DA-PROTO: DISTRIBUTION-ALIGNED PROTOTYPE LEARNING FOR CROSS-DOMAIN FEW-SHOT MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-Shot Medical Image Segmentation (FSMIS) aims to segment u nseen categories with only a few annotated samples. However, domain shifts across imaging devices, protocols, and modalities can severely degrade target-domain performance, motivating Cross-Domain FSMIS (CD-FSMIS). Existing methods mainly exploit cross-domain stable cues in frequency, anatomical structure, or semantic channels, while largely overlooking the support-query foreground distributional relationship. Notably, we observe that support and query images of the same category exhibit consistent foreground distributions within each domain, and this relationship remains stable across domain shifts such as CT $\rightarrow$ MR. Based on this observation, we propose Distribution-Aligned Prototype Learning for Cross-Domain Few-Shot Medical Image Segmentation (DA-Proto), which extracts a stable foreground distribution prior from target-domain support samples and uses it to modulate domain-shifted class prototypes into distribution-aligned prototypes. Extensive experiments across multiple cross-domain settings demonstrate that DA-Proto outperforms state-of-the-art methods. Code is available at: `https://anonymous.4open.science/r/DA_Proto-E8B2`.

1 INTRODUCTION

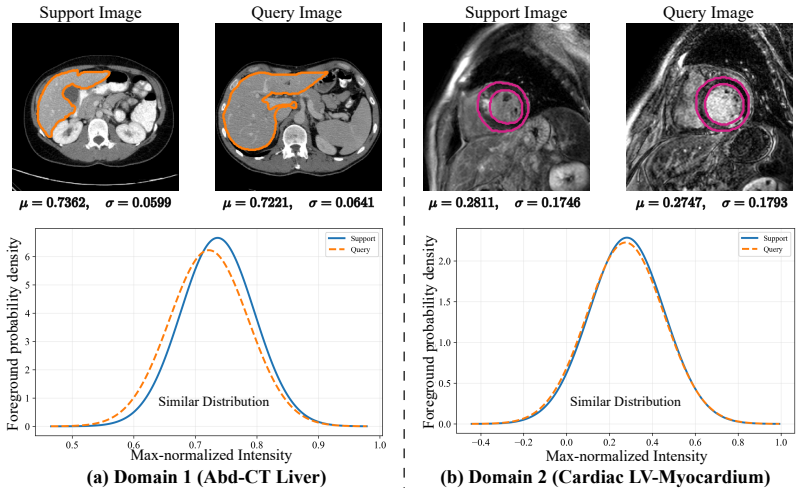
Automatic medical image segmentation plays an important role in clinical applications such as computer-aided diagnosis (Tsochatzidis et al., 2021) and treatment planning (Zhang et al., 2024). Despite the remarkable progress of deep learning-based segmentation models, their success often relies on large-scale annotated datasets. In practice, however, annotations for many organs or lesions are extremely limited, making conventional supervised methods (Cao et al., 2022; Zhou et al., 2025) prone to overfitting and poor generalization to unseen categories.

Few-Shot Medical Image Segmentation (FSMIS) provides a promising solution to this problem. The goal of FSMIS is to learn transferable segmentation knowledge from seen categories and rapidly adapt to novel categories at test time using only a few annotated support images. In this way, the model can segment corresponding query images even when only very limited annotations are available for the target category.

However, most existing FSMIS methods (Ouyang et al., 2022; Wang et al., 2019) are mainly designed for single-domain scenarios and are therefore vulnerable to the domain shifts (Kondrateva et al., 2021) commonly encountered in real-world clinical deployment. In the cross-domain setting (Bo et al., 2025b), a model is trained only on a source domain, while both the support set and query set at test time come from an unseen target domain. Due to differences in imaging devices, acquisition protocols, and imaging modalities, the source and target domains may exhibit substantial distribution shifts. For example, transferring a model from MRI to CT can alter image appearance and feature responses due to fundamentally different imaging mechanisms, thereby degrading the representation and generalization ability of the model in the target domain.

Existing Cross-Domain Few-Shot Medical Image Segmentation (CD-FSMIS) methods mainly focus on learning domain-invariant representations. For example, FAMNet (Bo et al., 2025b) exploits cross-domain information across different frequency bands to enhance domain robustness; C-Graph (Bo et al., 2025a) leverages the relative stability of anatomical shapes across domains and mod-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Domain 1 (Abd-CT Liver) **(b) Domain 2 (Cardiac LV-Myocardium)**

Figure 1: Illustration of our motivation. The foreground distributions of support and query images exhibit a stable relative relationship across different domains. (a) In Abdomen CT scans, the liver shows high foreground distributional consistency between support and query images. (b) A similar consistency is observed for the LV-Myocardium in Cardiac MRI scans. This observation motivates us to exploit foreground distribution as a statistical prior for cross-domain segmentation.

els category relationships through graph structures; DSM (Zhu et al., 2025b) dynamically selects semantically consistent feature channels to improve robustness. However, the relative foreground distributional relationship between support and query images has received much less attention.

Notably, we observe a key phenomenon: the foreground statistical distributions of support and query images within the same domain exhibit a stable relationship. As illustrated in Fig. 1, this relative support-query distributional relationship is preserved in both source and target domains. This observation motivates us to revisit prototype learning in cross-domain few-shot segmentation. Conventional prototype-based methods typically aggregate foreground features from support images to construct a class prototype and then match it against query features. However, since the feature extractor is mainly optimized on the source domain, feature shifts in the target domain can be directly propagated into the prototype representation, thereby weakening its ability to characterize the target category. Based on the above observation, we argue that the foreground distribution prior contained in target-domain support samples can be exploited to dynamically calibrate domain-shifted class prototypes, yielding representations that are better adapted to the target domain.

In this paper, we propose Distribution-Aligned Prototype Learning for Cross-Domain Few-Shot Medical Image Segmentation (DA-Proto) which consists of three main components. First, a Foreground Distribution Modeling (FDM) module captures foreground statistical priors from support samples. Second, a Distribution Feature Reconstruction (DFR) module regularizes the learned distribution representation by reconstructing foreground statistics. Finally, a Distribution-guided Prototype Modulator (DGPM) uses the learned prior to calibrate domain-shifted class prototypes, producing more target-adaptive representations. In summary, our contributions are as follows:

- We reveal a key phenomenon in CD-FSMIS setting: the support-query foreground distributional relationship remains consistently stable across different domains, providing a novel statistical prior for cross-domain category modeling.
- Based on this observation, we propose DA-Proto, a distribution-aligned prototype learning framework for cross-domain few-shot medical image segmentation. By integrating foreground distribution modeling, distribution reconstruction, and distribution-guided prototype modulation, DA-Proto explicitly injects target-domain distribution priors into class prototypes to alleviate prototype representation bias caused by domain shift.
- Extensive experiments demonstrate that DA-Proto consistently improves target-domain generalization across multiple cross-domain few-shot medical image segmentation settings. Ablation studies and mechanism analyses further validate the effectiveness of the foreground distribution prior and each component of the proposed framework.

2 Related Works

2.1 Few Shot Medical Image Segmentation (FSMIS)

FSMIS aims to segment unseen categories from only a few annotated samples. Most methods (Tang et al., 2021; Lin et al., 2023; Sun et al., 2022) follow episodic prototype learning to construct transferable category representations. PANet (Wang et al., 2019) introduces prototype alignment, while SSL-ALPNet (Ouyang et al., 2022), Q-Net (Shen et al., 2023), and RPT (Zhu et al., 2023) further improve prototype representations through adaptive local prototypes, query-guided refinement, and regional modeling, respectively. Recent methods such as GMRD (Cheng et al., 2024), CoW (Jiang & Zhang, 2025), and BePMI (Xu et al., 2026) explore multiple prototypes, hard-region modeling, and boundary-aware representations. However, these methods are developed for single-domain settings and do not explicitly address cross-domain shifts.

2.2 Cross Domain Few Shot Medical Image Segmentation (CD-FSMIS)

CD-FSMIS extends conventional FSMIS to scenarios with substantial source-target domain shifts, requiring models to adapt to unseen target-domain categories from only a few annotated support samples. Existing methods improve cross-domain robustness from different perspectives. FAMNet (Bo et al., 2025b) exploits transferable information across frequency bands to reduce domain-induced feature variations, while RobustEMD (Zhu et al., 2025a) enhances support-query feature matching under domain shift. DSM (Zhu et al., 2025b) selects semantically consistent channels that are more stable across domains, and C-Graph (Bo et al., 2025a) leverages anatomical structures and organ shapes to model cross-domain semantic relationships. Training-free SAM-based approaches have also been explored (Zhu & Zhang, 2025), but follow a different setting from our source-trained framework. Despite these advances, the stable foreground distributional relationship between support and query images remains underexplored. Since labeled support samples directly reveal category-specific foreground statistics in the target domain, we exploit this support-derived distribution as a statistical anchor to calibrate domain-shifted class prototypes.

