FAMNet: Frequency-aware Matching Network for Cross-domain Few-shot Medical Image Segmentation

Yuntian Bo, Yazhou Zhu, Lunbo Li, Haofeng Zhang[*]

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
{yuntian.bo, zyz_nj, lunboli, zhanghf}@njust.edu.cn

Abstract

Existing few-shot medical image segmentation (FSMIS) models fail to address a practical issue in medical imaging: the domain shift caused by different imaging techniques, which limits the applicability to current FSMIS tasks. To overcome this limitation, we focus on the cross-domain few-shot medical image segmentation (CD-FSMIS) task, aiming to develop a generalized model capable of adapting to a broader range of medical image segmentation scenarios with limited labeled data from the novel target domain. Inspired by the characteristics of frequency domain similarity across different domains, we propose a Frequency-aware Matching Network (FAMNet), which includes two key components: a Frequency-aware Matching (FAM) module and a Multi-Spectral Fusion (MSF) module. The FAM module tackles two problems during the meta-learning phase: 1) intra-domain variance caused by the inherent support-query bias, due to the different appearances of organs and lesions, and 2) inter-domain variance caused by different medical imaging techniques. Additionally, we design an MSF module to integrate the different frequency features decoupled by the FAM module, and further mitigate the impact of inter-domain variance on the model's segmentation performance. Combining these two modules, our FAMNet surpasses existing FSMIS models and Cross-domain Few-shot Semantic Segmentation models on three cross-domain datasets, achieving state-of-the-art performance in the CD-FSMIS task. Code is available at https://github.com/primebo1/FAMNet.

Introduction

To bridge the gap between limited labeled samples and the need for precise segmentation, few-shot medical image segmentation (FSMIS) (Guha Roy et al. 2019; Ouyang et al. 2022; Hansen et al. 2022; Zhu et al. 2023; Sun et al. 2022; Feng et al. 2021; Shen et al. 2023; Lin et al. 2023; Ding et al. 2023; Cheng et al. 2024) has emerged. By training on base categories, FSMIS models can leverage only a few annotated samples to segment new categories in medical images directly. Nevertheless, due to the limited generalization capability, they often exhibit diminished performance when tested on the data with domain shifts, which restricts their applicability to only a single domain.
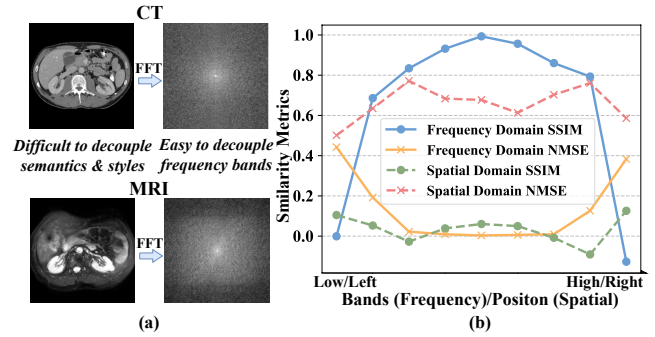
---

[*]Corresponding author.

Figure 1: Motivation of the proposed method. (a) CT and MRI scans in the spatial and frequency domains. Frequency spectra are processed using a Hamming window (Hamming 1977) and are center-shifted. (b) Quantitative metrics for the similarity of CT and MRI in the spatial and frequency domains using structural similarity index measure (SSIM) (Wang et al. 2004) and normalized mean square error (NMSE). Metrics are calculated using registered images.

Recently, some researchers have started investigating cross-domain few-shot semantic segmentation (CD-FSS) (Lei et al. 2022; Chen et al. 2024; Herzog 2024; Su et al. 2024; Nie et al. 2024; He et al. 2024), which has demonstrated impressive segmentation capabilities on datasets like Deepglobe (Demir et al. 2018) and FSS-1000 (Wei et al. 2019). Although this operation paves the way for cross-domain applications in few-shot scenarios, these models cannot be directly applied to the medical field due to the unique characteristics of medical images, *e.g.*, grayscale, intensity variations, and foreground-background imbalance. Meanwhile, existing domain generalization methods in medical imaging (Ouyang et al. 2021; Zhou et al. 2022; Xu et al. 2022; Su et al. 2023) mainly focus on domain randomization, neglecting the model itself and the few-shot setting.

