Cross-Domain Few-Shot Segmentation via Multi-view Progressive Adaptation

Supplementary Material

A1. Cross-Domain Few-Shot Segmentation

A1.1. Motivation

Few-Shot Segmentation (FSS) methods rely on abundant base category data to establish the capability of segmenting novel categories with only a few exemplars [10, 13, 38, 43, 50, 52, 72]. However, collecting sufficient annotated samples is often infeasible in data-scarce domains (*e.g.*, satellite imagery and medical screenings), making the FSS pipeline unsuitable. Cross-Domain Few-Shot Segmentation (CD-FSS) [31] addresses this challenge by meta-training models on a well-annotated source domain (*e.g.*, Pascal VOC [11]) and adapting them to data-scarce domains using a small support set, as illustrated in Fig. A1. CD-FSS provides an efficient solution for critical applications (*e.g.*, Tuberculosis detection and wildlife conservation), where large-scale data collection is labor-intensive, costly, and may involve privacy concerns [31]. The CD-FSS pipeline significantly reduces the burden of data collection and annotation in data-scarce target domains.

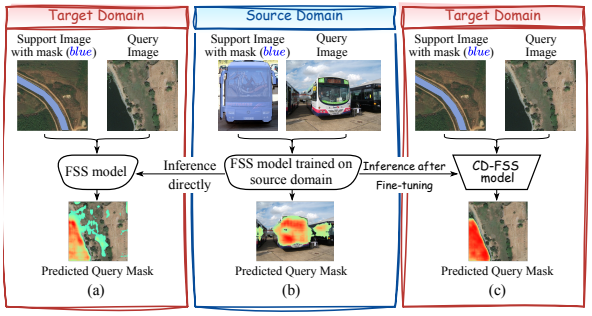

Figure A1. The formulation of the Cross-Domain Few-Shot Segmentation Task. (a) The segmentation performance significantly degrades when directly applying the source-trained model to target domains. (b) Meta-training the model on the source domain. (c) A well-designed adaptor effectively transfers the Few-Shot Segmentation capability to target domains.

A1.2. Two-Stage Pipeline Formulation

A feasible solution for CD-FSS is to perform meta-training on a data-rich source domain (*e.g.*, Pascal VOC [11]) and then adapt the source-trained model to data-scarce target domains. However, as shown in Fig. A1(a), the model often experiences a significant performance drop when directly applied to a different domain. This issue cannot be resolved simply by enhancing few-shot capabilities, as the primary challenge arises from the substantial domain gap. To address this, previous fine-tuning methods, such as IFA [40], enable the model to extract more support-query correspondences even with extremely limited data. This approach enhances the model's generalization ability for segmenting novel categories in target domains, as illustrated in Fig. A1(c). This formulates the two-stage CD-FSS pipeline.

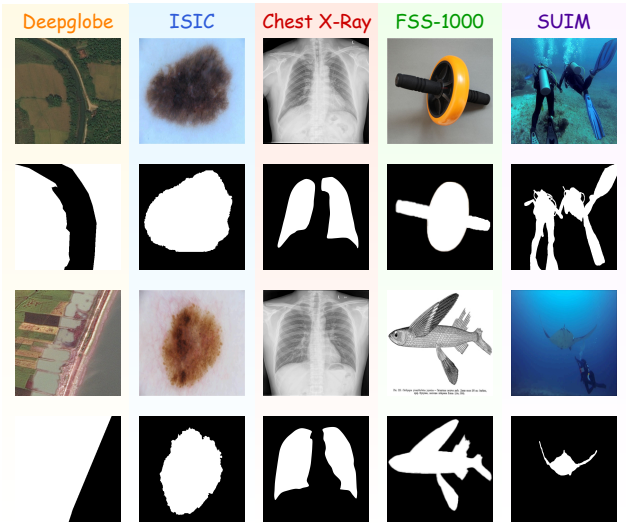

Figure A2. Examples of images and their corresponding ground-truth masks from five low-resource domains, encompassing a diverse range from satellite images, medical screenings, and minuscule everyday objects to underwater scenes.

A1.3. Evaluation Datasets

Fig. A2 shows two sample images and their masks from each of the five datasets.

A2. Method

A2.1. Revisiting the Pipeline of SSP

Motivated by the Gestalt principle [28], which states that pixels belonging to the same object are more similar than those from different objects, the given support may not always serve as a reliable reference for predicting the query mask. To address this, Fan *et al.* propose the Self-Support Prototype (SSP) module [13]. SSP first generates a support prototype P_s from the support feature F_s and mask M_s:

$$P_s = MAP(F_s, M_s), \tag{A1}$$

where MAP denotes the masked average pooling operation. Unlike traditional prototypical learning [10], which directly matches the support prototype with the query image, SSP adopts a two-step matching strategy. Specifically, SSP first utilizes the support prototype to locate the most similar region in the query image, followed by self-matching within the query image to predict the mask.

