TGCM:Cross-Domain Few-Shot Semantic Segmentation via one-shot Target Guided CutMix

HaiTao Wei[1 [0009−0008−6924−6486]], JianMing Liu*[1 [0009−0007−5755−2729]], Tong Chen[1 [0009−0008−9332−6902]], and WenLong Qiu[1 [0009−0006−8717−9884]]

School of Computer Information Engineering, Jiangxi Normal University, NanChang, China
{202241600194,liujianming,202241600177,202341600192}@jxnu.edu.cn

Abstract. The goal of few-shot semantic segmentation is to build a model using a small amount of annotated data to generalize to a new object class. When there are significant differences between the target domain and the source domain, segmentation performance usually deteriorates significantly. Existing methods mainly rely on annotated data from the source domain to achieve domain-independent feature extraction through feature transformation. However, bridging the huge domain gap without guidance from the target domain data remains highly challenging. To effectively address the domain shift problem in cross-domain few-shot semantic segmentation (CD-FSS), this paper introduces a new method called Target-guided Cross-domain Few-shot Semantic Segmentation (TGCM), which introduces a labeled sample from the target domain to guide model learning during training. Specifically, TGCM utilizes one-shot image and mask from the target domain as auxiliary data and employs the CutMix method for data augmentation on the source domain training data. Subsequently, the task-adaptive feature transformer module (TAFT) and domain channel alignment (DCA) module are introduced to translate the features of fused images into the feature space related to the target domain, reducing the domain drift effect caused by cross-domain discrepancies. Finally, we present a Dynamic Prediction (DP) strategy to help the model gradually improve segmentation performance. Experimental results show that our model achieves significant improvement in CD-FSS, with average accuracies higher by 4.71% and 2.95% on 1-shot and 5-shot, respectively, compared to the baseline methods in CD-FSS. The code and dataset are available on https://github.com/08-401/TGCM_ACCV.

Keywords: Few-Shot Learning · Cross-Domain Semantic Segemenation · CutMix.

1 Introduction

Deep neural networks are widely used for semantic segmentation of images, such as U-Net [1], FCN [2], and DeepLab [3]. However, these algorithms often require a large amount of annotated data during training, which is very time-consuming

and labor-intensive in some fields. To address this problem, *Guo* and *Yunhui* proposed the task of few-shot semantic segmentation (FSS), many excellent FSS models have emerged over the years. These include parameter-based models like CANET [5], CRNET [6], and PFENET [7], as well as prototype-based models like PANET [8]. However, these methods are limited to scenarios where the target domain and the source domain are in the same domain. When there is a significant difference between the target domain and the source domain, traditional FSS methods [5–8] experience a sharp decline in segmentation performance due to domain shift. To address this issue, the task of cross-domain few-shot semantic segmentation (CD-FSS) has been proposed [9]. The core challenge of CD-FSS is the significant domain gap between the source domain and the target domain, including differences in data distribution and label space. Additionally, there is no access to target domain data during training. Traditional domain adaptation methods usually rely on accessing target domain data during training to adjust the model, whereas CD-FSS requires the model to generalize without any access to target domain data. Therefore, learning universal features from the source domain and achieving accurate and effective segmentation in the target domain is key to this task. Inspired by Meta-FDMixup [17], we propose a novel approach

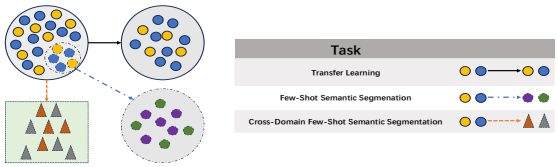

Fig. 1. The difference between cross-domain few-shot semantic segmentation and existing segmentation tasks.

to address cross-domain few-shot semantic segmentation by introducing a single Image and its Mask from the target domain during the training phase as auxiliary data. In practical applications, obtaining a single annotated target domain image is not difficult. The target domain categories during testing do not appear in the training phase, thus not violating the CD-FSS setup.

Based on the above setup, we propose a cross-domain few-shot segmentation method guided by a single annotated target domain image. First, in the training phase, the auxiliary data is mixed with the source domain dataset through CutMix, resulting in intermediate domain data that combines features from both the source and target domains. Second, we propose a domain channel alignment module to perform frequency domain analysis on features at the channel level, enhancing the correlation between intermediate domain data and the target domain. Then, the task-adaptive feature transformer (TAFT) module converts domain-specific features into domain-agnostic features, which are introduced into the ultra-correlation construction through center-pivot in 4D convolution and Cosine.

Our contributions are summarized as follows. **1)** We propose for the first time the introduction of labeled target data to guide the learning process of cross-domain few-shot semantic segmentation tasks.**2)** We propose a cross-domain few-shot semantic segmentation method TGCM, guided by a single target domain sample. TGCM utilizes the CutMix augmentation method to mix one target domain sample (only one image and its mask) with source domain data and employs the domain channel alignment module and task-adaptive feature transformer module to mitigate the domain shift in cross-domain tasks further, thereby improving the model's segmentation performance in the target domain.**3)** Extensive experimental results validate the rationality of our new setup and the effectiveness of our proposed method. Compared to the baseline [9] method in CD-FSS, our method achieves an average accuracy improvement of 4.71% and 2.95% on 1-shot and 5-shot tasks, respectively.