Two major challenges hinder the development of cross-domain few-shot medical image segmentation (CD-FSMIS), which we try to address in this paper: 1) Intra-domain variations: Medical images exhibit significant variability between individual organs, *e.g.*, size, fat content, and pathology, making it difficult to find similar support-query pairs, leading to support-query bias and reduced prototype representation in

prototypical networks. 2) Inter-domain variations:

Even within the same organ or region, the spatial domain similarity demonstrates low correlation across different domains, as illustrated in Figure 1(b). However, subtle distinctions are evident in the frequency domain, where inter-domain variations are primarily in high and low-frequency bands, while mid-frequency bands are relatively similar.

We therefore propose a novel method termed Frequency-aware Matching Network (FAMNet) for CD-FSMIS in this paper. Specifically, the core of our FAMNet, the Frequency-aware Matching (FAM) module, performs support-query matching in specific frequency bands, eliminating support-query bias by fusing foreground features and highlighting synergistic parts. Simultaneously, FAM incorporates frequency domain information within the feature space, reducing reliance on frequency bands with significant domain differences. This allows the model to focus more on resilient, domain-agnostic frequency bands, effectively addressing both intra-domain and inter-domain variations. Building upon the FAM module, we subsequently developed a Multi-spectral Fusion (MSF) module. While fusing the frequency-decoupled features from FAM, the MSF module extracts the critical information that remains in the domain-specific frequency bands after decoupling in the spatial domain. With FAM and MSF, our method not only demonstrates strong generalization capabilities but also effectively leverages domain-invariant interactive information from the sample space, showcasing excellent segmentation performance. In summary, our contributions are as follows:

- We extend few-shot medical image segmentation to a new task, termed cross-domain few-shot medical image segmentation, aimed at training a generalizable model to segment a novel class in unseen target domains with only a few annotated examples.

- We propose a novel FAM module that concurrently mitigates the adverse impacts of intra-domain and inter-domain variances on model performance. Moreover, an MSF module is introduced for multi-spectral feature fusion, further suppressing domain-variant information to enhance the model's generalizability.

- On three cross-domain datasets, our proposed method archives the state-of-the-art performance. The effectiveness and superiority of our method are further verified through various ablation studies and visualization.

Related Works

Few-shot Medical Image Segmentation

The FSMIS task has been proposed to address data scarcity typically found in medical scenarios, which aims to train models capable of segmenting novel organs or lesions with only a few annotated samples. Current FSMIS models can be categorized into two approaches: interactive networks (Guha Roy et al. 2019; Sun et al. 2022; Feng et al. 2021; Ding et al. 2023) and prototypical networks (Ouyang et al. 2022; Hansen et al. 2022; Shen et al. 2023; Zhu et al. 2023; Lin et al. 2023; Cheng et al. 2024). In the former category, SENet (Guha Roy et al. 2019) pioneered the use

of interactive networks in FSMIS tasks, followed by MR-rNet (Feng et al. 2021), GCN-DE (Sun et al. 2022), and CRAPNet (Ding et al. 2023). The core idea behind these models is to enhance support-query interaction through attention mechanisms. For the latter category, SSL-ALPNet (Ouyang et al. 2022) introduced a self-supervised framework that generates adaptive local prototypes and supervised by superpixel-based pseudo-labels during training. ADNet (Hansen et al. 2022) proposed a learnable threshold for segmentation and relied on a single foreground prototype to compute anomaly scores for all query pixels, rather than learning prototypes for each class. CATNet (Lin et al. 2023) utilized a cross-masked attention Transformer to enhance support-query interaction and improve feature representation. GMRD (Cheng et al. 2024) captured the complexity of prototype class distributions by generating multiple representative descriptors. Unfortunately, all existing FSMIS methods are limited to single-domain applications, neglecting the domain shifts encountered in medical imaging.

