

Cross-Domain Few-Shot Semantic Segmentation

Shuo Lei¹, Xuchao Zhang², Jianfeng He¹, Fanglan Chen¹, Bowen Du³ *, and Chang-Tien Lu¹

¹ Department of Computer Science, Virginia Tech, Falls Church, VA, USA

² NEC Laboratories America, Princeton, NJ, USA

³ State Key Laboratory of Software Development Environment, Beihang University, Beijing, China







Abstract. Few-shot semantic segmentation aims at learning to segment a novel object class with only a few annotated examples. Most existing methods consider a setting where base classes are sampled from the same domain as the novel classes. However, in many applications, collecting sufficient training data for meta-learning is infeasible or impossible. In this paper, we extend few-shot semantic segmentation to a new task, called Cross-Domain Few-Shot Semantic Segmentation (CD-FSS), which aims to generalize the meta-knowledge from domains with sufficient training labels to low-resource domains. Moreover, a new benchmark for the CD-FSS task is established and characterized by a task difficulty measurement. We evaluate both representative few-shot segmentation methods and transfer learning based methods on the proposed benchmark and find that current few-shot segmentation methods fail to address CD-FSS. To tackle the challenging CD-FSS problem, we propose a novel Pyramid-Anchor-Transformation based few-shot segmentation network (PATNet), in which domain-specific features are transformed into domain-agnostic ones for downstream segmentation modules to fast adapt to unseen domains. Our model outperforms the state-of-the-art few-shot segmentation method in CD-FSS by 8.49% and 10.61% average accuracies in 1-shot and 5-shot, respectively. Code and datasets are available at <https://github.com/slei109/PATNet>

Keywords: Few-Shot Learning, Cross-Domain Transfer Learning, Semantic Segmentation

1 Introduction

Deep neural networks for semantic segmentation, such as FCN [26], DeepLab [5] and PSPNet [52], typically require large-scale annotations for training, which is costly to obtain. To reduce such burden on data annotation, *Few-Shot Semantic Segmentation (FSS)* task has been proposed [33], which aims to learn a model that can perform segmentation on novel classes with only a few pixel-level annotated images. Although significant progress has been made in the FSS task [34,45,46,49,50], it is hard to apply existing methods to cross-domain scenarios.

* Corresponding author.

Task	Data Access during Training	Training and Testing Dataset	Example	
			Training Pair	Testing Pair
Cross-domain Segmentation	$X_s + X_t$	$X_s \neq X_t$ $Y_s = Y_t$		
Few-shot Segmentation	X_s	$X_s = X_t$ $Y_s \neq Y_t$		
Cross-domain Few-shot Segmentation	X_s	$X_s \neq X_t$ $Y_s \neq Y_t$		

large-scale training
 meta-training
 testing

Fig. 1. Differences between the cross-domain few-shot segmentation and existing tasks. X_s and X_t denote the data distribution in the source and target domain, respectively. Y_s represents the source label space and Y_t represents the target label space.

Since the methods still require a large number of base class samples for training, it is infeasible for low-resource domains where few training annotations can be obtained. For instance, it is too expensive to collect sufficient satellite images for meta-training purposes, remaining a large obstacle to applying the few-shot segmentation methods directly into the satellite image domain. To tackle the issue, we extend FSS to a new Cross-Domain Few-Shot Segmentation (CD-FSS) task that aims at generalizing the meta-knowledge from domains with sufficient training labels (*e.g.* PASCAL VOC [13]) to low-resource domains.

The conceptual comparisons between the existing tasks and our CD-FSS task are shown in Fig. 1. First, most works on cross-domain semantic segmentation (or domain adaptation for semantic segmentation) focus on the problem setting where the target domain data can be *accessed* during training and share the *same* label space as the source domain. For example, in the first row of Fig. 1, street photo-realistic synthetic images are usually used as training data for real-world urban scene understanding tasks. In contrast, we study the CD-FSS problem, where the source and target domains have completely *disjoint* label space and *cannot* access target domain data during the training stage. Second, the classic few-shot semantic segmentation only focuses on segmenting novel classes sampled from the *same* domain in the training stage. In other words, the input data distributions from source and target domains are the *same* while the label spaces are *disjoint* in the training and testing stages. In contrast, both data distributions and label spaces in the testing stage are *different* from the training stage in the CD-FSS task.

In this paper, we establish a new benchmark for the CD-FSS task to evaluate the cross-domain generalization ability of segmentation models under different domain gaps. It consists of four different domains characterized by domain shifts of different size: FSS-1000 [23], Deepglobe [11], ISIC2018 [10,42], and Chest X-ray datasets [4,21]. These datasets cover daily objects images, satellite images, dermoscopic images of skin lesions, and X-ray images, respectively. The selected datasets have class diversity and reflect the real-world scenario for few-shot semantic segmentation.

Furthermore, both representative few-shot segmentation methods and transfer learning based methods are evaluated on the proposed benchmark. Experiment results show that: 1) the performances of existing few-shot semantic segmentation methods degrade significantly under large domain shifts. Those methods even underperform the simple transfer learning baselines when the target domain is drastically different from the source domain; 2) meta-learning approaches are more effective than all transfer learning baselines in the setting of limited domain differences.