2 Related Work

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

Few-shot semantic segmentation. Currently, mainstream few-shot semantic segmentation algorithms are mainly categorized into metric-based and relation-based methods. Metric-based methods (such as PANet [8] and AMP [19]) utilize the nearest neighbor concept to achieve semantic segmentation of image pixels, enabling them to maintain good segmentation performance on datasets with certain domain gaps. These methods rely on constructing a stable metric space. On the other hand, relation-based methods (including HS-Net [18], CA-Net [5], RPMMs [20], PFENet [21], and PGNet [22]) freeze the encoder of the pre-trained model and only train the decoder. Metric-based methods focus more on separating foreground and background in images, while relation-based methods depend more on feature recognition in the pre-trained model. Therefore, the strength of the feature extractor determines the performance of relation-based methods. However, under the setting of the CD-FSS task, due to significant domain shifts in images, segmentation performance often experiences a noticeable decline.

Cross-domain few-shot semantic segmentation. In response to the cross-domain few-shot semantic segmentation task, *Lei et al.* proposed the PATNET [9] model, which uses a Transformation Module to convert domain-specific features into domain-agnostic features, addressing the issue of models trained in the source domain being unable to perform well in the target domain. Building on this, RestNet [35] introduced a new residual transformation network that retains intra-domain support-query feature information while promoting knowledge transfer, but its semantic segmentation performance on certain datasets is less than ideal. PMNet [36] proposed a more lightweight architecture based on a dense correlation matrix, achieves segmentation using a dense correlation matrix between query and support pixels. ABCDFSS [37] proposed a new perspective

on CD-FSS [9] by adding a task-adaptive network after each bottleneck block. It replaces the previous approach of learning a downstream segmentation network and test-time fine-tuning, instead of fine-tuning the adaptor during test-time. This method achieves cross-domain few-shot image segmentation through cross-view consistency and hierarchical fusion, but its performance is not ideal on datasets with many classes and few data per class.

Data augmentation for cross-domain few-shot learning. Currently, there is relatively limited research on using data augmentation methods in cross-domain few-shot learning tasks. Meta-FDMixup [17] partially exposes the target domain data and applies Mixup [32] operations to the data and labels of both the source and target domains using different proportions. Additionally, Meta-FDMixup [17] sets a large number of images and labels from the target domain as visible during training. This algorithm partially alleviates poor classification performance between the source and target domains in cross-domain few-shot learning tasks caused by large domain gaps. *Yuzuru Nakamura* proposed a few-shot adaptive object detection method with cross-domain CutMix [23]. However, the author employs feature-based adversarial learning to adjust the domain, and for effective fine-tuning, the training data features between the source and target domains **must be similar**. For example, a model trained using RGB images serves as the pre-trained model and is fine-tuned with infrared images. Both of these knowledge transfer methods, whether considering the task requirement of invisible target domain data or the size of domain gaps between the source and target domains, are not suitable for introducing cross-domain few-shot semantic segmentation tasks.

3 Method

3.1 Problem setting

We keep the same task setting as PATNet [9]. in the cross-domain few-shot semantic segmentation problem, the training sample D_{train} is composed of (X_s, Y_s) ,and the test sample D_{test} is composed of (X_t, Y_t) from the target domain, where the data distribution between the source domain and the target domain is different, and the target domain data cannot be accessed during the training period, that is $X_t, i.e, X_s \neq X_t, X_s \cap X_t = \emptyset$.

During the training process, $S = (X_i^s, Y_i^s)$ and $Q = (X_i^q, Y_i^q)$ are extracted from the training sample D_{train}, where i represents a class in the sample. During testing or meta-testing, D_{train} provides a support-set for the model, which is then evaluated using a set of queries from the target domain to assess segmentation performance.

3.2 Our approach

The goal of the cross-domain few-shot segmentation task is to train a few-shot semantic segmentation model in the source domain and generalize it to a target