2.3 Prototype Distribution Modeling

Recent studies have introduced distribution modeling into few-shot learning (Wang et al., 2021; Liu et al., 2025). These methods mainly model or calibrate prototypes from a probabilistic perspective. In contrast, we explicitly extract foreground distribution statistics from target-domain support samples and use them as a prior to modulate domain-shifted prototypes.

3 Methodology

3.1 Problem Formulation

Cross-Domain Few-Shot Medical Image Segmentation (CD-FSMIS) aims to transfer a segmentation model Θ, trained on a source domain $\mathcal{D}_s$ with labeled base categories $\mathcal{C}_b$, to an unseen target domain $\mathcal{D}_t$ containing novel categories $\mathcal{C}_n$, using only a few annotated support samples. The source and target domains exhibit distribution shifts due to differences in imaging modalities and acquisition conditions, while their category sets are mutually exclusive, i.e., $\mathcal{C}_b \cap \mathcal{C}_n = \emptyset$. The trained model is directly evaluated on $\mathcal{D}_t$ without retraining or fine-tuning.

Following the episodic meta-learning paradigm, each training episode is sampled from the source domain and consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set is defined as $\mathcal{S} = \{(\mathbf{I}_i^s, \mathbf{M}_i^s(c))\}_{i=1}^{S}$, where $\mathbf{I}_i^s$ denotes a support image, $\mathbf{M}_i^s(c) \in \{0,1\}^{H\times W}$ is the corresponding binary segmentation mask for category $c \in \mathcal{C}_b$, and S denotes the number of support shots. Following FAMNet, we use the 1-shot setting ($S = 1$). The query set is defined as $\mathcal{Q} = \{(\mathbf{I}_q, \mathbf{M}_q(c))\}$, where $\mathbf{I}_q$ denotes the query image and $\mathbf{M}_q(c)$ is its ground-truth segmentation mask.

At inference time, both support and query images are sampled from unseen target domain $\mathcal{D}_t$, with categories drawn from $\mathcal{C}_n$. Thus, the cross-domain setting refers to the shift between source-domain training and target-domain testing. Given a few annotated support samples from the target domain, the model infers the segmentation masks for unlabeled target-domain query samples.

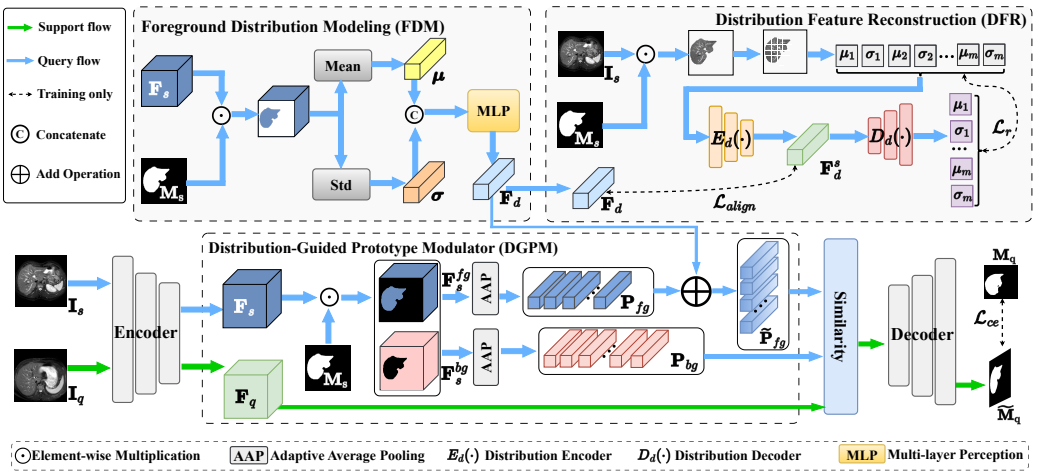


Figure 2: Overview of DA-Proto. (1)The FDM module captures the foreground statistical prior from the support sample; (2) The DFR module reconstructs foreground statistics to regularize the learned distribution representation during training; (3) The DGPM module then employs the learned distribution prior to calibrate the foreground prototype, producing a distribution-aligned prototype for similarity matching and query segmentation.

3.2 Network Overview

We follow the standard cross-domain few-shot medical image segmentation setting adopted in FAM-Net (Bo et al., 2025b), in which the model is trained episodically on a source domain. At test time, both the support and query images are sampled from the target domain. Given a support image $\mathbf{I}_s$, its corresponding foreground mask $\mathbf{M}_s \in \mathbb{R}^{H\times W}$, and a query image $\mathbf{I}_q$, a shared encoder f_θ extracts the support and query features as:

$$\mathbf{F}_s = f_\theta(\mathbf{I}_s), \quad \mathbf{F}_q = f_\theta(\mathbf{I}_q),$$

where $\mathbf{I}_s, \mathbf{I}_q \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{F}_s, \mathbf{F}_q \in \mathbb{R}^{H\times W\times C}$, C denotes the channel dimension, and H, W denote the spatial dimensions.

As illustrated in Fig. 2, DA-Proto consists of three main components. First, the Foreground Distribution Modeling (FDM) module extracts foreground statistical information from the support sample and learns a corresponding distribution representation $\mathbf{F}_d \in \mathbb{R}^{1\times C}$. Second, the Distribution Feature Reconstruction (DFR) module regularizes $\mathbf{F}_d$ by reconstructing the foreground statistics, encouraging the learned representation to preserve meaningful category-specific distributional information. DFR is used only during training. Finally, the Distribution-Guided Prototype Modulator (DGPM) uses the learned distribution prior to adaptively modulate the foreground prototype set $\mathbf{P}_{fg} = \{\mathbf{p}_{fg}^k\}_{k=1}^{N_{fg}}$, producing a distribution-aligned prototype set $\widetilde{\mathbf{P}}_{fg} = \{\tilde{\mathbf{p}}_{fg}^k\}_{k=1}^{N_{fg}}$, thereby alleviating prototype bias induced by domain shift. The modulated foreground prototype set $\widetilde{\mathbf{P}}_{fg}$ and the background prototype set $\mathbf{P}_{bg} = \{\mathbf{p}_{bg}^k\}_{k=1}^{N_{bg}}$ are then matched with the query features $\mathbf{F}_q$ to produce the final segmentation prediction $\widetilde{\mathbf{M}}_q \in \mathbb{R}^{H\times W}$. Here, N_{fg} and N_{bg} denote the numbers of foreground and background prototypes, respectively.

3.3 Foreground Distribution Modeling (FDM)

Motivated by the observation that support and query images of the same category exhibit consistent foreground distributions within each domain, we design a **Foreground Distribution Modeling (FDM)** module to capture the foreground distribution of the annotated support sample and use it as a category-specific statistical prior for subsequent prototype modulation.

Specifically, given the support feature $\mathbf{F}_s$ and its corresponding foreground mask $\mathbf{M}_s$, we first extract the valid foreground features:

$$\mathbf{F}_s^{fg} = \{\mathbf{F}_s(h,w) \mid \mathbf{M}_s(h,w) = 1\} \in \mathbb{R}^{n\times C}, \tag{1}$$

where n denotes the number of valid foreground locations; and C denotes the channel dimension.

We then compute the channel-wise mean $\boldsymbol{\mu}_s$ and standard deviation $\boldsymbol{\sigma}_s$ over all foreground locations to characterize the foreground feature distribution:

$$\boldsymbol{\mu}_s = \mathrm{Mean}\big(\mathbf{F}_s^{fg}\big) \in \mathbb{R}^{1\times C}, \qquad \boldsymbol{\sigma}_s = \mathrm{Std}\big(\mathbf{F}_s^{fg}\big) \in \mathbb{R}^{1\times C}. \tag{2}$$

The two statistics are concatenated to form a foreground statistical descriptor $\mathbf{Z}_f = [\boldsymbol{\mu}_s; \boldsymbol{\sigma}_s] \in \mathbb{R}^{1\times 2C}$. We further employ a two-layer MLP, denoted as $E_{mlp}(\cdot)$, to project $\mathbf{Z}_f$ into the prototype feature space and get the distribution representation $\mathbf{F}_d$:

$$\mathbf{F}_d = E_{mlp}(\mathbf{Z}_f) \in \mathbb{R}^{1\times C}. \tag{3}$$

The resulting $\mathbf{F}_d$ provides a compact representation of the support foreground distribution and is subsequently used as the statistical prior for prototype modulation. However, directly learning $\mathbf{F}_d$ from foreground feature statistics does not explicitly guarantee that it retains meaningful foreground distribution semantics. To this end, we further introduce the Distribution Feature Reconstruction module to regularize $\mathbf{F}_d$ using explicit statistics extracted from the support foreground.