A major challenge in CD-FSS is that the feature space learned from the source domain cannot be applied to the target domain. Concretely, existing methods learn a support-query matching/comparing model in a single domain and their basic assumption is that the pretrained encoder is powerful enough to embed the image into distinguishable features. However, the backbone only pretrained in the source domain fails in the target domain due to the different data distribution. To address this problem, we propose a novel Pyramid Anchor-based Transformation Module (PATM) to transform the domain-specific features into domain-agnostic ones. Thus, the downstream model can be well adapted to the novel domains by matching domain-agnostic features of support and query sets to make the segmentation. To further refine the predicted mask of the query image, we also propose a Task-adaptive Fine-tuning Inference (TFI) strategy for fast adaptation to unseen domain. In the testing phase, only PATM is updated with the prototype similarity between support images and query predictions to avoid over-fitting in few-shot scenarios. In this way, the predicted mask is refined with the calibrated features produced by the fine-tuned PATM.

Our main contributions are summarized as follows:

- We extend few-shot semantic segmentation to a new task, called Cross-Domain Few-Shot Semantic Segmentation (CD-FSS), which aims to segment a novel object class in *unseen domains* with only *a few* annotated examples.
- A practical evaluation benchmark for CD-FSS is established, consisting of four different domains. We also measure the task difficulty for each domain according to 1) domain shift and 2) discrimination between foreground and background classes.
- We propose a Pyramid Anchor-based Transformation Module (PATM) to transform the domain-specific features into domain-agnostic ones. Downstream segmentation modules can be adapted to unseen domains by learning with domain-agnostic features. A novel Task-adaptive Fine-tuning Inference (TFI) strategy is proposed to refine the prediction in unseen domains.
- We investigate a practical evaluation of few-shot segmentation methods and transfer learning based methods in the proposed benchmark. Results show that current few-shot segmentation methods fail to address CD-FSS and are even inferior to the transfer learning baseline methods when a large domain gap exists. In contrast, Our model outperforms the state-of-the-art few-shot segmentation method in CD-FSS by 8.49% and 10.61% average accuracies in 1-shot and 5-shot, respectively.

2 Related Work

The prior works related to this paper are summarized below in domain adaptation for semantic segmentation, cross-domain few-shot learning and few-shot semantic segmentation.

Domain adaptation for semantic segmentation. Recent works in domain adaptation for semantic segmentation are mainly divided into two directions. One group of studies aims to learn domain-invariant representations of instances by domain adversarial training [8,9,12,41]. Hoffman et al. [20] combine global and local alignment methods with adversarial training. Similar ideas are also explored using different techniques, such as distillation loss [9], output space alignment [40], class-balanced self-training [54], conservative loss [53], etc. The other group is learning from a pre-defined curriculum [31,51].

However, these methods operate in the setting where the target domain data can be *accessed* during training to drive the model adaptation and compensate for the domain shift. In addition, most existing works exploit photo-realistic synthetic data. Thus, the source and target domain share the *same* label space and still retain a high degree of visual similarity. In contrast, we study the cross-domain few-shot semantic segmentation problem, where the source and target domains have completely *disjoint* label space and *cannot* require target domain data during the training stage. The goal of this work is to learn a task-adaptive few-shot semantic segmentation model under large domain shifts.

Few-shot learning. Few-shot learning aims to learn a new concept representation from only a few annotated examples. Most existing works can be categorized into metric learning methods [44,37,35], gradient-based meta learners [29,14], and graph neural network [15,24] based methods. Yoon et al. [47] introduce a reference vector set to construct a linear transformer that performed task-specific null-space projection for classification, which is the theoretical basis of our method. In cross-domain few-shot learning [39,7,43], both data distribution and the label space in the meta-testing stage are different from the meta-training stage. Tseng et al. [43] propose feature-wise transformation layers to improve the generalization of metric-based few-shot classification approaches to unseen domains. Guo et al. [16] propose a harder cross-domain few-shot benchmark (BSCD-FSL), where there is a large shift between base and novel class domains. It covers several target domains with varying similarities to natural images. Our proposed benchmark can be seen as an extension of BSCD-FSL in the few-shot segmentation task to evaluate the cross-domain generalization ability of few-shot segmentation models under different domain shifts.

Few-shot semantic segmentation. In contrast to the domain adaptation for semantic segmentation, few-shot semantic segmentation has no access to the target domain during training stage. It aims at segmenting novel semantic objects in an image with only a few densely annotated examples. Based on the optimized module in the meta-training process, existing works can be divided into two groups, metric-based and relation-based methods. Specifically, metric-based methods (e.g. PANet [45] and AMP [34]) adopt non-parametric decoder and aim to train the encoder to construct a consistent metric space. In contrast,

relation-based methods (e.g. CaNet [49], RPMM [46], PGNet [48], PFENet [38] and HSNet [27]) freeze the pre-trained encoder during training process and train a decoder to compare the support and query samples. In other words, metric-based methods focus on separating foreground and background classes in each task, while relation-based methods focus on recognizing the foreground classes based on the pre-trained features. RePRI [3] foregoes meta-learning and adopts a transductive inference with a feature extraction trained on the base classes. However, these methods only focus on segmenting novel classes sampled from the same domain. They fail to generalize to unseen domains due to large discrepancy of the feature distribution across domains.

3 Benchmark

The proposed benchmark for CD-FSS consists of four datasets characterized by domain shifts of different sizes. The proposed benchmark includes images and pixel-level annotations from FSS-1000 [23], Deepglobe [11], ISIC2018 [10,42], and Chest X-ray datasets [4,21]. These datasets cover daily objects images, satellite images, dermoscopic images of skin lesions, and X-ray images, respectively. The selected datasets have class diversity and reflect the real-world scenario for few-shot semantic segmentation tasks. To provide a better overview, in Table 1, the task difficulty for each domain is measured from two aspects: 1) domain shift (cross the datasets) and 2) class distinction in a single image (within the dataset). Fréchet Inception Distance (FID) [19] is adopted to measure the domain shift [1] of these four datasets with respect to the PASCAL [13]. Since the discrimination between classes in a single image has an important impact on the segmentation task, we measure the similarity between foreground and background classes using KL-divergence. For more details, please refer to the supplementary material.