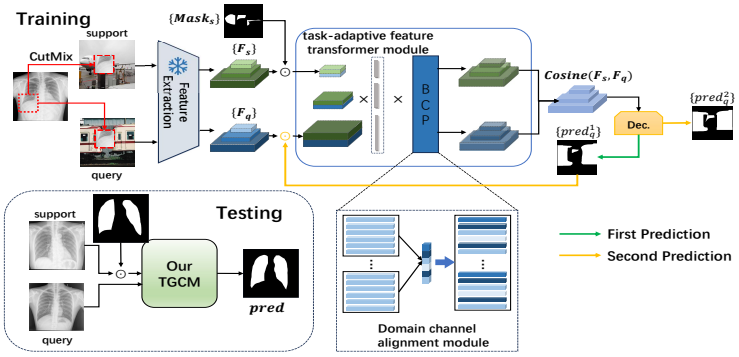

Fig. 2. Overview of our TGCM network. It mainly consists of three components: feature extraction backbone, domain channel alignment module, and task-adaptive feature transformer module. Firstly, we utilize the feature extraction backbone to extract image features from the support-set and query-set. Then, through the DFAT module, features constrained to specific domains are transformed and perturbed across different scales and channels to generate domain-agnostic, diversified features, aiming to reduce the negative transfer effects caused by domain shifts. Finally, we train convolutional decoders of various sizes to facilitate the generation of corresponding prediction masks. Additionally, we adopt a dynamic prediction (DP) strategy to enhance model learning.

domain that is significantly different from the source domain. The key issue to address is the performance degradation due to domain shift. Previous research relies on a sufficiently strong pre-trained encoder to embed features into a distinguishable form for the downstream matching module. However, this pre-trained backbone often significantly reduces segmentation performance due to unseen new classes during the training phase, especially in cases with a large domain gap, such as the difference in channel numbers between natural images and medical images. To reduce the drift phenomenon caused by this domain gap, as is shown in Figure 2, our model applies CutMix [16] to one image from the target domain, integrating it into the training data. It further reduces the domain gap through a domain channel alignment module and a task-adaptive feature transformer module.

3.3 CutMix's mixing strategies

CutMix [16] is an advanced data augmentation method proposed by *Sangdoo Yun* et al., which involves cutting a rectangular region from one image and pasting it onto another image to create a new combined image. This method is commonly used in object detection or image classification tasks, and the generated labels are calculated by the weighted average of areas from different cut regions. In order to diversify the features of the images, we propose target style data augmentation based on CutMix, aiming to make the features closer to the style of the target domain before transformation. Therefore, in this study, we will

use two different CutMix mixing strategies: pixel-based CutMix mixing strategy and feature space-based CutMix mixing strategy.

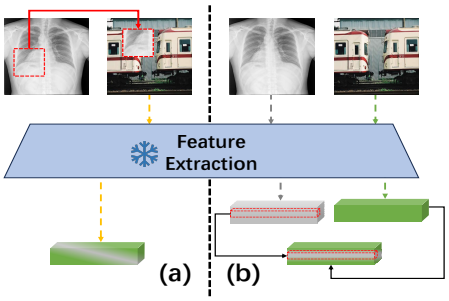

Fig. 3. CutMix's two mixing strategies.

The left half of Figure 3 illustrates the pixel-based CutMix mixing strategy. A sample image and its corresponding mask are extracted from the target domain to be used as auxiliary data. Then, a cropping ratio R is set to generate a cropping region, and a random position is chosen on the auxiliary data for cropping. Finally, the cropped region is pasted onto a fixed position on the source domain image to generate new mixed data.The right half of Figure 3 illustrates the feature-space based CutMix mixing strategy. The key difference from the previous strategy is that the latter involves passing the image through a feature extractor to obtain feature matrices. Next, regions within the feature matrices are cropped proportionally, and these cropped feature regions are pasted onto the feature matrix of the source domain image. It's important to note that the mask follows the same cropping strategy as the image.

3.4 TAFT and DCA module

To reduce the decline in segmentation performance caused by cross-domain issues, we improve the TAFT [33] and domain channel alignment (DCA) module to transform monotonous source domain image features into target domain-related features.

Domain Channel Alignment (DCA) Module. In "Diversified Arbitrary Style Transfer via Deep Feature Perturbation" [24], a deep feature perturbation method is proposed, which uses an orthogonal noise matrix to perturb image feature maps extracted from deep convolutional neural networks. We consider feature perturbation from another perspective and propose the Domain Channel Alignment (DCA) module. The DCA trains a stable perturbator using advanced semantic features derived from channel attention mechanisms, which subtly perturb features across different batches.

In cross-domain few-shot semantic segmentation tasks, the domain shift between the source and target domains can cause significant differences in data.

During training, because the target domain data is invisible, the encoder and decoder will adapt more to the domain and classes of the training data. If the encoder cannot provide high-level features of the target domain, good cross-domain segmentation performance cannot be achieved. Therefore, we use a unified batch channel attention mechanism to perform channel perturbations on certain intermediate layers l, introducing some variations and constraints to the features, thereby increasing the model's robustness and generalization performance to the data.

$$\hat{f}_l = bat\left\{\left(1-\sigma\left(Conv\left(AvgPool\left(\hat{f}_{li}\right)\right)\right)\right)\right\}_i^{k=512} * \hat{f}_{li} \tag{1}$$

Equation 1, $\hat{f}_l$ represents the masked feature. Since the masked feature varies in multiple sizes and channels, *bat* represents batches of equal length. in our experiment, 512 channels are used as one batch. *AvgPool* denotes the average pooling operation, facilitating compatibility with masked features of different sizes. The unified batch channel perturbation module allows sharing a perturbation mechanism across feature layers of different sizes. This helps the model learn a consistent perturbation ability for features of varying sizes and channels from the support-set and query-set. This also reduces the scale of training parameters.