3.4 Distribution Feature Reconstruction (DFR)

To explicitly regularize the distribution representation $\mathbf{F}_d$ learned by FDM, we design the **Distribution Feature Reconstruction (DFR)** module. DFR constructs explicit foreground statistics from the support image and uses them to provide both reconstruction and representation-level supervision.

Specifically, given the support image $\mathbf{I}_s$ and its foreground mask $\mathbf{M}_s$, we first identify the valid foreground locations as $\Omega_{fg} = \{(h,w) \mid \mathbf{M}_s(h,w) = 1\}$. We then spatially partition Ω_{fg} into m non-overlapping local regions $\{\Omega_j\}_{j=1}^m$, where $\Omega_{fg} = \bigcup_{j=1}^m \Omega_j$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. For each region, only the valid foreground pixels are retained:

$$\mathbf{I}_{s,j}^{fg} = \{\mathbf{I}_s(h,w) \mid (h,w) \in \Omega_j\}, \qquad j = 1, \ldots, m. \tag{4}$$

To reduce the influence of absolute intensity scale across regions, we further normalize each local foreground region by its maximum pixel intensity, where ϵ is a small constant for numerical stability:

$$\bar{\mathbf{I}}_{s,j}^{fg} = \frac{\mathbf{I}_{s,j}^{fg}}{\max\left(\mathbf{I}_{s,j}^{fg}\right) + \epsilon}. \tag{5}$$

For each normalized foreground region $\bar{\mathbf{I}}_{s,j}^{fg}$, we compute its mean μ_j and standard deviation σ_j. Then we concatenate the mean and standard deviation of all foreground regions to construct an explicit foreground statistical representation $\mathbf{Z}_s$:

$$\begin{cases} \mu_j = \mathrm{Mean}\left(\bar{\mathbf{I}}_{s,j}^{fg}\right), \qquad \sigma_j = \mathrm{Std}\left(\bar{\mathbf{I}}_{s,j}^{fg}\right), \quad j = 1, \ldots, m, \\ \mathbf{Z}_s = [\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_m, \sigma_m] \in \mathbb{R}^{1\times 2m}. \end{cases} \tag{6}$$

The resulting $\mathbf{Z}_s$ preserves the local intensity statistics of different foreground regions while providing a compact description of the overall foreground distribution. Next, we encode $\mathbf{Z}_s$ into a compact statistical representation $\mathbf{F}_d^s$ using a MLP-based encoder $E_s(\cdot)$:

$$\mathbf{F}_d^s = E_s\left(\mathbf{Z}_s\right) \in \mathbb{R}^{1\times C}. \tag{7}$$

To ensure that $\mathbf{F}_d^s$ retains the foreground statistics contained in $\mathbf{Z}_s$, a MLP-based decoder $D_s(\cdot)$ is employed to reconstruct the original statistical representation:

$$\widehat{\mathbf{Z}}_s = D_s\left(\mathbf{F}_d^s\right) \in \mathbb{R}^{1\times 2m}. \tag{8}$$

We employ a reconstruction loss $\mathcal{L}_r$ to minimize the mean ℓ_1 distance between the reconstructed statistics and the explicit statistics. This reconstruction constraint encourages the encoded representation $\mathbf{F}_d^s$ to preserve the foreground distribution information contained in the regional statistics:

$$\mathcal{L}_r = \frac{1}{2m}\left\|\widehat{\mathbf{Z}}_s - \mathbf{Z}_s\right\|_1. \tag{9}$$

Furthermore, we align the distribution representation $\mathbf{F}_d$ produced by FDM with the statistical representation $\mathbf{F}_d^s$ encoded from the support foreground statistics. We adopt cosine distance $\cos(\cdot)$ for this representation alignment:

$$\mathcal{L}_{align} = 1 - \cos\left(\mathbf{F}_d, \mathbf{F}_d^s\right). \tag{10}$$

By minimizing $\mathcal{L}_{align}$, the distribution feature learned from support features is encouraged to align with the representation encoded from explicit foreground statistics, thereby enriching $\mathbf{F}_d$ with more explicit distribution semantics.

Notably, DFR is used only during training. At inference time, the statistical encoding and reconstruction branches are removed, while the distribution feature $\mathbf{F}_d$ generated by FDM is directly passed to the subsequent Distribution-Guided Prototype Modulator.

3.5 Distribution-Guided Prototype Modulator (DGPM)

Given the foreground distribution representation $\mathbf{F}_d$ learned by FDM and regularized by DFR, we introduce the **Distribution-Guided Prototype Modulator (DGPM)** to calibrate the foreground prototypes with the support-derived distribution prior $\mathbf{F}_d$.

Specifically, given the support feature $\mathbf{F}_s$ and its foreground mask $\mathbf{M}_s$, we first extract the valid foreground and background features according to $\mathbf{M}_s$. Adaptive 1D pooling is then applied to construct foreground prototype set $\mathbf{P}_{fg}$ and background prototype set $\mathbf{P}_{bg}$:

$$\mathbf{P}_{fg} = \left\{\mathbf{p}_{fg}^k\right\}_{k=1}^{N_{fg}} \in \mathbb{R}^{N_{fg}\times C}, \qquad \mathbf{P}_{bg} = \left\{\mathbf{p}_{bg}^k\right\}_{k=1}^{N_{bg}} \in \mathbb{R}^{N_{bg}\times C}. \tag{11}$$

Since $\mathbf{F}_d \in \mathbb{R}^{1\times C}$ characterizes the foreground distribution, we use it only to modulate the foreground prototypes. Specifically, $\mathbf{F}_d$ is broadcast to each foreground prototype and added to $\mathbf{P}_{fg}$:

$$\widetilde{\mathbf{P}}_{fg} = \mathbf{P}_{fg} + \mathbf{F}_d \in \mathbb{R}^{N_{fg}\times C}, \tag{12}$$

where $\widetilde{\mathbf{P}}_{fg}$ denotes the distribution-aligned foreground prototype set, while the background prototype set $\mathbf{P}_{bg}$ remains unchanged.

We then compute the cosine similarities between the query feature $\mathbf{F}_q$ and all foreground/background prototypes, yielding foreground similarity maps $\mathbf{S}_{fg} \in \mathbb{R}^{N_{fg}\times H\times W}$ and background similarity maps $\mathbf{S}_{bg} \in \mathbb{R}^{N_{bg}\times H\times W}$:

$$\mathbf{S}_{fg}(j,h,w) = \cos\left(\widetilde{\mathbf{p}}_{fg}^j, \mathbf{F}_q(h,w)\right), \quad \mathbf{S}_{bg}(j,h,w) = \cos\left(\mathbf{p}_{bg}^j, \mathbf{F}_q(h,w)\right). \tag{13}$$

Since different query locations may correspond to different local prototypes, we perform pixel-wise Top-k selection along the prototype dimension. For each query location (h,w), we retain the k_{fg} largest foreground responses and k_{bg} largest background responses, and sum them to obtain the aggregated foreground and background similarity maps $\hat{\mathbf{S}}_{fg}, \hat{\mathbf{S}}_{bg} \in \mathbb{R}^{1\times H\times W}$:

$$\begin{cases} \hat{\mathbf{S}}_{fg}(h,w) = \sum \mathrm{TopK}\left(\mathbf{S}_{fg}(:,h,w), k_{fg}\right), \\ \hat{\mathbf{S}}_{bg}(h,w) = \sum \mathrm{TopK}\left(\mathbf{S}_{bg}(:,h,w), k_{bg}\right). \end{cases} \tag{14}$$

Finally, the aggregated , $\hat{\mathbf{S}}_{bg}$ and $\hat{\mathbf{S}}_{fg}$ are normalized by Softmax along the class dimension, yielding background query prediction $\widetilde{\mathbf{M}}_q^{bg} \in \mathbb{R}^{H\times W}$ and foreground query prediction $\widetilde{\mathbf{M}}_q^{fg} \in \mathbb{R}^{H\times W}$:

$$\widetilde{\mathbf{M}}_q = \left[\widetilde{\mathbf{M}}_q^{bg}, \widetilde{\mathbf{M}}_q^{fg}\right] = \mathrm{Softmax}\left(\left[\hat{\mathbf{S}}_{bg}, \hat{\mathbf{S}}_{fg}\right]\right), \tag{15}$$

3.6 Loss Function

To supervise the query segmentation, we employ a Cross-Entropy (Gordon-Rodriguez et al., 2020) loss $\mathcal{L}_{ce}$ between the predicted masks $(\widetilde{\mathbf{M}}_q^{bg}, \widetilde{\mathbf{M}}_q^{fg})$ and the ground-truth mask $\mathbf{M}_q$:

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{HW}\sum_{h,w}\left[\mathbf{M}_q^{bg}\log\widetilde{\mathbf{M}}_q^{bg} + \mathbf{M}_q^{fg}\log\widetilde{\mathbf{M}}_q^{fg}\right]. \tag{16}$$

In summary, the overall objective combines the segmentation loss with the alignment and reconstruction losses, weighted by λ_{align} and λ_r, respectively: $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ce}} + \lambda_{align}\mathcal{L}_{\mathrm{align}} + \lambda_r\mathcal{L}_r$.