Table 1. Conceptual difference between PASCAL and the four cross-domain datasets. The domain shift and class distinction in a single image is measured by FID and DisFB, respectively.

Dataset	perspective distortion	natural content	color depth	FID	DisFB
Deepglobe	×	×	3	213.58	0.143
ISIC	×	×	3	275.28	0.187
Chest X-ray	×	×	1	316.56	0.126
FSS-1000	✓	✓	3	238.41	0.112

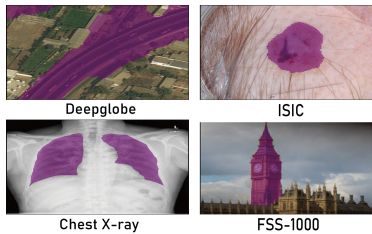


Fig. 2. Example of segmentation in the benchmark.

FSS-1000 [23] is a natural image dataset for few-shot segmentation, consisting of 1,000 object classes and each class has 10 samples. The official split for semantic segmentation is used in our experiment. We report the results on the official testing set, which contains 240 classes and 2,400 testing images.

Deepglobe [11] is a satellite image dataset. Each image is densely annotated at pixel-level with 7 categories: areas of urban, agriculture, rangeland, forest, water, barren, and unknown. As the ground-truth label is only available in the

training set, thus we adopt the official training set to report the results, which contains 803 images. The images have a fixed spatial resolution of 2448×2448 pixels. To increase the number of testing images and reduce the size of images, we cut each image into 6 pieces. As the categories labeled in this dataset have no regular shape, the cutting operation has little effect on the segmentation. After filtering the single class images and the ‘unknown’ class, we get 5,666 images to report the results and each image has 408×408 pixels.

ISIC2018 [10,42] is a dataset on lesion images, consisting of skin cancer screening samples. Every lesion image contains exactly one primary lesion. As the ground-truth label is only available in the training set, thus we report the results on the official training set, containing 2,596 images. The images have a spatial resolution around 1022×767 . As a common practice we down-size the images to 512×512 pixels.

Chest X-ray [4,21] is an X-ray image dataset for Tuberculosis. It includes 566 images with a resolution of 4020×4892 , which are collected from 58 cases with a manifestation of Tuberculosis and 80 normal cases. Due to the large size of image, we down-size the images to 1024×1024 pixels as a common practice.

4 Problem Setting

The cross-domain few-shot semantic segmentation (CD-FSS) problem can be formalized as follows. We have a source domain $(\mathcal{X}_s, \mathcal{Y}_s)$ and a target domain $(\mathcal{X}_t, \mathcal{Y}_t)$, where \mathcal{X} is the input data distribution and \mathcal{Y} is the label space. In CD-FSS, the input data distribution in source domains \mathcal{X}_s is different from target domains and the label space in source domains has no overlap with target domains \mathcal{X}_t , i.e., $\mathcal{X}_s \neq \mathcal{X}_t$, $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$.

Suppose that the model is trained on the source domain, CD-FSS aims to use the trained model to perform segmentation on the novel classes in the target domain with only a few annotated images per class. The training set \mathcal{D}_{train} is constructed from $(\mathcal{X}_s, \mathcal{Y}_s)$ and the testing set \mathcal{D}_{test} is constructed from $(\mathcal{X}_t, \mathcal{Y}_t)$. We align training and testing with the episodic paradigm [44] to handle the few-shot scenario. Specifically, given a N -way K -shot learning task, both the training set \mathcal{D}_{train} and testing set \mathcal{D}_{test} consist of several episodes. Each episode is constructed by 1) a support set $\mathcal{S} = \{(\mathbf{I}_i^s, \mathbf{M}_i^s)\}_{i=1}^{N \times K}$ and 2) a query set $\mathcal{Q} = \{(\mathbf{I}_i^q, \mathbf{M}_i^q)\}_{i=1}^Q$, where \mathbf{I} is an image, \mathbf{M} is a corresponding mask and Q is the number of query samples. Note that the model is trained on \mathcal{D}_{train} from the source domain and has *no access* to the target domain data. During the testing (or meta-testing) process, the model is presented with a support set and a query set from the target domain is used to evaluate the model performance.

5 Model

The main challenge in CD-FSS is to reduce the performance degradation brought by domain shifts. Previous works focus on learning a support-query matching

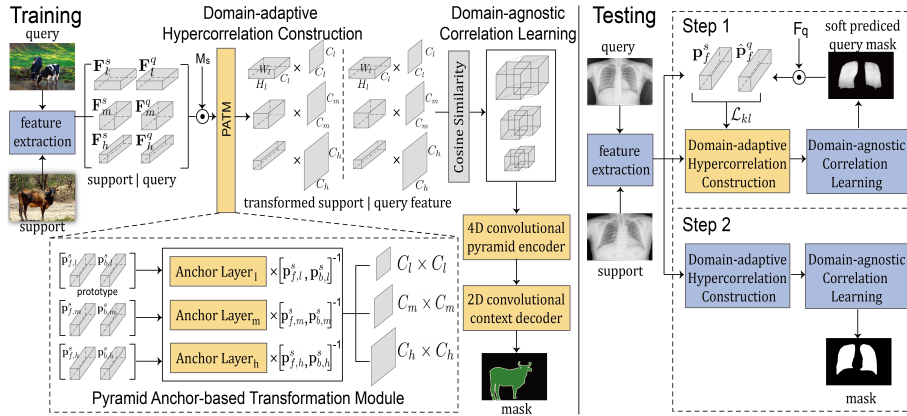


Fig. 3. Overview of our method in a 1-way 1-shot example. After obtaining the pyramid features of support and query images, PATM is introduced to transform the domain-specific hypercorrelations into domain-agnostic ones by producing linear transformation matrices. Then, the transformed features are fed into Domain-agnostic Correlation Learning part for the final query segmentation mask prediction. In the testing phase, the anchor layers are fine-tuned with the foreground prototype similarity between support and query predictions. Yellow parts are trainable and blue parts are frozen.