Task-adaptive Feature Transformer (TAFT). Inspired by previous work on the Task-adaptive Feature Transformer [33], we use a linear transformation matrix as a transformation mapper to linearly convert task-specific high-level features into task-independent features. First, we need to calculate the foreground and background prototype representations from the support-set's mask $mask^s$. For example, in a 1-shot scenario, the foreground prototype of the l-th layer is calculated as follows:

$$C_{s,l}^f = \frac{\sum_i F_{i,l}^s B_l(mask_i^s)}{\sum_i B_l(mask_i^s)} \tag{2}$$

where B_l is bilinear interpolation, and $F_{i,l}^s$ and $mask_l^s$ denote the 2D spatial positions. Suppose A represents the weight matrix of the anchor layer, C_s represents the prototype matrix of the support-set image, and $C_s = \left[\frac{C_s^f}{\|C_s^f\|}, \frac{C_s^b}{\|C_s^b\|}\right]$. We construct the transformation matrix I by finding a matrix that satisfies $I * C_s = A$. By calculating feature matrices of different dimensions and scales from various layers of ResNet-50 and using three linear layers to expand the transformation matrix, we convert the feature matrix pixel by pixel into task-independent features. Here, $A = \left[\frac{a^f}{\|a^f\|}, \frac{a^b}{\|a^b\|}\right]$ and a is an anchor vector with a length consistent with the feature channel. Therefore, we can calculate the non-square matrix $C_{s,l}$ using: $C_{s,l}^+ = \left\{C_{s,l}^T C_{s,l}^+\right\}^{-1} C_{s,l}^T$. Thus, we obtain $I_l = A_l C_{s,l}^+$.

$$\dot{F}_{s,l} = C_l * F_{s,l} \tag{3}$$

Thanks to the center-pivot 4D convolution work proposed in Hyper-correlation Squeeze for Few-Shot Segmentation [18], we incorporate it into subsequent hypercorrelation construction. For each layer l, the pair of transformed mask support

features $\dot{F}_{s,l}$ 3 and query features $F_{q,l}$ are obtained, where the relationship between $\dot{F}_{s,l}$ and $F_{s,l}$ is shown in Equation 4. Using cosine similarity, we construct a 4D correlation tensor $T_l \in \mathbb{R}^{H_l W_l H_l W_l}$

$$T_l\left(i,j\right) = Relu\left(\frac{I_l F_{q,l}\left(i\right) \cdot I_l \dot{F}_{s,l}\left(j\right)}{\|I_l F_{q,l}\left(i\right)\| \|I_l \dot{F}_{s,l}\left(j\right)\|}\right) \tag{4}$$

where i and j denote 2D spatial positions of $F_{q,l}$ and $\dot{F}_{s,l}$, respectively.

3.5 Dynamic prediction strategy

To further optimize and accelerate the model's ability to recognize features in segmented images, we improved upon methods from ResNet [35] and introduced a strategy called dynamic prediction. This strategy integrates foreground features into the target prototypes of the support-set, refining the hierarchical representation of both foreground and background prototypes. The motivation behind this is that our proposed model is also based on a relation-based method with an encoder that freezes pre-trained weights. Hence, under cross-domain conditions, the strength of background and foreground prototype representations often determines the semantic segmentation capabilities of relation-based methods. The model identifies the first prediction results exceeding a certain threshold as masks, separates the foreground and background of query-set images, and generates corresponding prototype representations.

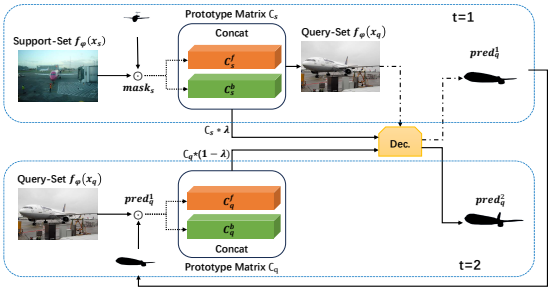

Fig. 4. Dynamic prediction strategy structure chart.