Cross-domain Few-shot Semantic Segmentation

Expanding on few-shot semantic segmentation (FSS), recent studies (Herzog 2024; He et al. 2024; Su et al. 2024; Nie et al. 2024; Chen et al. 2024; Lei et al. 2022) focus on CD-FSS, considering a more practical setting where both label space and data distribution are disjoint between the training and testing datasets. PATNet (Lei et al. 2022) employs a Pyramid-Anchor-Transformation module (PATM) to map domain-specific features into domain-agnostic ones. PMNet (Chen et al. 2024) proposes a lightweight matching network to densely exploit pixel-to-pixel and pixel-to-patch correlations between support-query pairs. DRAdapter (Su et al. 2024) utilizes local-global style perturbation to train an adapter that rectifies diverse target domain styles to the source domain, maximizing the utilization of the well-optimized source domain segmentation model. Nevertheless, existing CD-FSS models often suffer from substantial performance degradation when applied to medical images due to significant differences from natural images, such as color, intensity, and foreground-background imbalance.

Methodology

Problem Setting

The CD-FSMIS task aims to construct a generalizable model Θ to segment novel organs or lesions in an unseen domain with few annotated medical images. To elaborate, the model Θ is optimized using a single source domain dataset $\mathcal{D}^s$ encompassing the base categories $\mathcal{C}_{base}$. Subsequently, the model's performance is assessed on a target domain dataset $\mathcal{D}^t$, which comprises novel target categories $\mathcal{C}_{target}$ with only a few labeled images. It is crucial to highlight that the sets of categories $\mathcal{C}_{base}$ and $\mathcal{C}_{target}$ are disjoint, *i.e.* $\mathcal{C}_{base} \cap \mathcal{C}_{target} = \emptyset$, and a domain shift exists between the source domain $\mathcal{D}^s$ and the target domain $\mathcal{D}^t$. During the training phase, the model has no access to the target domain.