Table 1: Quantitative comparison on the Cross-Modality tasks in terms of Dice score (%, mean ± std). The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Reference	Abdominal CT → MRI					Abdominal MRI → CT				
		Liver	LK	RK	Spleen	Mean	Liver	LK	RK	Spleen	Mean
SSL-ALP	TMI'22	73.35 ± 04.22	62.01 ± 05.77	70.70 ± 04.43	57.59 ± 10.06	65.91 ± 03.64	76.47 ± 05.52	63.86 ± 13.38	57.51 ± 12.51	67.79 ± 08.39	66.41 ± 08.86
PATNet	ECCV'22	67.39 ± 04.41	58.37 ± 07.54	64.81 ± 04.95	55.01 ± 07.30	61.39 ± 04.60	67.60 ± 06.41	45.37 ± 10.33	50.92 ± 10.84	50.07 ± 11.29	53.49 ± 07.14
QNet	IntelliSys'23	48.95 ± 13.30	53.23 ± 08.98	62.11 ± 11.74	56.62 ± 07.88	55.23 ± 04.47	74.10 ± 03.84	64.22 ± 06.92	59.23 ± 14.41	65.75 ± 07.62	65.82 ± 04.92
RPT	MICCAI'23	66.93 ± 09.21	56.99 ± 10.51	57.18 ± 16.16	56.61 ± 07.35	59.43 ± 06.37	**79.87 ± 01.26**	58.62 ± 03.57	60.37 ± 14.08	68.89 ± 09.25	66.94 ± 04.75
IFA	CVPR'24	72.69 ± 03.24	53.24 ± 08.03	51.28 ± 07.46	54.38 ± 08.59	57.90 ± 04.36	71.36 ± 05.94	56.72 ± 16.45	49.15 ± 21.78	60.63 ± 08.06	59.46 ± 11.39
TPM	MICCAI'25	65.58 ± 05.03	61.46 ± 07.72	68.99 ± 08.62	57.03 ± 07.22	63.26 ± 05.39	72.74 ± 07.58	53.78 ± 15.17	54.54 ± 13.58	62.52 ± 05.02	60.89 ± 05.35
FAMNet	AAAI'25	**75.23 ± 04.73**	62.64 ± 08.87	68.19 ± 09.12	62.32 ± 10.29	67.10 ± 06.29	74.79 ± 05.54	57.76 ± 09.70	56.19 ± 18.84	63.38 ± 06.93	63.03 ± 08.01
DSM	TIP'25	69.19 ± 08.26	61.01 ± 11.23	71.16 ± 10.01	54.69 ± 11.05	64.01 ± 07.33	77.69 ± 03.94	59.18 ± 04.62	59.53 ± 21.75	69.90 ± 10.27	66.57 ± 08.35
C-Graph	TMI'26	68.73 ± 04.07	69.99 ± 08.85	74.66 ± 11.60	62.46 ± 11.02	68.96 ± 07.11	71.74 ± 08.67	66.14 ± 04.29	64.63 ± 06.75	67.36 ± 05.95	67.47 ± 04.55
Ours	—	74.56 ± 03.09	**73.26 ± 04.30**	**85.75 ± 04.84**	**64.98 ± 11.33**	**74.64 ± 04.41**	79.23 ± 04.77	**72.99 ± 07.75**	**71.65 ± 13.73**	**74.55 ± 03.93**	**74.61 ± 05.84**

Table 2: Quantitative comparison on the Cross-Sequence tasks in terms of Dice score (%, mean ± std). The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Reference	Cardiac LGE → b-SSFP				Cardiac b-SSFP → LGE			
		LV-BP	LV-MYO	RV	Mean	LV-BP	LV-MYO	RV	Mean
SSL-ALP	TMI'22	84.89 ± 02.89	52.92 ± 06.19	74.97 ± 04.94	70.93 ± 04.97	70.15 ± 05.02	50.25 ± 03.68	62.24 ± 03.64	60.88 ± 03.79
PATNet	ECCV'22	76.29 ± 08.11	24.77 ± 05.22	56.88 ± 12.80	52.65 ± 08.07	51.58 ± 16.03	27.63 ± 08.22	46.60 ± 04.53	41.94 ± 08.36
QNet	IntelliSys'23	85.08 ± 02.27	40.38 ± 05.71	71.90 ± 01.84	65.78 ± 01.39	68.10 ± 07.32	41.93 ± 07.21	56.65 ± 07.48	55.56 ± 05.86
RPT	MICCAI'23	83.06 ± 02.43	48.55 ± 05.16	71.76 ± 05.89	67.79 ± 03.58	75.50 ± 05.60	49.05 ± 04.01	60.18 ± 05.50	61.58 ± 03.29
IFA	CVPR'24	78.03 ± 10.50	44.69 ± 06.61	65.96 ± 10.58	62.89 ± 08.72	71.15 ± 11.02	46.54 ± 03.99	54.01 ± 12.77	57.23 ± 08.37
TPM	MICCAI'25	85.31 ± 02.70	40.41 ± 10.59	74.60 ± 03.56	66.77 ± 04.60	64.98 ± 10.70	45.37 ± 11.51	61.00 ± 06.97	57.12 ± 08.01
FAMNet	AAAI'25	87.34 ± 02.04	49.76 ± 01.23	74.57 ± 03.26	70.56 ± 01.96	**76.76 ± 06.36**	47.54 ± 03.99	56.92 ± 05.98	60.40 ± 03.75
DSM	TIP'25	87.85 ± 03.65	49.79 ± 03.33	73.53 ± 09.31	70.39 ± 05.18	67.93 ± 01.86	45.62 ± 04.71	50.90 ± 11.38	54.82 ± 04.72
C-Graph	TMI'26	86.62 ± 04.80	52.69 ± 04.41	75.88 ± 04.05	71.73 ± 03.37	70.82 ± 06.63	52.21 ± 05.54	62.57 ± 07.36	61.87 ± 05.27
Ours	—	**92.05 ± 01.86**	**66.90 ± 01.37**	**86.08 ± 03.58**	**81.68 ± 01.78**	71.64 ± 04.19	**56.86 ± 04.43**	**67.13 ± 01.83**	**65.21 ± 02.89**

4 EXPERIMENTS

4.1 DATASETS AND SETTING

Datasets. We evaluate DA-Proto on two widely used medical image segmentation benchmarks, covering cross-modality and cross-sequence domain shifts. (1) The **Cross-Modality** benchmark consists of **Abd-MR** (Kavur et al., 2021) with T2-SPIR 3D MR scans for 20 patients and **Abd-CT** (Landman et al., 2015) with 3D CT scans from 30 patients, evaluating Liver, Left Kidney (LK), Right Kidney (RK), and Spleen. (2) The **Cross-Sequence** benchmark (Zhuang et al., 2022) contains 3D scans from 45 patients for **Cardiac LGE** and 45 patients for **Cardiac bSSFP**, evaluating LV-Myocardium (LV-MYO), Left Ventricle (LV), and Right Ventricle (RV).

Setting. We construct **four cross-domain tasks**: Abd-CT → MR and Abd-MR → CT for cross-modality evaluation, and Cardiac LGE → bSSFP and Cardiac bSSFP → LGE for cross-sequence evaluation. For a task $\mathcal{S} \rightarrow \mathcal{T}$, the model is trained episodically only on the source domain $\mathcal{S}$ and directly evaluated on the unseen target domain $\mathcal{T}$, where both support and query samples are drawn from $\mathcal{T}$. We employ Dice Similarity Coefficient (DSC) (Setiawan, 2020) as the evaluation metric.