As shown in Figure 4, using $mask_s$ and image features $f_\Phi\left(x_s\right)$ to calculate the foreground and background prototype representations of the support-set, C_l in Equation 5 is obtained by concatenating $\left[C_{s,l}^f, C_{s,l}^b\right]$. After obtaining the first prediction $pred_q^1 \leq T$, a parameter λ is introduced to adjust the weight ratio of the prototype representations between the support-set and query-set. Using two different index matrices of $pred_q^1$ as foreground and background to make a second segmentation prediction, $C_{q,l}^f, C_{q,l}^b$ are obtained for the second

segmentation prediction. If the first prediction mask $pred_q^1 < T$, the second prediction action is abandoned.

$$C_l = \begin{cases} \left[C_{s,l}^f, C_{s,l}^b\right] & T = 1 \\ \lambda * \left[C_{s,l}^f, C_{s,l}^b\right] + (1-\lambda) * \left[C_{q,l}^f, C_{q,l}^b\right] & T = 2 \end{cases} \tag{5}$$

3.6 Loss function

Maximum Mean Discrepancy (MMD) [39, 40] measures the distance between the mean embeddings of two different distributions (source domain and target domain) in the reproducing kernel Hilbert space. As shown in Equation 7, x^m is the source domain dataset after CutMix, x^q is the target domain dataset, and $\phi(\cdot)$ is the feature mapping function. By constructing the DAC module to align the features, MMD is introduced as a loss function in the model training to enforce channel perturbations for this module.

$$L_{MMD}(x^m, x^q) = \| \frac{1}{n} \sum_{i=1}^{n} \phi(x^m) - \frac{1}{m} \sum_{j=1}^{n} \phi(x^q) \|_H^2 \tag{6}$$

Recall that we performed CutMix operations on the images from the source domain and target domain during the training phase. Similarly, we need a multi-task learning mechanism consisting of a segmentation task for the mixed images and embedded target domain images to optimize our model. The loss function is defined as follows:

$$L = L_t + R^2 * L_p + p * L_{MMD} \tag{7}$$

where L_p is the cross-entropy loss between the predicted mask of the mixed image's target domain part and the local ground truth, and R is the mixing ratio of CutMix, usually appearing in squared form. L_t is the cross-entropy loss between the predicted mask of the entire mixed image and its ground truth.

4 Experiment

4.1 DataSet

In this paper, we use the PASCAL VOC [10] dataset as the source domain and introduce four datasets with certain domain shifts: Chest-X [11,12] for X-ray images of lungs, ISIC2018 [13,14] for skin lesion images, DeepGlobe [38] for satellite images, and FSS-1000 [15] for everyday object images. Chest-X contains 566 X-ray images, including 58 cases of tuberculosis and 80 normal cases. ISIC2018 consists of lesion images for skin cancer screening, and we use only 2596 images with ground truth to report experimental results. FSS-1000 is a natural image dataset specifically designed for few-shot semantic segmentation, consisting of 1000 object classes, each with 10 samples. We report experimental results on

240 classes and 2400 images. DeepGlobe comprises 803 satellite images, each of size 2448*2448 pixels, with dense pixel-level annotations for 7 categories: forest, agriculture, rangeland, water body, urban area, unknown, and barren land. To reduce the size and quantity of individual images, we divide each image into 6 blocks, filtered out single-class images and images with unknown regions, and used the remaining 5666 images to report experimental results. All images will be uniformly resized to 400*400 pixels during the training and testing processes.

4.2 Experimental procedure

Baseline. To evaluate our model, we conducted semantic segmentation tests on cross-domain datasets using existing few-shot semantic segmentation models: AMP [19], PANet [8], PFENet [7], RePRI [25], HSNet [18], as well as cross-domain few-shot semantic segmentation models: PATNet [9], RestNet [35], HDMNet [34], ABCDFSS [37], and PMNet [36]. We used publicly available code and followed the default training methods and configurations of these models.
Performance Evaluation. Performance Evaluation. For performance evaluation on different datasets, we conducted meta-training on the PASCAL VOC [10] dataset, extracted a portion of the target domain data during validation to select the best-performing parameters, and retained them locally. Regarding segmentation performance, we will average the **m**ean **I**ntersection **o**ver **U**nion (**mIoU**)of 5 runs [26] with different random seeds. For each class, the IoU calculation formula is $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP, and FN represent the numbers of true positive, false positive, and false negative pixels in the predicted segmentation mask, respectively.

4.3 Implementation details

We employed ResNet-50 [27] as the feature extractor for our model, using weights pre-trained on ILSVRC [28] as the initial values, which were frozen during training. For the ResNet-50 backbone, we utilized hierarchical features from different conv layers to construct a feature pyramid, with channel and size dimensions, respectively. To reduce memory consumption and ensure uniform training, we set the default input image size to 400*400. This model was implemented using the PyTorch [29] framework and optimized using the Adam [30] optimizer with a learning rate of $5e-4$.

4.4 Experimental results

We conducted cross-domain semantic segmentation tests on FSS-1000 [15], ISIC2018 [13, 14], DeepGlobe [38], and Chest-X [11, 12] datasets with 1-shot and 5-shot settings, as indicated by the average results across these datasets. Figure 5 illustrates some segmentation comparison results of the model across different datasets. It is noteworthy that all methods were trained on PASCAL VOC and tested on the CD-FSS dataset.