Our proposed MPA (1-shot) introduced in Sec. 4 integrates the SSP module to predict the query prototype P_q from the support prototype P_s and query feature F_q:

$$\hat{P}_q = SSP(F_q, P_s). \tag{A2}$$

Compared to previous methods, SSP produces a more representative prototype for the query [13], which is crucial for our design. Additional implementation details on integrating SSP into our method are provided in the supplementary materials.

A2.2. Gestalt Principle on Target Domains

Fan *et al.* [13] validate the presence of the Gestalt principle [28] in daily object datasets through cosine similarity statistics. Following their approach, we conduct similar experiments across four target domains in the data-scarce FSS datasets. The statistical results are presented in Tab. A1. Our findings confirm that, in all datasets, pixels within the same object exhibit significantly higher similarity compared to cross-object pixels. This indicates that the Gestalt principle remains valid, and the SSP method remains effective. For the FSS-1000 dataset, we compute cosine similarity using all images. For the remaining datasets, we randomly select 200 images per category, except for ISIC class 2, where we use 100 images due to data limitations.

Table A1. Cosine similarity for cross/intra-object pixels in four data-scarce target domains.

ForeGround Pixels Similarity							
Deepglobe		ISIC		Chest X-Ray		FSS-1000	
cross	intra	cross	intra	cross	intra	cross	intra
0.497	**0.552**	0.512	**0.526**	0.528	**0.554**	0.494	**0.563**

A2.3. Additional Details on Applying SSP

We adopt the Self-Support Prototype (SSP) [13] as our baseline while incorporating its key design components. To mitigate the impact of cluttered backgrounds, we integrate the adaptive self-support background prototype [13] into our framework. Additionally, we employ self-support refinement to enhance prediction accuracy, as demonstrated to be effective in [13]. For fair comparisons, all experiments in Sec. 5.4 are conducted under the same setup.

A3. Details of Sequential Chain of MPA

For j^{th} query image I_{q_j}, the computations proceed as follows ($1 \leq j \leq N$):

$$P_{q_j}^{seq} = SSP(F_{q_j}, P_s^{seq_{(j-1)}}), \quad P_s^{seq_j} = SSP(F_s, P_{q_j}^{seq}),$$
$$\hat{M}_{q_j}^{seq} = \sigma(Sim(F_{q_j}, P_{q_j}^{seq})), \hat{M}_s^{seq_j} = \sigma(Sim(F_s, P_s^{seq_j})). \tag{A3}$$

$P_{q_1}^{seq}$, $P_s^{seq_1}$, $\hat{M}_{q_1}^{seq}$ and $\hat{M}_s^{seq_1}$ in Eq. 2, 3, and 5 can be treated as $P_{q_j}^{seq}$, $P_s^{seq_j}$, $\hat{M}_{q_j}^{seq}$ and $\hat{M}_s^{seq_j}$ when $j = 1$. Then, corresponding supervisions can be obtained:

$$\mathcal{L}_{q_j}^{seq} = BCE(\hat{M}_{q_j}^{seq}, M_{q_j}), \ \mathcal{L}_s^{seq_j} = BCE(\hat{M}_s^{seq_j}, M_s). \tag{A4}$$

A3.1. Entire algorithm of MPA

The overall MPA algorithm is presented in Algo. A1.

A4. Analysis

A4.1. Visulization.

We provide more visualization examples in Fig. A3, and MPA predicts accurate segmentations.

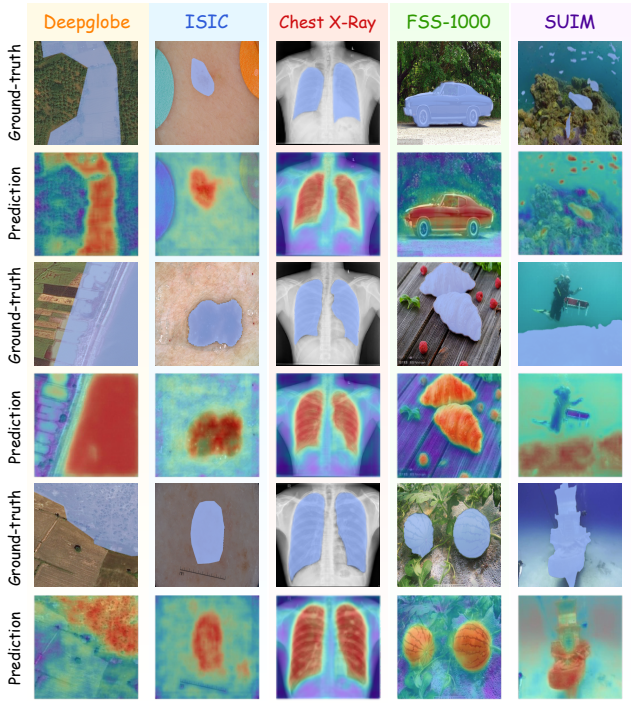

Figure A3. More qualitative visualizations of MPA.