model and their basic assumption is that the pretrained encoder is powerful enough to embed the image into distinguishable features for the downstream matching model. However, the backbone only pretrained in the source domain fails in the target domain, especially under the large domain gap, like daily life object images to X-ray images. To address the problem, our model learns to transform the domain-specific features into domain-agnostic ones. In this way, the downstream model can be well adapted to the novel domain by matching domain-agnostic features of support and query sets to make the segmentation.

As shown in Fig. 3, our method consists of three major parts, feature extraction backbone, domain-adaptive hypercorrelation construction and domain-agnostic correlation learning. Given support and query images, we first extract all the intermediate features with feature extractor. Then, we introduce a particularly novel module in the Domain-adaptive Hypercorrelation Construction part, dubbed Pyramid Anchor-based Transformation Module (PATM), to transform the domain-specific features into domain-agnostic ones. Next, we compute multi-level correlation maps with all transformed feature maps to feed into Domain-agnostic Correlation Learning part. Two off-the-shelf modules, 4D convolutional pyramid encoder and 2D convolutional context decoder [27], are adopted to produce the prediction mask in a coarse-to-fine manner with efficient 4D convolutions. In the testing phase, we also propose a Task-adaptive Fine-tuning Inference (TFI) strategy to encourage the model to fast adapt to the target domain

by fine-tuning PATM with \mathcal{L}_{kl} loss, which measures the foreground prototype similarity between support and query predictions.

5.1 Pyramid Anchor-based Transformation Module

The core idea of Pyramid Anchor-based Transformation Module (PATM) aims at learning pyramid anchor layers to transform the domain-specific features into domain-agnostic ones. Intuitively, if we can find a transformer to transform the domain-specific features into a domain-agnostic metric space, it will reduce the detrimental effects brought by the domain drift. Since the domain-agnostic metric space is constant, it will be much easier for the downstream segmentation modules to make predictions in such a stable space.

Ideally, features belonging to the same class will produce similar results when they are transformed in the same way. Thus, if we transform the support features to the corresponding anchor points in the domain-agnostic space, then by using the same transformation, we can also make query features belonging to the same class transform close to the anchor points in the domain-agnostic space. Inspired by TAF-T module [32], we adopt a linear transformation matrix as the transformation mapper since it introduces fewer learnable parameters. As shown in Fig. 3, we use the anchor layer and the prototype set of the support image to compute the transformation matrix. Let \mathbf{A} represent the weight matrix of the anchor layer and \mathbf{P} denote the prototype matrix of the support image. We construct the transformation matrix \mathbf{W} by finding a matrix such that $\mathbf{W}\mathbf{P} = \mathbf{A}$.

Specifically, for an 1-way 1-shot task, once the intermediate feature maps in L layers of the support image, $\{\mathbf{F}_l^s\}_{l=1}^L$, are obtained, we can calculate the foreground prototype of each feature map $\mathbf{F}_l^s \in \mathbb{R}^{C_l \times H_l \times W_l}$ with the support mask $\mathbf{M}^s \in \{0, 1\}^{H \times W}$ via masked average pooling, i.e. $\mathbf{p}_{f,l}^s = \frac{\sum_i \mathbf{F}_l^s \zeta_l(\mathbf{M}^s)_i}{\sum_i \zeta_l(\mathbf{M}^s)_i}$, where $\mathbf{p}_{f,l}^s \in \mathbb{R}^{C_l}$ and i is 2D spatial positions of the feature map. $\zeta_l(\cdot)$ denotes a function that bilinearly interpolates input tensor to the spatial size of the feature map \mathbf{F}_l^s at intermediate layer l by expanding along channel dimension, $\zeta_l: \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{C_l \times H_l \times W_l}$. Similarly, the background prototype $\mathbf{p}_{b,l}^s$ for \mathbf{F}_l^s can be obtained in the same way and the prototype matrix \mathbf{P}_l^s is defined as $[\frac{\mathbf{p}_{f,l}^s}{\|\mathbf{p}_{f,l}^s\|}, \frac{\mathbf{p}_{b,l}^s}{\|\mathbf{p}_{b,l}^s\|}]$. Accordingly, the anchor weight matrix \mathbf{A}_l is defined as $[\frac{\mathbf{a}_{f,l}}{\|\mathbf{a}_{f,l}\|}, \frac{\mathbf{a}_{b,l}}{\|\mathbf{a}_{b,l}\|}]$, where $\mathbf{a}_{\cdot,l} \in \mathbb{R}^{C_l}$. In general, \mathbf{P}_l^s is a non-square matrix and we can calculate its generalized inverse [2] with $\mathbf{P}_l^{s+} = \{\mathbf{P}_l^{sT} \mathbf{P}_l^s\}^{-1} \mathbf{P}_l^{sT}$. Thus, the transformation matrix at intermediate layer l is computed as $\mathbf{W}_l = \mathbf{A}_l \mathbf{P}_l^{s+}$, where $\mathbf{W}_l \in \mathbb{R}^{C_l \times C_l}$.