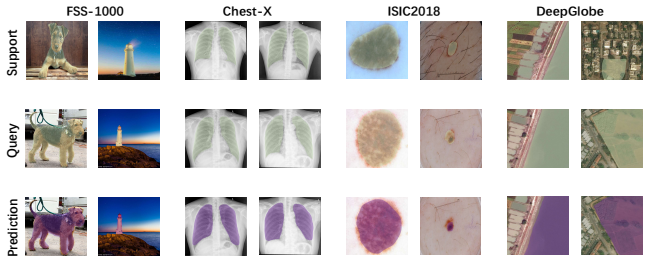

Fig. 5. Our model's segmentation performance is demonstrated on the CD-FSS dataset under a 1-way 1-shot setting. It is important to note that we train on the PASCAL VOC dataset, test on the cross-domain dataset, and during testing, the target domain data is unseen. The masks of the support-set and the Ground Truth of the query-set are overlaid in yellow, while the predictions of the query-set are overlaid in purple. Viewing the images in color mode is recommended.

In the case of 1-shot learning, few-shot learning can learn discriminative relationships between features from extremely limited samples. Additionally, when the source domain and target domain are relatively similar, such as in the FSS-1000 dataset, relation-based methods often outperform metric-based methods in terms of segmentation performance. However, in the ISIC and Chest-X datasets, metric-based methods tend to perform better. This indicates that when there is a large domain gap between the source and target domains, metric-based methods are more effective than relation-based methods because the encoder may acquire better meta-transfer learning capability than the decoder.

Table 1 compares our work on the CD-FSS benchmark with PATNet [9]. Contrasting with PATNet [9] and RestNet [35], which also belong to the relation-based method, TGCM exhibits significant advantages over the baseline PATNet on the Chest-X dataset, with remarkable improvements of 8.9% and 8.66% in 1-shot and 5-shot settings, respectively. On the DeepGlobe dataset, we outperform RestNet by 17.69% and 15.16% in 1-shot and 5-shot settings, respectively. Reporting the results of PATNet with Task-adaptive Fine-tuning Inference (TFI) strategy, we achieve a performance lead of 2.5% and 2.12% in 1-shot and 5-shot tests without any model fine-tuning at the testing stage, surpassing PATNet+TFI. On the ISIC2018 dataset, we achieve improvements of 3.15% and 1.81% in 1-shot and 5-shot settings, respectively, compared to ABCDFSS [37] and PATNet [9]. With a segmentation performance of 60.77% in average 1-shot mIoU, we achieve the best performance. In the 5-shot setting, our performance is only slightly lower by 0.06% compared to the best-performing ABCDFSS [37]. However, on datasets with smaller domain gaps between the source and target domains, such as the FSS-1000 dataset, our performance ranks only fourth.This indicats that some modules in the TGCM may have negative effects on domain adaptation and feature perturbation in datasets with smaller domain gaps.

Table 1. Segmentation performance of our proposed method and other methods under 1-shot and 5-shot Settings (%).Please note that the TGCM network employs a feature-space-based CutMix mixing strategy on the DeepGlobe dataset, while pixel-based CutMix mixing strategy is used for the remaining datasets. Additionally, on this dataset, PATNet reports the test results with the use of the Task-adaptive Fine-tuning Inference (TFI) strategy for fine-tuning anchor layers at the testing stage.

Methods	**DeepGlobe**		**ISIC2018**		**Chest-X**		**FSS-1000**		**Average**	
	1shot	**5shot**	**1shot**	**5shot**	**1shot**	**5shot**	**1shot**	**5shot**	**1shot**	**5shot**
Few-shot Segmentation Semantic Methods										
AMP [19]	37.61	40.61	28.41	30.41	51.23	53.04	57.18	59.24	43.61	45.83
PFENet [21]	16.88	18.01	23.50	23.83	27.22	27.57	70.87	70.52	34.62	34.98
HSNet [18]	29.65	35.08	31.20	35.10	51.88	54.36	77.53	80.99	47.57	51.38
RePRI [25]	25.03	27.51	23.27	26.23	65.08	65.48	70.96	74.23	46.09	48.34
Cross-Domain Few-shot Semantic Segmentation Methods										
PMNet [18]	31.10	41.60	-	-	70.40	74.00	**84.60**	**86.30**	-	-
HDMNet [34]	25.40	39.10	33.00	35.00	30.60	31.30	75.10	78.60	41.00	46.00
PATNet [9]	37.89	42.97	41.16	53.58	66.61	70.20	78.59	81.23	56.06	61.99
RestNet [35]	22.70	29.90	42.30	51.10	70.40	73.70	81.50	84.90	54.20	59.90
ABCDFSS [37]	**42.60**	**49.00**	45.70	53.30	**79.80**	**81.40**	74.60	76.20	60.70	**65.00**
Ours	40.39*	45.06*	**48.85**	**55.39**	75.51	78.86	78.31	80.31	**60.77**	64.94

Table 2. Ablation study on CutMix ratio on **ISIC2018**.