A4.2. Effectiveness of Two Branches of DMP

We perform ablation studies on our designed multi-view prediction and supervisory branches, as shown in Tab. A2.

Algorithm A1 Multi-view Progressive Adaptation (1-shot).

Require: support image I_s, support ground-truth mask M_s, backbone model $\mathcal{G}_b$, and number of augmented queries N.
Ensure: Adapted target model $\mathcal{G}_t$.

$\mathcal{G}_b = \mathcal{G}_s$, $N = 1$.
while not reach the maximum epoch **do**
  initialize:
    $j^{seq} = 0$, $j^{par} = 1$, $\mathcal{L} = 0$.
  /* Augment N queries: */
  Augment N query images I_{q_1}, ..., I_{q_N} and ground-truth masks M_{q_1}, ..., M_{q_N} from I_s and M_s.
  /* Derive fundamental information: */
  Extract support feature F_s and query feature F_q from $\mathcal{G}_t$.
  Calculate P_s by Eqn. 2.
  Predict $\hat{M}_s$ by Eqn. 3.
  Calculate $\mathcal{L}_{bs}$ by Eqn. 4.
  /* Sequential multi-view predictions: */
  while $j^{seq} < N$ **do**
    Calculate $P^{seq}_{q_{(j^{seq}+1)}}$ and $P^{seq_{(j^{seq}+1)}}_{s}$ by Eqn. A3 (supp.).
    Predict $\hat{M}^{seq}_{q_i}$ and $\hat{M}^{seq_{(j^{seq}+1)}}_{s}$ by Eqn. A3 (supp.).
    Calculate $\mathcal{L}^{seq}$. by Eqn. 7.
    $j^{seq} = j^{seq} + 1$.
  end while
  /* Parallel multi-view predictions: */
  while $j^{par} < N$ **do**
    Calculate $P^{par}_{q_{(j^{par}+1)}}$ and $P^{par_{(j^{par}+1)}}_{s}$ by Eqn. 6.
    Predict $\hat{M}^{par}_{q_{(j^{par}+1)}}$ and $\hat{M}^{par_{(j^{par}+1)}}_{s}$ by Eqn. 6.
    Calculate $\mathcal{L}^{par}_{s}$ and $\mathcal{L}^{par}_{q}$ by Eqn. 8.
    $j^{par} = j^{par} + 1$.
  end while
  /* Model parameter updating: */
  Compute $\mathcal{L}$ via Eqn. 9 and optimize target model $\mathcal{G}_t$.
  /* Curriculum updating: */
  while meet the updating criteria **do**
    $N = N + 1$.
  end while
end while

The performance gains on Deepglobe and ISIC confirm the effectiveness of both sequential and parallel chains, reinforcing our assumption that comprehensively leveraging augmented views is crucial for adaptation.

Table A2. Ablation studies for two branches of MPA. Our proposed sequential and parallel multi-view prediction chains both show significant effectiveness.

Adaptation w/o Source-Training		
Incorporated design	Deepglobe	ISIC
Baseline	42.1	42.2
+ Sequential	50.1	69.3
+ Sequential + Parallel	**53.1**	**71.1**

A4.3. Effectiveness of Implicit Progressive Strategy of HPA

To validate our assumption and reaffirm the conclusion from IFA [40] that previous prediction errors accumulate and propagate to subsequent views, we analyze the sequential prediction chain. To clearly demonstrate the segmentation results, we establish a prediction chain in which six query images of the same category are predicted sequentially. The 1^{st} and 6^{th} query images are identical for result comparison, while the others differ. As illustrated in Fig. A4, the segmentation results of the 6^{th} query image are significantly worse than those of the 1^{st} query image.

A4.4. Hyper-Parameter Values

We conduct experiments to determine the value of λ_{bs}, λ^{seq}, λ^{par}_{s}, λ^{par}_{q}, and N, which balance the loss terms and limit the maximum number of augmented query views discussed in Sec. 4. Specifically, we utilize MPA to adapt the backbone model to the two datasets, Deepglobe and ISIC. (1-shot setup with a ResNet-50 backbone). As shown in Tab. A3, the performances of both datasets remain robust to variations in these hyper-parameters.