4.2 IMPLEMENTATION DETAILS

We train DA-Proto for 10 epochs with 5,000 episodes per epoch using SGD (learning rate 0.001, momentum 0.9), with a decay factor of 0.9 every 1,000 iterations. We use a COCO-pretrained ResNet-50 backbone (Lin et al., 2014b;a) with $C = 512$. Training episodes are constructed using 3D superpixel pseudo-labels (Ouyang et al., 2022) without ground-truth annotations. For evaluation, we adopt 5-fold cross-validation, using the third patient in each fold as the support patient and the remaining patients as queries. For each target organ in each fold, the middle slice and its mask from the support patient are used as the support pair. We set number of local regions $m = 16$, the loss weights (λ_{align} and λ_r) to 0.1 and 1.0, $N_{fg} = 200$, $N_{bg} = 400$, $k_{fg} = 10$, and $k_{bg} = 20$. All methods are re-trained and evaluated under the same data splits and protocol for fair comparison.

4.3 COMPARISON WITH SOTA METHODS

We compare our proposed method (DA-Proto) with representative few-shot segmentation methods, including SSL-ALP (Ouyang et al., 2022), PATNet (Lei et al., 2022), QNet (Shen et al., 2023), RPT (Zhu et al., 2023), IFA (Nie et al., 2024), TPM (Kim et al., 2025) , FAMNet (Bo et al., 2025b), DSM (Zhu et al., 2025b), and C-Graph (Bo et al., 2025a). To ensure a fair comparison, we retrained all methods using the same data splits and evaluation protocol, and conducted 5-fold cross-validation.

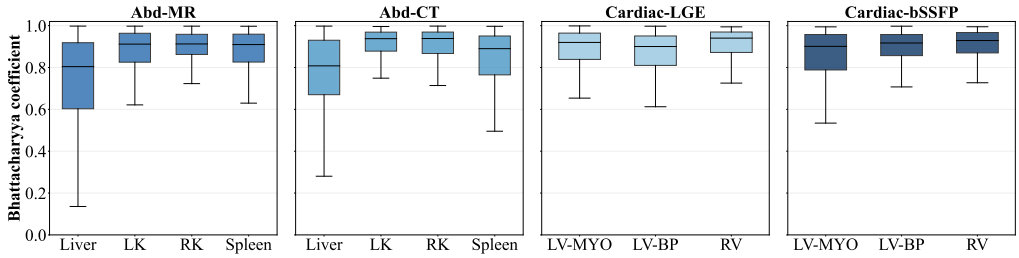


Figure 3: Bhattacharyya coefficient distributions between support and query foregrounds in target domains. Higher values indicate stronger foreground distribution consistency.

Table 3: Components ablation of DA-Proto under Abd-CT → Abd-MR setting. The first row denotes the baseline, DSC denotes the mean Dice (%) and Δ denotes the improvement over the baseline (the first row).

FDM	DFR	DGPM	DSC ± Std	Δ
			55.28 ± 05.97	—
✓			68.08 ± 07.66	+12.80
✓	✓		70.19 ± 06.89	+14.91
✓	✓	✓	**74.64 ± 04.41**	**+19.36**
✓		✓	71.75 ± 05.48	+16.47
✓	✓	✓	**74.64 ± 04.41**	**+19.36**

Table 4: Ablation study of the loss terms in DA-Proto under the Abd-CT → Abd-MR setting. DSC denotes the mean Dice score (%). $\mathcal{L}_{ce}$ is the primary segmentation loss, while $\mathcal{L}_{align}$ and $\mathcal{L}_r$ denote the alignment and reconstruction losses in the DFR module, respectively.

$\mathcal{L}_{ce}$	$\mathcal{L}_{align}$	$\mathcal{L}_r$	DSC
✓			71.75 ± 05.48
✓	✓		71.14 ± 05.03
✓		✓	71.10 ± 05.21
✓	✓	✓	**74.64 ± 04.41**

The results are summarized in Tab. 1 and Tab. 2. For each class, we report the mean Dice score and standard deviation (Std) across five folds. The *Mean* column in Tab. 1 and Tab. 2 reports the class-averaged Dice and its Std across the five folds, where a lower Std indicates more stable performance across different folds.

As shown in Tab. 1 and Tab. 2, our DA-Proto achieves the best mean Dice scores across all four cross-domain tasks. On the cross-modality tasks, DA-Proto obtains mean Dice scores of 74.64% and 74.61% on Abd-CT → MR and Abd-MR → CT, outperforming the second-best method by 5.68 and 7.14 percentage points, respectively. On the cross-sequence tasks, DA-Proto achieves 81.68% and 65.21% on Cardiac LGE → bSSFP and Cardiac bSSFP → LGE, with improvements of 9.95 and 3.34 percentage points over the corresponding second-best results. Moreover, our DA-Proto also generally maintains relatively low standard deviations across five folds, demonstrating robust and stable performance under diverse domain shifts.

4.4 Ablation

Unless otherwise specified, experiments are conducted on the Abd-CT → Abd-MR task by default.

Validation of Foreground Distribution Consistency. To quantitatively validate support-query foreground distribution consistency, we compute the Bhattacharyya coefficient (BC) between their foreground intensity distributions, where a higher BC indicates greater similarity. As shown in Fig. 3, high BC values are consistently observed across Abd-MR, Abd-CT, Cardiac-LGE, and Cardiac-bSSFP, with most categories showing median values around 0.9 or above. These results support that the support-query foreground distributional relationship remains relatively stable across different domains, motivating its use as a category-specific prior for prototype modulation.

Component Ablation. To evaluate the effectiveness of each component, we progressively introduce FDM, DFR, and DGPM into baseline. As shown in Tab. 3, adding FDM improves the mean DSC from 55.28% to 68.08%, yielding a gain of 12.80 percentage points and demonstrating the benefit of explicitly modeling foreground distribution priors. Building on FDM, introducing DFR further improves the performance to 70.19%, indicating that reconstruction-based regularization helps preserve meaningful foreground distribution information in the learned representation.

On the other hand, adding DGPM on top of FDM increases the mean DSC to 71.75%, demonstrating the effectiveness of distribution-guided prototype modulation. Combining all three modules achieves the best performance of 74.64%, outperforming the baseline by 19.36 percentage points.

Table 5: Ablation study on category-specific distribution prior injection under the Cross-Modality task (Abd-CT → Abd-MR). Each row denotes the foreground distribution prior used in the DGPM module to modulate distribution-aligned foreground prototypes. "Matched" uses the distribution prior corresponding to the current target class in support samples.

Distribution Prior	Liver	LK	RK	Spleen	Mean
None	70.12 ± 03.29	40.95 ± 03.71	44.70 ± 03.79	50.90 ± 08.89	51.67 ± 03.52
Random	28.58 ± 13.50	59.32 ± 11.74	66.71 ± 13.77	53.61 ± 15.16	52.05 ± 08.74
Liver Prior	**74.56 ± 03.09**	67.57 ± 04.84	81.03 ± 07.41	62.63 ± 12.46	71.45 ± 06.40
LK Prior	59.85 ± 08.42	**73.26 ± 04.30**	84.89 ± 05.79	61.89 ± 12.56	69.97 ± 05.57
RK Prior	58.48 ± 09.62	72.96 ± 05.43	**85.75 ± 04.84**	63.53 ± 11.90	70.18 ± 05.68
Spleen Prior	62.05 ± 08.72	71.53 ± 05.81	83.14 ± 03.65	**64.98 ± 11.33**	70.42 ± 05.52
Matched Prior (Ours)	**74.56 ± 03.09**	**73.26 ± 04.30**	**85.75 ± 04.84**	**64.98 ± 11.33**	**74.64 ± 04.41**

Table 6: Efficiency comparison on the Abd-CT → Abd-MR task. We report the mean Dice score (%) with Std, the number of parameters, FLOPs, and average slice processing time (ASPT) for segmentation performance, model complexity, computational cost, and inference efficiency, respectively.

Method	mDice ↑ ± Std ↓	Params (M) ↓	FLOPs (G) ↓	ASPT (ms/slice) ↓
TPM	63.26 ± 05.39	40.22	**102.3**	06.70 ± 00.61
FAMNet	67.10 ± 06.29	29.68	226.2	19.47 ± 00.15
DSM	64.01 ± 07.33	45.54	354.4	67.96 ± 01.25
C-Graph	68.96 ± 07.11	56.73	754.8	111.6 ± 02.28
DA-Proto (Ours)	**74.64 ± 04.41**	**20.41**	149.5	12.01 ± 00.36

Overall, FDM extracts the foreground distribution prior, DFR regularizes its representation, and DGPM exploits it for prototype modulation, demonstrating their complementary effects.