R-**ratio**	**1-shot**	**5-shot**
0.3	46.61	51.87
0.4	**48.85**	**55.39**
0.5	45.54	54.33

Table 3. Ablation study on TAFT and DCA on **Chest-X**.

TAFT	**DCA**	**Params**	**1-shot**		**5-shot**	
			Acc	**Latency**	**Acc**	**Latency**
×	×	-	64.12	330	69.18	438
✓	×	**2.58M**	64.93	379	77.26	502
×	✓	**0.42M**	64.10	346	71.03	467
✓	✓	**3.01M**	**75.51**	395	**78.86**	532

4.5 Ablation study

We conducted extensive ablation experiments to investigate the effects of CutMix ratio, CutMix mixing strategy, TAFT and DCA module. All ablation studies were performed using RestNet-50 as the backbone.

Effect of CutMix ratio. To investigate the influence of CutMix ratios on segmentation performance, we set a parameter R to compare values of 0.3, 0.4, and 0.5. The parameter R represents the aspect ratio of the CutMix target domain image. For example, when $R = 0.5$, a randomly cropped image of size 200*200 will be pasted onto the fixed position of the training image in the source domain. From Table 2, it can be observed that when the cropping ratio is 0.4, the segmentation performance for both 1-shot and 5-shot scenarios is optimal, indicating that the CutMix ratio can affect the segmentation performance on cross-domain datasets to a certain extent. We believe that when the CutMix ratio is smaller, the generated mixed images tend to preserve more local features

Table 4. Table 4:Ablation study on CutMix mixing strategies on **DeepGlobe**.

CutMix mixing strategy	1-shot	5-shot
Pixel-based	23.08	29.11
Feature space-based	**40.39**	**45.05**

and label information of the original images but may reduce the model's learning of global features. Conversely, when the CutMix ratio is larger, the generated mixed images contain more external information, which can help the model learn more diverse visual features.However, it may also lead to more discontinuities in the mixed images, affecting the model's learning effectiveness.

Effect of TAFT and DCA module. Table 3 provides detailed ablation experiment results on the channel alignment (DCA) module and the task-adaptive feature transformer (TAFT) module.The experiment is split into separate tests for the TAFT and DCA modules. We observed that adding the TAFT module improved the segmentation performance by 10.81% and 8.08% in 1-shot and 5-shot settings, respectively. However, the model with only the DCA module showed smaller improvements. Compared to the final segmentation performance of the model, the DCA module provided a more stable channel perturbation effect, benefiting the TAFT module. Overall, the use of both the TAFT and DCA modules significantly enhances the model's cross-domain segmentation capability. However TAFT requires Singular Value Decomposition on dense matrices, resulting in the Latency of the DCA module being less than that of TAFT.

Effect of CutMix mixing strategy. Table 4 provides the results of ablation experiments on the CutMix mixing strategy. We can observe that the pixel-based CutMix mixing strategy performs less effectively in semantic segmentation on the DeepGlobe dataset, which exhibits significant domain shifts between the source and target domains, compared to the feature space-based CutMix mixing strategy. The main difference between these two mixing strategies lies in whether they mix images or feature spaces. The method proposed in this study is built upon a robust feature extractor. In cases where the feature extractor is frozen and there is a significant domain shift in the target domain images, PATNet [9] proposed the TFI strategy, which involves fine-tuning different anchor layers during the testing phase. In contrast, our study enhances the model's adaptability to unseen domain images during the training phase using the feature space-based CutMix mixing strategy.

As shown in Figure 6, we use class activation mapping [41] (CAM) to visualize the images in the DeppGlobe dataset on different mixing strategies. The CAM performs a linear weighted sum of the features in different Spaces. By performing upsampling on the CAM to the size of the input image, the most relevant regions to the support-set can be seen. Since the difference between remote sensing images and natural images is significant, both PATNet [9] and ABCDFSS [37] are in the testing phase, fine-tuning the anchor layer at low latitude and adapters, respectively. We find that TGCM cannot effectively fit the data by embedding

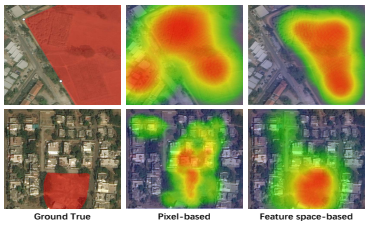

Fig. 6. CAM on different CutMix mixing strategies

the target domain image, and can significantly reduce the MMD [39] distance from the target domain image by embedding the target domain features directly. This method can greatly alleviate the obvious differences between remote sensing images and natural images, and can achieve better segmentation results **without any fine-tuning**.