For the subsequent hypercorrelation construction, a pair of transformed query and masked support features $\hat{\mathbf{F}}_l^s$ at each layer forms a 4D correlation tensor $\mathbf{C}_l \in \mathbb{R}^{H_l \times W_l \times H_l \times W_l}$ using cosine similarity:

$$\mathbf{C}_l(i, j) = \text{ReLU} \left(\frac{\mathbf{W}_l \mathbf{F}_l^q(i) \cdot \mathbf{W}_l \hat{\mathbf{F}}_l^s(j)}{\|\mathbf{W}_l \mathbf{F}_l^q(i)\| \|\mathbf{W}_l \hat{\mathbf{F}}_l^s(j)\|} \right) \quad (1)$$

where i and j denote 2D spatial positions of \mathbf{F}_l^q and $\hat{\mathbf{F}}_l^s$, respectively.

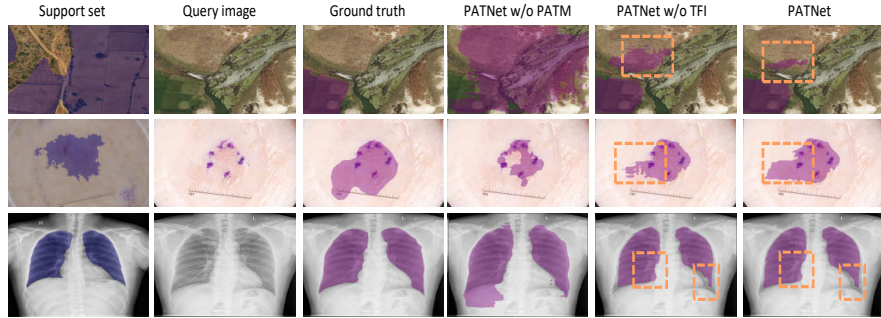


Fig. 4. Visual comparison results of several 1-shot tasks. For each task, the first three columns show the ground truth of support and query sets. The next two columns represent the prediction mask without anchor layers and the prediction mask without fine-tuning, respectively. The last column shows the final predicted segmentation after fine-tuning with \mathcal{L}_{kl} . Best viewed in colors.

To avoid adding too many learnable parameters, we set three anchor layers for low-, medium- and high-level feature maps respectively. Note that only three anchor layers are introduced for different feature dimensions. Even though feature maps with the same dimension share one anchor layer, each of them still can obtain its unique transformation matrix with its own prototype set.

5.2 Task-adaptive Fine-tuning Inference

To further refine the prediction mask of query images, we propose a Task-adaptive Fine-tuning Inference (TFI) strategy for fast adaptation to unseen domains in the testing phase. The motivation is that if the model can predict a good segmentation mask for the query image, the foreground class prototype of the segmented query image should be similar to that of the support set. Different from optimizing all parameters in the model, we only fine-tune the anchor layers to avoid overfitting in few-shot scenarios. Fig. 3 shows the pipeline of the strategy. In the testing phase, during step 1, only anchor layers are updated accordingly using the proposed \mathcal{L}_{kl} , which measures the similarity between the foreground class prototype of support and query sets. In step 2, all layers in the model are frozen and make the final prediction for query images. In this way, the model is encouraged to fast adapt to the target domain and the predicted mask is refined with calibrated features produced by fine-tuned anchor layers.

Formally, given a sequence of L intermediate feature maps of the query image $\{\mathbf{F}_l^q\}_{l=1}^L$ and its predicted probability map $\hat{\mathbf{M}}$, we compute the foreground class prototype of the query image at layer l with the probability map $\hat{\mathbf{M}}_l = \zeta_l(\hat{\mathbf{M}})$ by applying the soft masked average pooling method. Thus, the loss function \mathcal{L}_{kl} for fine-tuning the model can be computed as follows:

$$\mathcal{L}_{kl} = \sum_{l=1}^L D_{KL}(\mathbf{p}_{f,l}^s || \hat{\mathbf{p}}_{f,l}^q), \text{ where } \hat{\mathbf{p}}_{f,l}^q = \frac{\sum_i \mathbf{F}_{l,i}^q \hat{\mathbf{M}}_{l,i} \mathbb{1}[\hat{\mathbf{M}}_{l,i} \geq \tau]}{\sum_i \hat{\mathbf{M}}_{l,i}} \quad (2)$$

Here, $D_{KL}(\cdot)$ denotes the Kullback-Leibler divergence loss function and i denotes 2D spatial positions of the feature map. $\mathbb{1}(\cdot)$ is an indicator function to extract the binary predicted mask from $\hat{\mathbf{M}}_l$, outputting value 1 if the argument is true or 0 otherwise. Pixels will be predicted as the foreground class if their values are larger than threshold τ . We set $\tau = 0.5$ in our experiments.

6 Experiment

6.1 Evaluation Setup

Datasets. We use PASCAL VOC 2012 [13] with SBD [17] augmentation as training domain and then evaluate the trained models on the proposed benchmark introduced in Section 3.