Loss Ablation. To investigate the contribution of each loss term, we use the segmentation loss $\mathcal{L}_{ce}$ as the baseline and separately introduce the alignment loss $\mathcal{L}_{align}$ and reconstruction loss $\mathcal{L}_r$. As shown in Tab. 4, using either $\mathcal{L}_{align}$ or $\mathcal{L}_r$ alone yields performance comparable to the baseline, with mean DSC of 71.14% and 71.10%, respectively. In contrast, combining both losses improves the mean DSC from 71.75% to 74.64% while reducing the Std from 5.48 to 4.41, demonstrating their complementary effects. Specifically, $\mathcal{L}_r$ encourages the distribution representation to preserve foreground statistics, while $\mathcal{L}_{align}$ further aligns the learned prior with the current feature space. Their joint supervision therefore leads to a more effective and stable distribution representation.

Effect of Foreground Distribution Consistency. To verify the role of foreground distribution consistency , we vary the source of the distribution prior used for prototype modulation, including **no prior**, a **random prior**, four **fixed category priors**, and a **matched prior** from the corresponding support category. As shown in Tab. 5, the random prior slightly improves the mean Dice but causes unstable changes, including a clear performance drop and increased Std on Liver. Each fixed category prior benefits its corresponding target class most, suggesting that the learned prior contains category-specific information. Notably, the matched setting achieves the highest mean Dice of 74.64%, clearly outperforming the no-prior, random-prior, and fixed-prior variants. These results show that DA-Proto benefits from category-consistent foreground distribution information rather than arbitrary feature injection, supporting its use for reliable prototype modulation.

Efficiency Comparison. To evaluate efficiency, we compare DA-Proto with TPM (Kim et al., 2025), FAMNet (Bo et al., 2025b), DSM (Zhu et al., 2025b), and C-Graph (Bo et al., 2025a) in terms of mDice, parameters, FLOPs, and ASPT on an NVIDIA RTX 3090 (24GB). In Tab. 6, DA-Proto achieves the highest mDice of 74.64% with only 20.41M parameters and 12.01 ms per slice. Compared with C-Graph, it reduces parameters and FLOPs by 64.0% and 80.2%, respectively, while improving mDice by 5.68 percentage points. Overall, DA-Proto demonstrates superior performance.

5 Conclusion

In this paper, we proposed DA-Proto, a novel CD-FSMIS framework that exploits stable support-query foreground distribution consistency as a category-specific prior. By integrating FDM, DFR, and DGPM, DA-Proto extracts category-specific foreground distribution priors, regularizes the learned distribution representations through reconstruction and alignment, and uses them to modulate domain-shifted prototypes, leading to improved cross-domain performance and robustness.

6 AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including improving grammar, clarity, fluency, and academic style. We did not use generative AI tools to generate research ideas, experimental results, data, figures, or scientific claims. All AI-assisted text was carefully reviewed and revised by the authors to ensure technical accuracy and consistency with the original content. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yuntian Bo, Tao Zhou, Zechao Li, Haofeng Zhang, and Ling Shao. Contrastive graph modeling for cross-domain few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 45: 2173–2186, 2025a.

Yuntian Bo, Yazhou Zhu, Lunbo Li, and Haofeng Zhang. Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In *AAAI*, volume 39, pp. 1889–1897, 2025b.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pp. 205–218. Springer, 2022.

Ziming Cheng, Shidong Wang, Tong Xin, Tao Zhou, Haofeng Zhang, and Ling Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 43(6):2202–2214, 2024.

Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, Geoff Pleiss, and John Patrick Cunningham. Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. 2020.

Jianchao Jiang and Haofeng Zhang. Concentrate on weakness: Mining hard prototypes for few-shot medical image segmentation. In *IJCAI*, pp. 1242–1250, 2025.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.

Hyeongji Kim, Stine Hansen, and Michael Kampffmeyer. Tied prototype model for few-shot medical image segmentation. In *MICCAI*, pp. 651–661, 2025.

Ekaterina Kondrateva, Marina Pominova, Elena Popova, Maxim Sharaev, Alexander Bernstein, and Evgeny Burnaev. Domain shift in computer vision models for mri data analysis: an overview. In *Thirteenth International Conference on Machine Vision*, volume 11605, pp. 126–133. SPIE, 2021.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Workshop Challenge*, volume 5, pp. 12, 2015.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *ECCV*, pp. 73–90, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014b.

Yi Lin, Yufan Chen, Kwang-Ting Cheng, and Hao Chen. Few shot medical image segmentation with cross attention transformer. In *MICCAI*, pp. 233–243. Springer, 2023.

Jie Liu, Jiayi Shen, Pan Zhou, Jan-Jakob Sonke, and Efstratios Gavves. Probabilistic prototype calibration of vision-language models for generalized few-shot semantic segmentation. In *ICCV*, pp. 21155–21165. IEEE, 2025.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *CVPR*, pp. 3380–3390, 2024.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.

Agung W Setiawan. Image segmentation metrics in skin lesion: accuracy, sensitivity, specificity, dice coefficient, jaccard index, and matthews correlation coefficient. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pp. 97–102. IEEE, 2020.

Qianqian Shen, Yanan Li, Jiyong Jin, and Bin Liu. Q-net: Query-informed few-shot medical image segmentation. In *IntelliSys*, pp. 610–628, 2023.

Liyan Sun, Chenxin Li, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, Yizhou Yu, and John Paisley. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140:105067, 2022.

Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *ICCV*, pp. 3898–3908. IEEE, 2021.

Lazaros Tsochatzidis, Panagiota Koutla, Lena Costaridou, and Ioannis Pratikakis. Integrating segmentation information into cnn for breast cancer diagnosis of mammographic masses. *Computer Methods and Programs in Biomedicine*, 200:105913, 2021.

Haochen Wang, Yandan Yang, Xianbin Cao, Xiantong Zhen, Cees Snoek, and Ling Shao. Variational prototype inference for few-shot semantic segmentation. In *WACV*, pp. 525–534, 2021.

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

Bin Xu, Yazhou Zhu, Shidong Wang, Yang Long, and Haofeng Zhang. Few-shot medical image segmentation via boundary-extended prototypes and momentum inference. *Computer Vision and Image Understanding*, 263:104571, 2026.

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171: 108238, 2024.

Xueying Zhou, Ge Jin, Yang Liu, Juncheng Li, Jun Wang, Shihui Ying, Yanyan Zheng, and Jun Shi. Multi-resolution based dual-channel unet with cross clique for medical image dense prediction. *Expert Systems with Applications*, 276:127190, 2025.

Yazhou Zhu and Haofeng Zhang. Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In *MICCAI*, pp. 326–336. Springer, 2025.

Yazhou Zhu, Shidong Wang, Tong Xin, and Haofeng Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *MICCAI*, pp. 271–280, 2023.

Yazhou Zhu, Minxian Li, Qiaolin Ye, Shidong Wang, Tong Xin, and Haofeng Zhang. Robustemd: Domain robust matching for cross-domain few-shot medical image segmentation. *Artificial Intelligence in Medicine*, pp. 103197, 2025a.

Yazhou Zhu, Shidong Wang, Tao Zhou, Zechao Li, Haofeng Zhang, and Ling Shao. Cross-domain few-shot medical image segmentation via dynamic semantic matching. *IEEE Transactions on Image Processing*, 34:6669–6682, 2025b.

Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, Cheng Ouyang, et al. Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis*, 81:102528, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 EXPERIMENTAL SETUP

All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24GB memory. DA-Proto adopts a COCO-pretrained ResNet-50 (Lin et al., 2014b;a) as the backbone, with the feature dimension set to $C = 512$. The model is trained for 10 epochs with 5,000 episodes per epoch using SGD with an initial learning rate of 0.001 and momentum of 0.9. The learning rate is decayed by a factor of 0.9 every 1,000 iterations.

For the proposed modules, we set the number of local foreground regions to $m = 16$. The weights of the alignment and reconstruction losses are set to $\lambda_{align} = 0.1$ and $\lambda_r = 1.0$, respectively. The numbers of foreground and background prototypes are set to $N_{fg} = 200$ and $N_{bg} = 400$, while the corresponding Top-k values are $k_{fg} = 10$ and $k_{bg} = 20$.

