules with 1-shot performance averaged over four datasets.

4.1 Experimental Setup

To fairly compare the cross-domain segmentation performance of DMTNet with PATNet [Lei *et al.*, 2022], we choose PASCAL VOC 2012 with SBD augmentation as the source domain, and ISIC2018, Chest X-ray, Deepglobe, and FSS-1000 as the target domains.

ISIC2018 is a skin cancer screening dataset consisting of lesion images. The dataset contains three types of skin lesions and a total of 2,596 images, each with one primary lesion region. The initial spatial resolution is approximately 1022 × 767 and we uniformly reduce it to 512 × 512.

Chest X-ray is a Tuberculosis X-ray image dataset with 566 annotated images, which is collected from 58 abnormal cases with a manifestation of Tuberculosis and 80 normal cases. The initial spatial resolution is 4020 × 4892 and we downsize it to 1024 × 1024.

Deepglobe is a satellite image dataset. Each image is densely annotated at pixel level with 7 categories: urban land, agricultural land, rangeland, forestland, water, barren land, and unknown. Following PATNet, we partition each image into 6 pieces and filter the single class images and the 'unknown' class, leading to a dataset of 5,666 images with a spatial resolution of 408 × 408.

FSS-1000 is a natural image dataset specialized for FSS, which contains 1,000 categories and each category has 10 annotated images. Each image contains only one segmentation target and has a resolution of 224 × 224.

We use the standard Intersection over Union (IoU) metric by averaging the prediction results across 5 runs with different random seeds. Each run is composed of 1,200 tasks for ISIC2018, Chest X-ray, and Deepglobe. And for FSS-1000, each run contains 2,400 tasks. We evaluate our model under 1-way 1-shot and 1-way 5-shot settings.

4.2 Implementation Details

Following previous works, we employ ResNet-50 and VGG-16 pre-trained on ILSVRC as our backbones. For ResNet-50, the features from the conv3_x, conv4_x, and conv5_x are extracted to produce feature maps. The channel dimensions of the three anchor layers are set to 512, 1024, and 2048, respectively. For the VGG-16 backbone, the features from the conv4_x to conv5_x are extracted to produce feature maps. The channel dimensions of the three anchor layers all are set to 512. In the meta-training stage, we use Adam optimizer to train DMTNet for 19 epochs with a learning rate of 1e-3. In the self-finetuning of the meta-testing stage, we use Adam optimizer with a learning rate of 1e-6 for ISIC2018, Deepglobe and FSS-1000, 1e-1 for Chest X-ray. All input

Ft_{low}	Ft_{mid}	Ft_{high}	$\text{Ft}_{encoder}$	$\text{Ft}_{decoder}$	Average
√					57.69
	√				58.72
		√			58.95
			√		**59.74**
				√	53.16

Table 3: Ablation results on the choice of fine-tuning parameters for TSF under 1-way 1-shot setting.

images are resized to 400 × 400 resolution.

4.3 Comparison with State-of-the-Art Methods

We compare the performance of our model with several state-of-the-art methods. These methods are categorized into three groups: transfer learning, few-shot semantic segmentation, and cross-domain few-shot semantic segmentation methods. All methods are trained on PASCAL VOC and tested on the four datasets. The experimental results are shown in Table 1. Because the whole ISIC is seen as one class in DAM [Chen *et al.*, 2023], we do not compare their report results on ISIC here (under the same ISIC dataset setting, our model exceeds DAM by 2.25% in the 5-shot setting). We can see that under both (1-way) 1-shot and (1-way) 5-shot settings, the performance of our model ranks at the top on the average results of the four datasets, reaching 59.74% MIoU in the 1-shot setting and 66.01% MIoU in the 5-shot setting. We show the 1-way segmentation visualization results of our model on four datasets in Figure 2.

When compared with the state-of-the-art PATNet [Lei *et al.*, 2022], our model exceeds PATNet by 3.68% and 4.02% in the 1-shot and 5-shot settings respectively. This suggests that using query-specific transformation matrices and the dual hypercorrelation construction can avoid overfitting to base classes, leading to further enhancement of the segmentation results. In addition, when using VGG-16 as the backbone, our model significantly improves performance, leading to the second-best results on Chest X-ray and Deepglobe.