Baseline. To evaluate the performance of existing few-shot semantic segmentation models on CD-FSS, we adopt eight representative few-shot segmentation models: AMP [34], CaNet [49], PANet [45], RPMMs [46], PGNet [48], PFENet [38], RePRI [3] and HSNNet [27]. We use the publicly available codes and follow the default training configuration of these models. For CaNet [49] method, we iteratively optimize the predicted results for 4 times after the initial prediction at inference time, which is same as their recommended settings. For a fair comparison, we also adopt ResNet-50 [18] as a feature extractor in PANet [45] to be our baseline model, denoted as PANet*. An alternative way to tackle CD-FSS is based on transfer learning, where an initial model is trained on the source dataset in a standard supervised learning way and reused on the novel datasets. We adapt the FCN [26] and DeeplabV3 [6] to serve as baselines by fine-tuning their last k layers on the support set, denoted as “Ft-last- k ”. For example, “Ft-last-1_{FCN}” represents the performance of fine-tuning the last-1 (fc-8) fully connected layers of FCN-32s pretrained on PASCAL VOC. In addition, the trained segmentation networks followed by the base classifier are also evaluated on the benchmark. The base classifier is trained to map dense features from the support set to their corresponding labels and uses it to generate the predicted mask in the query set. We experimented with various classifiers including 1-NN and logistic regression. For more details, please refer to the supplementary materials.

Training and testing strategy. We meta-train all methods on all the classes of PASCAL VOC with SBD augmentation and meta-test the trained models on each dataset of the proposed benchmark. For each evaluation, we average the mean-IoU of 5 runs [44] with different random seeds. Each run contains 1200 tasks for all datasets except FSS-1000. FSS-1000 has 2400 tasks in each run, which is the same as the setting in [23,27].

6.2 Evaluation Metric

Mean intersection over-union (mIoU), which is defined as the mean IoUs of all image categories, was employed as the metric for performance evaluation.

For each category, the IoU is calculated by $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP and FN respectively denote the number of true positive, false positive and false negative pixels of the predicted segmentation masks.

6.3 Implementation Details

We adopt VGG-16 [36] and ResNet-50 [18] as feature extractors, which are initialized with weights pre-trained on ILSVRC [30] and kept frozen during training, following previous works [25,45,49,27]. For the VGG backbone, we use feature maps from conv4_x to conv5_x, and after the last max-pooling layer. The channel dimensions of the three anchor layers are set to 512. For the ResNet backbone, we use feature maps from conv3_x, conv4_x and conv5_x. The channel dimensions of the three anchor layers are set to 512, 1024 and 2048, respectively. To reduce the memory consumption and speed up training process, we set spatial sizes of both support and query images to 400×400 . We implement the model in PyTorch [28] and utilize the Adam [22] optimizer with a learning rate of 1e-3. At inference, all images are resized to a fixed 400×400 resolution. An Adam optimizer is used to fine-tune PATM, with a learning rate of 1e-3 for Deepglobe and ISIC, 5e-5 for Chest X-ray and FSS-1000. For each task, a total of 50 iterations are performed. More details can be found in the supplementary material.

6.4 Baseline Performance Analysis

Meta-learning based results. Table 2 shows the results using mIoU, in terms of different datasets, methods, and shot levels in the benchmark. The results reveal that the performance of existing few-shot semantic segmentation methods degrades significantly under domain shifts, especially under large domain gaps. The main reason is that the frozen pretrained encoder cannot generate distinguishable features for the downstream decoder when a large domain gap exists. Furthermore, when the target domain is similar to the source domain, like on FSS-1000, the relation-based methods generally perform better than the metric-based methods. But when the domain gap becomes larger (*e.g.* Deepglobe and Chest X-ray), the metric-based methods are more effective than the relation-based methods. For instance, PANet surpasses HSNet by 5.87% (1-shot) and 14.95% (5-shot) on Chest X-ray, but underperforms HSNet by 8.38% (1-shot) and 9.31% (5-shot) on FSS-1000. This indicates that if the target domain is drastically different from the source domain, it may be more effective to make the encoder obtain the meta-transfer ability than the decoder. Finally, we observed that all the methods achieved the best performance on the FSS-1000 dataset among the four selected datasets because the data distribution of the FSS-1000 is most similar to the source dataset (PASCAL VOC) compared to the other datasets.

Transfer learning based results. We observe that the base classifier methods significantly outperform simple fine-tuning methods on CD-FSS. The main reason is that limited samples in support set are insufficient for the deep segmentation networks to be adapted to a novel distribution. Furthermore, when

Table 2. Mean-IoU of 1-way 1-shot and 5-shot results of meta-learning and transfer learning methods on the CD-FSS benchmark. **Note that all methods are trained on PASCAL VOC and tested on CD-FSS.** Bold denotes the best performance among *all* methods and underlined shows the best performance in *each* method group. * denotes the model implemented by ourselves.