To ensure a fair comparison, all competing methods are re-trained and evaluated by us under the same data splits, training settings, and evaluation protocol as DA-Proto. We do not directly use the results reported in the original papers. Instead, all methods are reproduced and evaluated under the same 5-fold cross-validation setting and inference protocol, and the corresponding results are recorded from our unified experimental environment.

A.2 TRAINING STRATEGY

Following SSL-ALP (Ouyang et al., 2022) and the training strategy adopted in FAMNet (Bo et al., 2025b), we construct training episodes using pseudo-labels generated by 3D superpixel clustering, without using ground-truth segmentation masks during training.

Specifically, for each training episode, one 2D slice is selected from a 3D volume as the support image, and the corresponding superpixel pseudo-label is used as the support mask. Nearby slices are then sampled to construct the query set. To increase the diversity of episodic training and simulate appearance variations between support and query samples, mild perturbations are applied to the query images, including translation, rotation, and pixel-intensity perturbation. The resulting support-query pairs are then used for episodic optimization.

A.3 INFERENCE PROTOCOL

At inference time, DA-Proto performs segmentation in a 2D slice-wise manner and is directly evaluated on the unseen target domain without additional fine-tuning. We adopt 5-fold cross-validation for evaluation. For each fold, the third patient is selected as the support patient, while the remaining patients are used as query patients.

Under the 1-way 1-shot setting, the middle slice corresponding to the target category is selected from the support patient as the single annotated support image. This support image and its corresponding foreground mask are then used to segment the target category in all query slices from the remaining patients. Notably, both the support and query samples are drawn from the same unseen target domain during inference.

A.4 EVALUATION METRIC

We use the Dice Similarity Coefficient (Setiawan, 2020) (DSC) as the primary evaluation metric. Given a predicted foreground mask $\hat{M}$ and the corresponding ground-truth mask M, DSC is defined as

$$\mathrm{DSC}(\hat{M}, M) = \frac{2|\hat{M} \cap M|}{|\hat{M}| + |M|}, \tag{17}$$

Table 7: Effect of the number of foreground and background prototypes N_{fg}, N_{bg} on the mean Dice Score (%). The best and second-best results are highlighted in **bold** and underlined, respectively.

N_{fg}	N_{bg}	Abd-CT → Abd-MR				
		Liver	LK	RK	Spleen	Mean
50	100	71.43 ± 01.93	73.45 ± 05.23	79.61 ± 08.30	61.50 ± 13.42	71.50 ± 06.16
100	200	74.32 ± 02.39	73.13 ± 04.25	83.26 ± 06.88	63.80 ± 11.77	73.63 ± 05.30
150	300	**74.57 ± 02.85**	**74.33 ± 03.74**	85.05 ± 05.15	**65.46 ± 11.57**	**74.85 ± 04.66**
200	400	74.56 ± 03.09	73.26 ± 04.30	85.75 ± 04.84	64.98 ± 11.33	74.64 ± 04.41
250	500	74.39 ± 03.24	73.44 ± 05.22	**85.87 ± 04.48**	64.54 ± 12.35	74.56 ± 04.52
300	600	74.22 ± 03.29	73.50 ± 05.76	85.65 ± 04.30	64.27 ± 12.55	74.41 ± 04.56

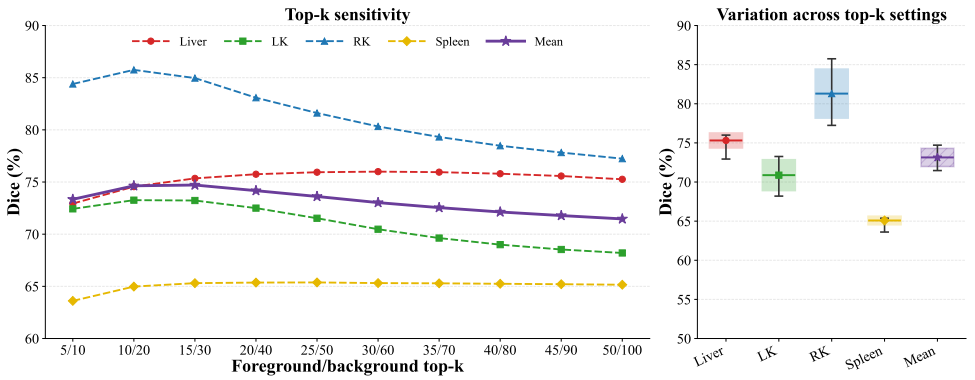


Figure 4: Sensitivity analysis of DA-Proto to different foreground/background Top-k settings on the Abd-CT → Abd-MR task. Left: class-wise and mean Dice scores under different Top-k configurations, where the foreground/background ratio is fixed at $1:2$. Right: distribution of Dice scores across all evaluated Top-k settings for Liver, LK, RK, Spleen, and the overall mean, illustrating the performance variation of each category.

where $|\cdot|$ denotes the number of foreground pixels. DSC ranges from 0 to 1, with a higher value indicating a larger overlap between the predicted segmentation and the ground truth. In our experiments, DSC is reported in percentage form. We further report the mean and standard deviation (Std) across five folds to evaluate both segmentation performance and stability.

B Additional Ablation

B.1 Sensitivity to the Number of Prototypes

To investigate the sensitivity of DA-Proto to the number of prototypes, we vary the numbers of foreground and background prototypes while keeping their ratio fixed at $1:2$. Specifically, we evaluate $(N_{fg}, N_{bg}) \in \{(50,100),(100,200),(150,300),(200,400),(250,500),(300,600)\}$, while keeping all other settings unchanged. The experiments are conducted on the Abd-CT → Abd-MR task.

As shown in Tab. 7, the performance initially improves as the number of prototypes increases, since a larger prototype set can capture more diverse local foreground and background patterns. However, excessively increasing the number of prototypes provides limited additional benefit and may introduce redundant representations. Overall, DA-Proto maintains stable performance across a wide range of prototype numbers, indicating that DA-Proto is not highly sensitive to this hyperparameter.

B.2 Sensitivity to the Top Number of Prototypes

To evaluate the sensitivity of DA-Proto to the Top-k prototype selection, we vary the foreground/background Top-k values while keeping their ratio fixed at $1:2$ and all other settings unchanged. As shown in Fig. 4, different Top-k settings lead to some class-wise performance variations, particularly for RK. However, the overall mean Dice changes only moderately across a wide range of settings and consistently remains above 70%. These results indicate that DA-Proto is rel-

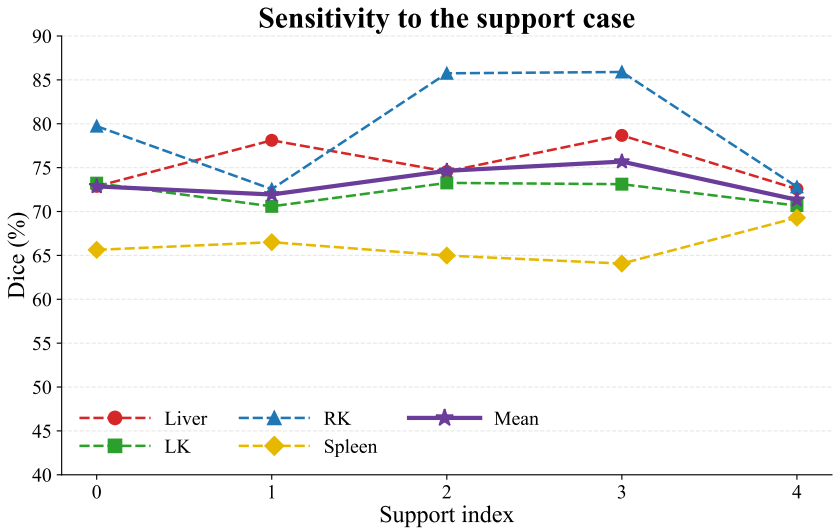


Figure 5: Sensitivity analysis of DA-Proto to the selected support sample on the Abd-CT → Abd-MR task. We vary the support index while keeping all other settings unchanged, and report the Dice scores for Liver, LK, RK, Spleen, and their mean Dice score.

Table 8: Comparison under 1-shot and 5-shot settings on the Cross-Modality task (Abd-CT → Abd-MR). Results are reported in Dice score (%). The best and second-best results are highlighted in **bold** and underlined, respectively.