Our approach adheres to the episode-based meta-learning paradigm. For each meta-learning task, we randomly divide
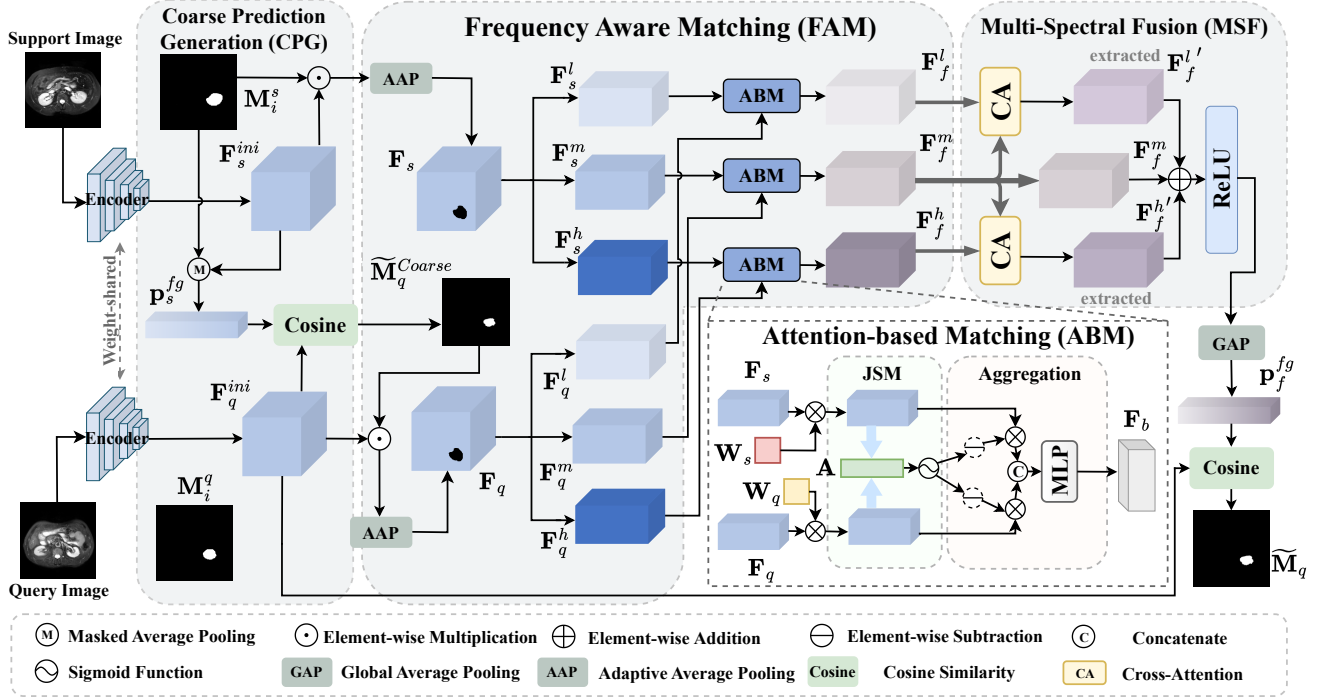
Figure 2: The overall architecture of our method, consists of three main technical components: the Coarse Prediction Generation (CPG) module, the Frequency-aware Matching (FAM) module, and the Multi-Spectral Fusion (MSF) module. Note that in ABM, JSM denotes joint space matching. In the case of DAFBs, the attention matrix is directly utilized for attention weighting. Conversely, for DSFBs, an element-wise subtraction is applied prior to the weighting process.

the data into multiple episodes. Each episode $(\mathcal{S}, \mathcal{Q})$ comprises: 1) a support set $\mathcal{S} = \{x_i^s, \mathbf{M}_i^s\}_{i=1}^{K}$ containing K support samples, and 2) a query set $\mathcal{Q} = \{x_i^q, \mathbf{M}_i^q\}_{i=1}^{N_q}$ containing N_q query samples, where x_i denotes the i-th image, and $\mathbf{M}_i$ denotes the corresponding segmentation ground truth. During inference, the model's segmentation performance is evaluated by providing a support set and a query set from the novel target domain.

Overall Architecture

The proposed network is depicted in Figure 2, which can be briefly divided into three main parts: 1) a Coarse Prediction Generation (CPG) module for generating a coarse prediction of the query mask, 2) a Frequency-aware Matching (FAM) module for performing frequency-aware matching between support and query foreground features, and 3) a Multi-Spectral Fusion (MSF) module for fusing features based on their respective frequency bands.

First, the support and query images are fed into a weight-sharing feature encoder to extract their corresponding feature maps. Next, a coarse prediction of the query foreground mask is obtained using the CPG module. Then, the support foreground feature computed in CPG, the generated coarse query mask, and the extracted query feature are input into the proposed FAM module for frequency-aware matching. In this module, features are divided into three frequency bands, each independently fuses the support and query features through distinct weighting mechanisms

guided by matching results, yielding fused features for each band. To reintegrate these multi-spectral features and further suppress the influence of domain-specific frequency bands (DSFBs), the divided features are fused through the MSF module. A global average pooling operation is then performed to obtain the foreground prototype required by the prototypical network. Finally, we compute the cosine similarity between the foreground prototype and the query feature to produce the final prediction of the query mask.

Feature Extraction

We use a ResNet-50 (He et al. 2016) feature encoder $\mathcal{E}_\theta(\cdot)$ pre-trained on MS-COCO (Lin et al. 2014) as a weight-shared backbone to extract the support and query feature maps, where θ denotes the backbone parameters.