Methods	Backbone	Deepglobe		ISIC		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
<i>Transfer Learning Methods</i>											
Ft-last-1 _{FCN}	Vgg-16	29.80	<u>32.25</u>	15.17	19.75	<u>33.63</u>	48.08	32.51	53.62	27.78	38.43
Ft-last-2 _{FCN}	Vgg-16	32.90	35.34	17.52	21.65	36.35	53.85	32.15	57.44	29.82	42.07
Ft-last-3 _{FCN}	Vgg-16	32.91	35.54	17.91	25.58	45.61	56.05	33.32	<u>60.86</u>	32.34	44.51
1NN _{FCN}	Vgg-16	32.42	38.63	15.68	23.66	46.26	52.70	41.51	46.64	33.97	40.41
Linear _{FCN}	Vgg-16	<u>33.56</u>	38.75	15.51	<u>30.65</u>	37.69	50.07	41.09	49.16	31.96	42.16
Ft-last-1 _{Deeplab}	Res-50	28.11	28.65	11.08	16.57	30.43	35.54	25.14	35.86	23.69	29.41
Ft-last-2 _{Deeplab}	Res-50	24.09	36.74	10.22	17.56	31.16	51.57	20.68	42.50	21.29	37.10
1NN _{Deeplab}	Res-50	32.28	35.96	<u>21.44</u>	26.04	<u>47.76</u>	57.93	<u>45.81</u>	55.95	<u>36.82</u>	43.97
Linear _{Deeplab}	Res-50	32.95	<u>39.69</u>	19.42	30.04	43.52	<u>60.29</u>	40.50	58.36	34.10	<u>47.10</u>
<i>Few-Shot Segmentation Methods</i>											
AMP [34]	Vgg-16	<u>37.61</u>	40.61	28.42	30.41	51.23	53.04	57.18	59.24	43.61	45.83
PGNet [48]	Res-50	10.73	12.36	21.86	21.25	33.95	27.96	62.42	62.74	32.24	31.08
PANet* [45]	Res-50	36.55	45.43	25.29	33.99	57.75	<u>69.31</u>	69.15	71.68	47.19	<u>55.10</u>
CaNet [49]	Res-50	22.32	23.07	25.16	28.22	28.35	28.62	70.67	72.03	36.63	37.99
RPMs [46]	Res-50	12.99	13.47	18.02	20.04	30.11	30.82	65.12	67.06	31.56	32.85
PFENet [38]	Res-50	16.88	18.01	23.50	23.83	27.22	27.57	70.87	70.52	34.62	34.98
RePRI [3]	Res-50	25.03	27.41	23.27	26.23	<u>65.08</u>	65.48	70.96	74.23	46.09	48.34
HSNet [27]	Res-50	29.65	35.08	<u>31.20</u>	<u>35.10</u>	51.88	54.36	<u>77.53</u>	<u>80.99</u>	<u>47.57</u>	51.38
PATNet	Vgg-16	28.74	34.83	33.07	45.83	57.83	60.55	71.60	76.17	47.81	54.35
PATNet	Res-50	37.89	<u>42.97</u>	41.16	53.58	66.61	70.20	78.59	81.23	56.06	61.99

the target domain is similar to the source domain (*e.g.* FSS-1000), those meta-learning based methods outperform transfer learning based methods with a large margin. In contrast, the base classifier methods surprisingly achieve comparable performance when a large domain shift gap exists. For example, the pre-trained Deeplab with a simple linear classifier achieves 39.69% on Deepglobe for 5-shot, outperforming most few-shot segmentation methods. It is worth noting that RePRI [3] is also a kind of transfer learning method designed for few-shot segmentation tasks. It performs well on Chest X-ray and FSS-1000, but fails on Deepglobe and ISIC. This indicates that it is inefficient only to fine-tune the classifier during inference. Generating distinguishable features for the downstream segmentation modules is a key to reducing the performance degradation brought by domain shifts.

6.5 Experimental Results of PATNet

As shown in Table 2, across all the datasets, our model outperforms both meta-learning methods and transfer learning based methods with a sizable margin. Specifically, our 1-shot and 5-shot results respectively achieve **8.49%** and

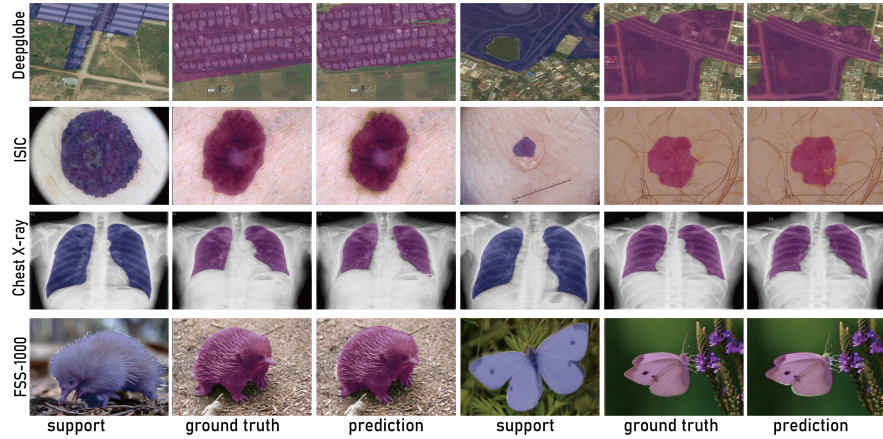


Fig. 5. Qualitative results of our model in 1-way 1-shot segmentation on CD-FSS. Note that the model is trained with PASCAL. Support labels are overlaid in blue. Prediction and ground truth of query images are in plum. Best view in color and zoom in.

10.61% of mean-IoU improvements over HSNet (achieves the best performance among meta-learning methods on CD-FSS), 21.96% and 14.89% of mean-IoU improvements over DeeplabV3 combined with a linear classifier (achieves the best performance among transfer learning based methods on CD-FSS), verifying its superiority on the CD-FSS task. In particular, our model outperforms recent methods with a sizable margin under large domain gaps, surpassing HSNet by 14.73% (1-shot) and 15.84% (5-shot) on Chest X-ray, and 9.96% (1-shot) and 18.48% (5-shot) on ISIC. In addition, we present some of the qualitative results of the proposed model for 1-way 1-shot segmentation in Fig. 5. These results validate that the proposed method can significantly improve the generalization ability under large domain gaps while achieving a comparable accuracy in a similar domain shift.

6.6 Ablation Study

We conduct extensive ablation studies to investigate the impacts of PATM and TFI strategy. All ablation study experiments are performed with ResNet-50.

Effect of pyramid anchor layers. To study the effect of the number of pyramid anchor layers in PATM, we compare our method with and without the anchor layers. We also form an explicit transformation module using a unique anchor layer for each intermediate feature map. From Table 3 we can observe that introducing the anchor layers for feature transformation improves the segmentation performance with 8.25% and 7.64% gain in 1-shot and 5-shot, respectively. This suggests that our proposed PATM is able to enhance the generalization ability by transforming the domain-specific features

Table 3. Ablation study on PATM on CD-FSS. Results are averaged over 4 datasets for 1-shot and 5-shot.