	1-shot					5-shot				
Method	Liver	LK	RK	Spleen	Mean	Liver	LK	RK	Spleen	Mean
TPM	65.58	61.46	68.99	57.03	63.26	67.70	61.78	70.80	57.93	64.55
FAMNet	**75.23**	62.64	68.19	62.32	67.10	75.56	66.73	72.56	64.59	69.86
C-Graph	68.73	69.99	74.66	62.46	68.96	71.09	70.80	75.92	63.42	70.31
DA-Proto	74.56	**73.26**	**85.75**	**64.98**	**74.64**	**78.77**	**74.02**	**86.43**	**65.90**	**76.28**

atively robust to the choice of Top-k and does not rely heavily on a specific prototype selection configuration.

B.3 Sensitivity to the Selected Support Sample

To evaluate the robustness of DA-Proto to support sample selection, we vary the support index while keeping all other settings unchanged. As shown in Fig. 5, different support samples lead to some class-wise performance variations. However, the mean Dice remains relatively stable across different support indices and consistently stays above 70%. This indicates that DA-Proto is not highly sensitive to the selection of a specific support sample and maintains robust segmentation performance under different support choices.

C Additional Experiments

C.1 5-Shot Performance

To further evaluate the effectiveness of DA-Proto when more annotated support samples are available, we extend the standard 1-shot setting to a 5-shot setting on the Abd-CT → Abd-MR task. As shown in Tab. 8, all methods benefit from additional support samples to varying degrees. DA-Proto consistently achieves the best performance under both settings, with its mean Dice increasing from 74.64% in the 1-shot setting to 76.28% in the 5-shot setting. Improvements are observed across all four target organs, including Liver, LK, RK, and Spleen. Under the 5-shot setting, DA-Proto further outperforms the second-best C-Graph by 5.97 percentage points in mean Dice.

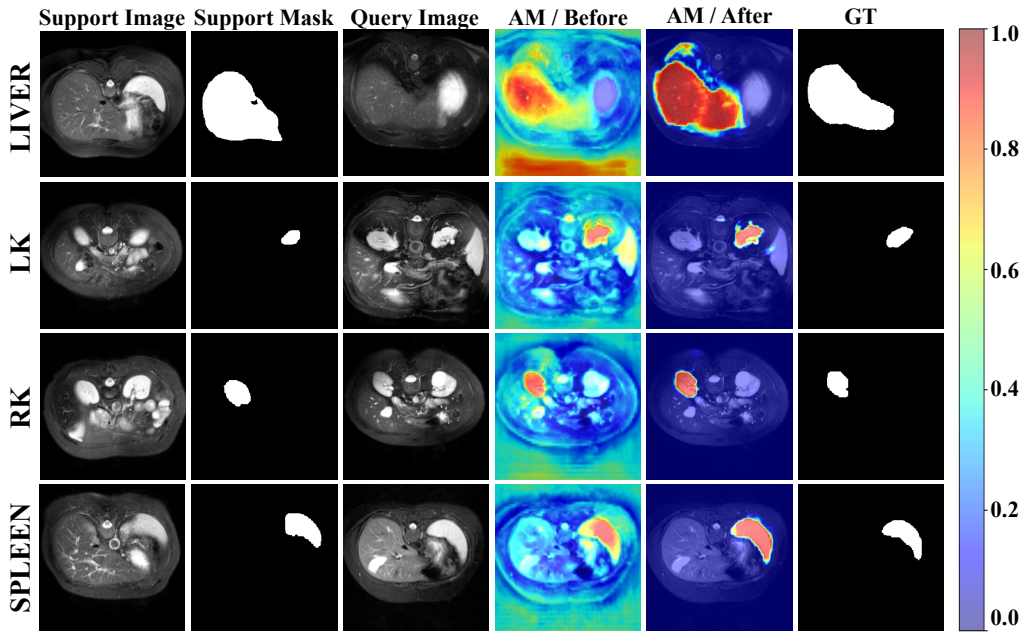


Figure 6: Activation Maps (AMs) before and after distribution prior injection. The values in the AM are normalized to a range of 0-1. From the left to right: support image, support mask, query image, AMs before and distribution prior injection, and the ground truth of query image.

These results indicate that DA-Proto can effectively exploit additional target-domain support information. Multiple annotated support samples provide richer category-specific foreground statistics, enabling a more reliable distribution prior for prototype modulation. This further demonstrates that DA-Proto generalizes well beyond the 1-shot setting and remains effective when more support annotations are available.

D VISUALIZATIONS

D.1 ACTIVATION MAPS

To intuitively demonstrate the effect of foreground distribution prior injection, we visualize the activation maps before and after distribution prior modulation for four abdominal organs. As shown in Fig. 6, the responses before distribution prior injection are relatively scattered with noticeable background activations. In contrast, after injecting the distribution prior, the responses become more concentrated on the target organs with reduced irrelevant activations. These results show that the foreground distribution prior suppresses ambiguous responses in irrelevant regions and strengthens the model's focus on the target foreground, providing qualitative evidence for the effectiveness of prior-guided prototype modulation.

D.2 VISUAL SEGMENTATION RESULTS

To provide an intuitive qualitative comparison of segmentation performance, we further visualize the predictions of DA-Proto and several representative competing methods, including TPM (Kim et al., 2025), FAMNet (Bo et al., 2025b), DSM (Zhu et al., 2025b), and C-Graph (Bo et al., 2025a). As shown in Figs. 9–12, from left to right, we present the support image, segmentation results of TPM, FAMNet, DSM, C-Graph, and our proposed DA-Proto, followed by the ground-truth (GT) mask. Across different target categories and domain shifts, DA-Proto generally achieves better segmentation results, with more complete foreground regions and fewer erroneous predictions in surrounding areas. These qualitative results are consistent with the quantitative comparisons and further demonstrate the effectiveness of DA-Proto under cross-domain settings.

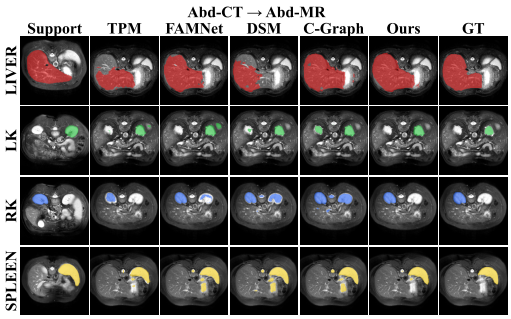


Table 9: Qualitative comparison between our DA-Proto and other methods on the Cross-Modality dataset (Abdomen CT → MR). From top to bottom, the rows correspond to Liver, Left Kidney (LK), Right Kidney (RK) and Spleen.

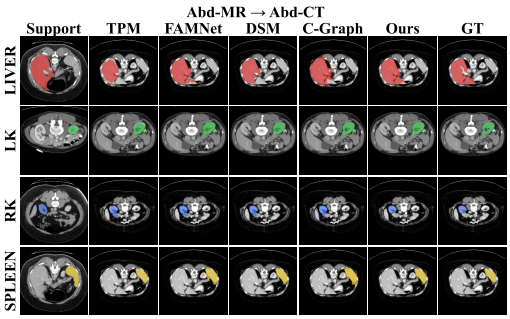


Table 10: Qualitative comparison between our DA-Proto and other methods on the Cross-Modality dataset (Abdomen MR → CT). From top to bottom, the rows correspond to Liver, Left Kidney (LK), Right Kidney (RK) and Spleen.

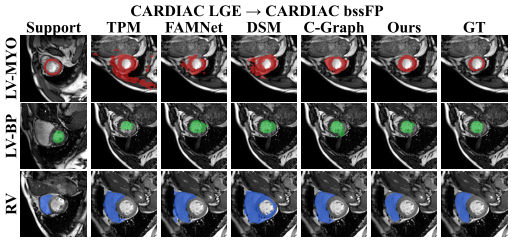


Table 11: Qualitative comparison between our DA-Proto and other methods on the Cross-Sequence dataset (Cardiac LGE → Cardiac bssFP). From top to bottom, the rows correspond to the Left Ventricle Myocardium (LV-MYO), Left Ventricle Blood Pool (LV-BP) and Right Ventricle (RV).

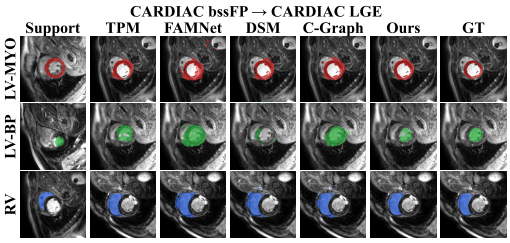


Table 12: Qualitative comparison between our DA-Proto and other methods on the Cross-Sequence dataset (Cardiac bssFP → Cardiac LGE). From top to bottom, the rows correspond to the Left Ventricle Myocardium (LV-MYO), Left Ventricle Blood Pool (LV-BP) and Right Ventricle (RV).