5 Conclusion

In this paper, we propose a CutMix-based cross-domain few-shot semantic segmentation algorithm. The network uses an Image and Mask from the target domain as assistant data and fuses different CutMix strategies with the training data of the source domain to guide the model to perform cross-domain few-shot semantic segmentation. Extensive experiments demonstrate that the TGCM model outperforms existing techniques on datasets with significant domain gaps between the source and target domains. However, during the training process, we used a very small amount of target domain data as auxiliary data and mixed it with the training data, which breaks the CD-FSS task setting. This work provides a new perspective for the community and makes a small contribution to further in-depth research on real-world applications. We also believe that there will be better methods to address these issues.

Acknowledgments. This work was financially supported by the Natural Science Foundation of China (No.62266022), the Natural Science Foundation of Jiangxi Province (No.20242BAB25110). Corresponding author: JianMing Liu(liujianming@jxnu.edu.cn).

References

1. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.

2. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
3. Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.
4. Guo, Yunhui, et al. "A broader study of cross-domain few-shot learning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer International Publishing, 2020.
5. Zhang, Chi, et al. "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
6. Liu, Weide, et al. "Crnet: Cross-reference networks for few-shot segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
7. Jhou, Fong-Ci, et al. "Mask Generation with Meta-Learning Classifier Weight Transformer Network for Few-Shot Image Segmentation." 2023 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). IEEE, 2023.
8. Wang, Kaixin, et al. "Panet: Few-shot image semantic segmentation with prototype alignment." proceedings of the IEEE/CVF international conference on computer vision. 2019.
9. Lei, Shuo, et al. "Cross-domain few-shot semantic segmentation." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
10. Hoiem, Derek, Santosh K. Divvala, and James H. Hays. "Pascal VOC 2008 challenge." World Literature Today 24.1 (2009): 1-4.
11. Candemir, Sema, et al. "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration." IEEE transactions on medical imaging 33.2 (2013): 577-590.
12. Jaeger, Stefan, et al. "Automatic tuberculosis screening using chest radiographs." IEEE transactions on medical imaging 33.2 (2013): 233-245.
13. Codella, Noel, 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).
14. Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. 2018; 5: 180161." Search in 2 (2018).
15. Li, Xiang, et al. "Fss-1000: A 1000-class dataset for few-shot segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
16. Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
17. Fu, Yuqian, Yanwei Fu, and Yu-Gang Jiang. "Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data." Proceedings of the 29th ACM international conference on multimedia. 2021.
18. Zhang, Wenchao, et al. "HSNet: A hybrid semantic network for polyp segmentation." Computers in biology and medicine 150 (2022): 106173.
19. Siam, Mennatullah, Boris N. Oreshkin, and Martin Jagersand. "Amp: Adaptive masked proxies for few-shot segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

20. Yang, Boyu, et al. "Prototype mixture models for few-shot semantic segmentation." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer International Publishing, 2020.
21. Tian, Zhuotao, et al. "Prior guided feature enrichment network for few-shot segmentation." IEEE transactions on pattern analysis and machine intelligence 44.2 (2020): 1050-1065.
22. Zhang, Chi, et al. "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
23. Nakamura, Yuzuru, et al. "Few-shot adaptive object detection with cross-domain cutmix." Proceedings of the Asian Conference on Computer Vision. 2022.
24. Wang, Zhizhong, et al. "Diversified arbitrary style transfer via deep feature perturbation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
25. Boudiaf, Malik, et al. "Few-shot segmentation without meta-learning: A good transductive inference is all you need?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
26. Vinyals, Oriol, et al. "Matching networks for one shot learning." Advances in neural information processing systems 29 (2016).
27. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
28. Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115 (2015): 211-252.
29. Paszke, Adam, et al. "Automatic differentiation in pytorch." (2017).
30. Liu, Mingrui, et al. "Adam $^+$: A Stochastic Method with Adaptive Variance Reduction." arXiv preprint arXiv:2011.11985 (2020).
31. Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.
32. Huang, Lang, Chao Zhang, and Hongyang Zhang. "Self-adaptive training: beyond empirical risk minimization." Advances in neural information processing systems 33 (2020): 19365-19376.
33. Seo, Jun, et al. "Task-adaptive feature transformer for few-shot segmentation." arXiv preprint arXiv:2010.11437 (2020).
34. Peng, Bohao, et al. "Hierarchical dense correlation distillation for few-shot segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
35. Huang, Xinyang, Chuang Zhu, and Wenkai Chen. "Restnet: Boosting cross-domain few-shot segmentation with residual transformation network." arXiv preprint arXiv:2308.13469 (2023).
36. Chen, Hao, et al. "Dense affinity matching for few-shot segmentation." Neurocomputing 577 (2024): 127348.
37. Herzog, Jonas. "Adapt Before Comparison: A New Perspective on Cross-Domain Few-Shot Segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
38. Demir, Ilke, et al. "Deepglobe 2018: A challenge to parse the earth through satellite images." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018.
39. Gretton, Arthur, et al. "A kernel two-sample test." The Journal of Machine Learning Research 13.1 (2012): 723-773.
40. Gretton, Arthur, et al. "A kernel method for the two-sample-problem." Advances in neural information processing systems 19 (2006).

41. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.