Table A3. Impact of the hyper-parameters λ_{bs}, λ^{seq}, λ^{par}_{s}, λ^{par}_{q}, and N which denote the weight of $\mathcal{L}_{bs}$, $\mathcal{L}^{seq}$, $\mathcal{L}^{par}_{s}$, $\mathcal{L}^{par}_{q}$, and the maximum number of augmented query views.

Adapt Backbone Model to Deepglobe w/o Source-Training									
λ_{bs}	0.1	**0.2**	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	51.7	**53.1**	52.1	52.0	51.9	51.8	52.2	50.9	52.1
λ^{seq}	**0.1**	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	**53.1**	52.6	51.3	52.0	51.5	52.0	52.6	52.4	51.1
λ^{par}_{s}	0.1	0.2	0.3	**0.4**	0.5	0.6	0.7	0.8	0.9
mIoU	70.9	70.3	70.5	**71.1**	70.2	70.6	69.8	71.0	69.4
λ^{par}_{q}		0.5		**1.0**		1.5		2.0	
mIoU		70.5		**71.1**		70.2		70.7	
N	1	2	3	4	5	**6**	7	8	9
mIoU	50.6	51.0	51.8	51.4	52.6	**53.1**	52.2	51.1	50.9

Adapt Backbone Model to ISIC w/o Source-Training									
λ_{bs}	0.1	**0.2**	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	69.9	**71.1**	70.4	70.7	70.1	69.8	69.2	70.5	70.6
λ^{seq}	**0.1**	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	**71.1**	70.8	70.2	70.5	69.3	69.9	70.4	70.0	69.8
λ^{par}_{s}	0.1	0.2	0.3	**0.4**	0.5	0.6	0.7	0.8	0.9
mIoU	70.9	70.3	70.5	**71.1**	70.2	70.6	69.8	70.9	69.8
λ^{par}_{q}		0.5		**1.0**		1.5		2.0	
mIoU		70.5		**71.1**		70.2		70.3	
N	1	2	3	4	5	**6**	7	8	9
mIoU	70.6	70.7	70.9	70.8	70.9	**71.1**	70.8	70.2	69.9

Table A4. Performances of IFA with different augmentation operations. MPA consistently surpasses all variants of IFA.

Dataset	Ours	IFA	IFA+easy aug.	IFA+difficult aug.
Deepglobe	**53.1**	50.3	48.5	49.1
ISIC	**71.1**	66.3	62.8	64.0

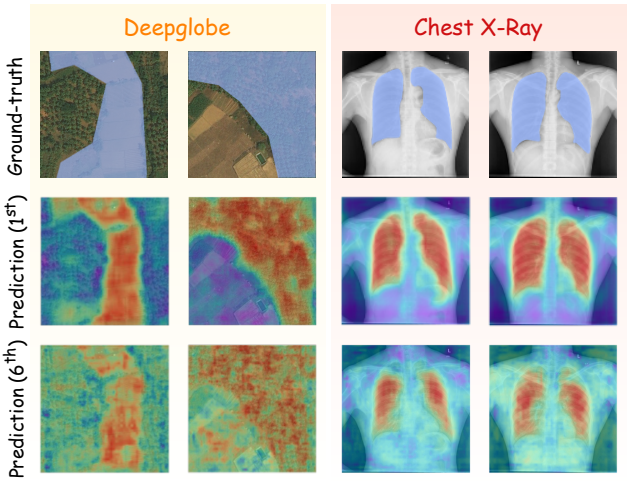

Figure A4. Qualitative illustrations over samples from Deepglobe and Chest X-Ray. For each pair of query image (ground truth highlighted by blue), Rows 2-3 show the corresponding segmentation heatmaps of the 1^{st} query image and 6^{th} query image within the same sequential multi-view prediction chain, respectively. Best viewed in color.

A4.5. Differences from IFA

(i) Motivation and design. As observed in Fig. 1, **directly training** over multiple views is suboptimal because strongly augmented views cannot be effectively used. Our proposed MPA addresses this by **progressively** leveraging **multiple augmented views**, which is fundamentally different from IFA [40] that utilizes only **a single augmented view**. **(ii) Performance.** We evaluate IFA under both easy and difficult augmentation settings as in Tab. A4. MPA outperforms both variants. This shows that IFA with easy augmentation tends to overfit, while IFA with difficult augmentation suffers from a large "support–query" gap. Moreover, due to the superior design of MPA, the source-training can be skipped to save time (refer to Tab. 9), which is not achievable with IFA.