4.4 Ablation Study

Component Analysis. We conduct several ablation experiments to verify the effectiveness of the three key modules of DMTNet, *i.e.* SMT, DHC and TSF. Table 2 shows the impact of each module on the model performance. We can see that using all three modules proposed in this paper achieves the best results, and removing any of them would lead to a drop in the average performance across four datasets. Figure 3 further shows the segmentation visualization results of removing the TSF and the TSF+DHC modules, demonstrating the impact of our proposed module in a more intuitive way.

Fine-tuning Parameters for TSF. We conduct quantitative experiments to select the fine-tuning parameters for the test-time self-finetuning strategy. Following PATNet [Lei *et al.*, 2022], we choose the low-, medium- and high-level anchor layers. Additionally, we believe that the encoder/decoder should have an important impact on the cross-domain segmentation performance as it hierarchically fuses/reconstructs the 4D correlation maps, therefore we also choose the two fusing/de-fusing layers in the encoder/decoder. The results are presented in Table 3. We can observe that fine-tuning all but the decoder layers can lead to performance improvement, demonstrating the effectiveness of the TSF. Fine-tuning the encoder layers achieves the best performance of 59.45%, indicating the effectiveness of bringing target domain information into the encoder for better fusion of correlation maps.

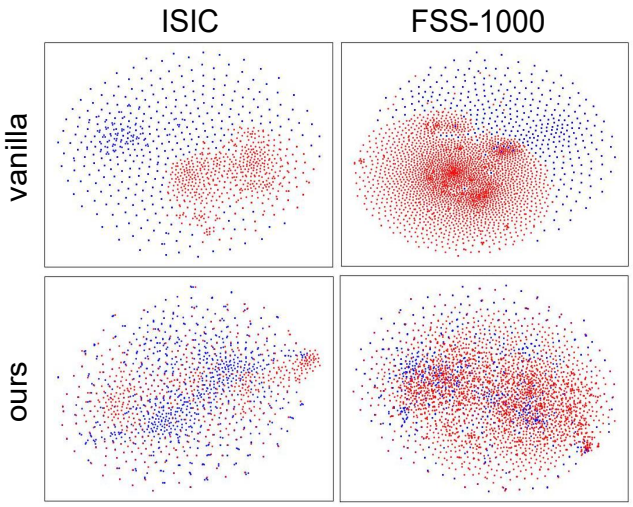

Figure 4: Visualization results w.r.t. SMT. The first and second rows represent the feature distributions before and after applying SMT, respectively. The red dots represent the PASCAL VOC dataset and the blue dots represent the ISIC or FSS-1000 datasets.

Visualization w.r.t. SMT. We visualize the image feature distributions before and after SMT to more clearly illustrate the effect of SMT in bridging the domain gaps, as shown in Figure 4. We show the feature distributions of ISIC and FSS-1000 datasets for comparison with PASCAL VOC. We can see that the feature distributions of ISIC and FSS-1000 are closer to the PSCAL VOC after using SMT, which proves the effectiveness of our module in bridging domain distances and better generalization to the target domains.

5 Conclusion

We propose DMTNet for cross-domain few-shot semantic segmentation. DMTNet first exploits an SMT module to calculate each support and query image a transformation matrix based on its own prototype, and transform their domain-specific features into domain-agnostic ones self-adaptively. Then, a DHC module is used to explore the dual hypercorrelation between the query image with the foreground and background of the support image in the domain-agnostic space, based on which a foreground and background prediction mask are obtained and supervised during meta-training, respectively. During meta-testing, a TSF strategy is used to further improve the segmentation performance by only refining a few parameters of DMTNet. Extensive experiments show that DMTNet is effective and achieves state-of-the-art performance on four datasets with different domain gaps.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62276129 & No. 62206127), and the Natural Science Foundation of Jiangsu Province (No. BK20220890).

References

[Boudiaf *et al.*, 2020] Malik Boudiaf, Hoel Kervadec, Imtiaz Masud Ziko, Pablo Piantanida, Ismail Ben Ayed, and José Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13974–13983, 2020.