Method	CD-FSS		#params to train
	1-shot	5-shot	
w/o PAT	47.57	51.38	2.574M
explicit PAT	54.16	59.38	2.602M
PATNet	56.06	61.99	2.581M

Table 4. Ablation study on the choice of fine-tuning anchor layers on Deepglobe.

Ft _{high}	Ft _{med}	Ft _{low}	1-shot	5-shot
×	×	×	35.10	40.72
✓	×	×	37.52	42.03
×	✓	×	34.56	39.74
×	×	✓	37.89	42.97

into domain-agnostic ones. One may ask why not make each feature map have its own anchor layer. We compare the results with the explicit transformation module, introducing the anchor layer for each intermediate feature map (denoted as ‘explicit PAT’ in Table 3). Performance degradation from PATNet to explicit PAT indicates that the light-weight anchor layers are more reliable to construct the domain transformation matrices in few-shot scenarios. Thus, we only introduce one anchor layer for each feature dimension and feature maps with the same dimension share one anchor layer to compute their corresponding transformation matrices.

Choice of fine-tuning anchor layers. Table 4 provides a quantitative evaluation of the TFI strategy. We present the results of fine-tuning each anchor layer: low-, medium- and high-level feature dimensions, respectively. We observe that fine-tuning the anchor layer of the low feature map achieves the best performance, indicating that the correlation patterns from low intermediate CNN layers are crucial in effective domain transfer. Qualitative results on how TFI affects the final prediction are provided in Fig. 4. We adopt fine-tuning the anchor layer for low dimensions to report all the experiment results.

7 Conclusion

In this paper, we extend few-shot semantic segmentation to a new task, called Cross-Domain Few-Shot Semantic Segmentation (CD-FSS), which aims to learn a model that can segment the novel classes in *unseen* domains with only *a few* pixel-level annotated images. Moreover, a new benchmark for CD-FSS is established to evaluate the cross-domain generalization ability of few-shot segmentation models under different domain shifts. Experiments show that SOTA few-shot segmentation models do not generalize well to categories from different domains, due to the large discrepancy of the feature distribution across domains. In addition, we propose a novel model, PATNet, to tackle the CD-FSS problem by transforming domain-specific features into domain-agnostic ones for downstream segmentation modules to fast adapt to unseen domains. Extensive experimental results show that our method outperforms the prior art with a sizable margin under domain shifts. We believe this work will help the community understand existing methods in a practical way and dive into further advances for real-world applications.

References

1. Adler, T., Brandstetter, J., Widrich, M., Mayr, A., Kreil, D., Kopp, M., Klambauer, G., Hochreiter, S.: Cross-domain few-shot learning by representation fusion. arXiv preprint arXiv:2010.06498 (2020)
2. Ben-Israel, A., Greville, T.N.: Generalized inverses: theory and applications, vol. 15. Springer Science & Business Media (2003)
3. Boudiaf, M., Kervadec, H., Masud, Z., et al.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13979–13988 (2021)
4. Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging* **33**(2), 577–590 (2013)
5. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 834–848 (2018)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019)
8. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1992–2001 (2017)
9. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2018)
10. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
11. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
12. Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X.: Sfsdan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 982–991 (2019)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International journal of computer vision* **88**(2), 303–338 (2010)
14. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400 (2017)
15. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043 (2017)

16. Guo, Y., Codella, N.C., Karlinsky, L., Smith, J.R., Rosing, T., Feris, R.: A new benchmark for evaluation of cross-domain few-shot learning. arXiv preprint arXiv:1912.07200 (2019)
17. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision. pp. 991–998. IEEE (2011)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6629–6640 (2017)
20. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
21. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging **33**(2), 233–245 (2013)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2869–2878 (2020)
24. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002 (2018)
25. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. arXiv preprint arXiv:2007.06309 (2020)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
27. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6941–6952 (2021)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
29. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
31. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7374–7383 (2019)
32. Seo, J., Park, Y.H., Yoon, S.W., Moon, J.: Task-adaptive feature transformer for few-shot segmentation. arXiv preprint arXiv:2010.11437 (2020)
33. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)

34. Siam, M., Oreshkin, B.N., Jagersand, M.: Amp: Adaptive masked proxies for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5249–5258 (2019)
35. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4136–4145 (2020)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in neural information processing systems. pp. 4077–4087 (2017)
38. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI) (2020)
39. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., et al.: Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096 (2019)
40. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
41. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1456–1465 (2019)
42. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 180161 (2018)
43. Tseng, H.Y., Lee, H.Y., Huang, J.B., Yang, M.H.: Cross-domain few-shot classification via learned feature-wise transformation. arXiv preprint arXiv:2001.08735 (2020)
44. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. pp. 3630–3638 (2016)
45. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9197–9206 (2019)
46. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. arXiv preprint arXiv:2008.03898 (2020)
47. Yoon, S.W., Seo, J., Moon, J.: Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In: International Conference on Machine Learning. pp. 7115–7123. PMLR (2019)
48. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9587–9595 (2019)
49. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5217–5226 (2019)
50. Zhang, X., Wei, Y., Yang, Y., Huang, T.S.: Sg-one: Similarity guidance network for one-shot semantic segmentation. IEEE Transactions on Cybernetics (2020)

51. Zhang, Y., David, P., Foroosh, H., Gong, B.: A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE transactions on pattern analysis and machine intelligence* (2019)
52. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2881–2890 (2017)
53. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 568–583 (2018)
54. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 289–305 (2018)