[Candemir *et al.*, 2014] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul Kumar Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Kiran Antani, George R. Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33:577–590, 2014.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017.

[Chen *et al.*, 2023] Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, Yingming Li, Jungong Han, and Zhongfei Zhang. Dense affinity matching for few-shot segmentation, 2023.

[Codella *et al.*, 2019] Noel C. F. Codella, Veronica M Rotemberg, Philipp Tschandl, M. E. Celebi, Stephen W. Dusza, David 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.

[Demir *et al.*, 2018] 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209, 2018.

[Fan *et al.*, 2022] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European Conference on Computer Vision*, 2022.

[Hoffman *et al.*, 2017] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *ArXiv*, abs/1711.03213, 2017.

[Huang *et al.*, 2021] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6887–6898, 2021.

[Huang *et al.*, 2023] Xinyang Huang, Chuang Zhu, and Wenkai Chen. Restnet: Boosting cross-domain few-shot segmentation with residual transformation network, 2023.

[Lei *et al.*, 2022] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European Conference on Computer Vision*, 2022.

[Li *et al.*, 2021] Gen Li, V. Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8330–8339, 2021.

[Liu *et al.*, 2020a] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4164–4172, 2020.

[Liu *et al.*, 2020b] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. *ArXiv*, abs/2007.06309, 2020.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[Long *et al.*, 2017] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Neural Information Processing Systems*, 2017.

[Lu *et al.*, 2022] Yuhang Lu, Xinyi Wu, Zhenyao Wu, and Song Wang. Cross-domain few-shot segmentation with transductive fine-tuning. *ArXiv*, abs/2211.14745, 2022.

[Min *et al.*, 2021] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmenation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6921–6932, 2021.

[Okazawa, 2022] Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *European Conference on Computer Vision*, 2022.

[Pan *et al.*, 2019] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1863–1871, 2019.

[Peng *et al.*, 2021] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021.

[Peng *et al.*, 2022] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2595, 2022.

[Rakelly *et al.*, 2018] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *International Conference on Learning Representations*, 2018.

[Richter *et al.*, 2016] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *ArXiv*, abs/1608.02192, 2016.

[Ros *et al.*, 2016] Germán Ros, Laura Sellart, Joanna Materzynska, David Vázquez, and Antonio M. López. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[Seo *et al.*, 2020] Jun Seo, Younghyun Park, Sung Whan Yoon, and Jaekyun Moon. Task-adaptive feature transformer for few-shot segmentation. *ArXiv*, abs/2010.11437, 2020.

[Shaban *et al.*, 2017] Amirreza Shaban, Shray Bansal, Z. Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *ArXiv*, abs/1709.03410, 2017.

[Siam *et al.*, 2019] Mennatullah Siam, Boris N. Oreshkin, and Martin Jägersand. Amp: Adaptive masked proxies for few-shot segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5248–5257, 2019.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, 2017.

[Tavera *et al.*, 2021] A. Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1959–1968, 2021.

[Tian *et al.*, 2020] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1050–1065, 2020.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Neural Information Processing Systems*, 2016.

[Wang *et al.*, 2019] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9196–9205, 2019.

[Wang *et al.*, 2022] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7055–7064, 2022.

[Wei *et al.*, 2019] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2866–2875, 2019.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems*, 2021.

[Yang *et al.*, 2020] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. *ArXiv*, abs/2008.03898, 2020.

[Zhang *et al.*, 2018] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50:3855–3865, 2018.

[Zhang *et al.*, 2019a] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9586–9594, 2019.

[Zhang *et al.*, 2019b] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5212–5221, 2019.

[Zhang *et al.*, 2021] Gengwei Zhang, Guoliang Kang, Yunchao Wei, and Yi Yang. Few-shot segmentation via cycle-consistent transformer. In *Neural Information Processing Systems*, 2021.

[Zhao *et al.*, 2016] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.

[Zou *et al.*, 2018a] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. *ArXiv*, abs/1810.07911, 2018.

[Zou *et al.*, 2018b] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision*, 2018.