The support and query feature maps are denoted as $\mathbf{F}_s^{ini} = \mathcal{E}_\theta(x_i^s)$ and $\mathbf{F}_q^{ini} = \mathcal{E}_\theta(x_i^q)$, and $\mathbf{F}_s^{ini}, \mathbf{F}_q^{ini} \in \mathbb{R}^{C \times H \times W}$, where C denotes the channel depth of the feature, and H and W denote the height and width of the feature respectively.

Coarse Prediction Generation (CPG) Module

A prototypical network-based method is performed to obtain a coarse segmentation mask of the query image. Given a support image x_i^s and its corresponding foreground binary mask $\mathbf{M}^s$, the support foreground prototype $\mathbf{p}_s^{fg}$ can be generated by using Masked Average Pooling (MAP). Mathematically,

this process can be denoted as:

$$\mathbf{p}_s^{fg} = \frac{\sum_{u,v} \mathbf{F}_s^{ini}(u,v)\mathbf{M}^s(u,v)}{\sum_{u,v} \mathbf{M}^s(u,v)}, \qquad (1)$$

where (u,v) is the index of pixels on the feature map.

Following this, we directly use the computed $\mathbf{p}_s^{fg}$ and the extracted query feature map $\mathbf{F}_q^{ini}$ to predict a coarse query foreground mask $\widetilde{\mathbf{M}}_q^{Coarse}$:

$$\widetilde{\mathbf{M}}_q^{Coarse} = 1.0 - \sigma\left(d(\mathbf{F}_q^{ini}, \mathbf{p}_s^{fg}) - \tau\right), \qquad (2)$$

where $d(a,b) = -\alpha cos(a,b)$ is the negative cosine similarity with a fixed scaling factor $\alpha = 20$ (Wang et al. 2019), $\sigma(\cdot)$ denotes the Sigmoid activation and τ is a learnable threshold introduced by (Hansen et al. 2022).

Frequency-Aware Matching (FAM) Module

Multi-Spectral Decoupling of Foreground Features. To mitigate the discrepancy between support and query foregrounds, we first extract the foreground features from $\mathbf{f}_s$ and $\mathbf{f}_q$ using the support mask and coarse query mask. Since the number of foreground pixels in the support and query images often differs, we apply Adaptive Average Pooling (AAP) (Liu et al. 2018) to standardize the number of foreground pixels to a fixed value N. This process can be formulated as:

$$\begin{cases} \mathbf{F}_s = \text{AAP}(\mathbf{F}_s^{ini} \odot \mathbf{M}^s, N) \\ \mathbf{F}_q = \text{AAP}(\mathbf{F}_q^{ini} \odot \mathcal{R}(\widetilde{\mathbf{M}}_q^{Coarse}), N) \end{cases}, \qquad (3)$$

where $\odot$ denotes the Hadamard product, and $\mathbf{F}_s, \mathbf{F}_q \in \mathbb{R}^{C \times N}$ denote the extracted foreground features, and $\text{AAP}(a,n)$ is the AAP operation that adjusts the input feature map a to a fixed output size n along the last dimension, and $\mathcal{R}$ denotes the mathematical function that rounds decimals to 0 or 1.

For each foreground feature, we utilize the two-dimensional Fast Fourier Transform (FFT) to convert the signal from the spatial domain to the frequency domain while preserving spatial information. We employ a reshape function ρ to transform the feature into a square, $\rho : \mathbb{R}^{C \times N} \to \mathbb{R}^{C \times \sqrt{N} \times \sqrt{N}}$, with ρ^{-1} serving as its inverse. For the support foreground feature, this process is formalized as:

$$\phi_s = \text{SC}(\mathcal{F}(\rho(\mathbf{F}_s))), \qquad (4)$$

where $\mathcal{F}$ denotes FFT, $\phi_s \in \mathbb{C}^{C \times \sqrt{N} \times \sqrt{N}}$ denotes the frequency domain feature representation, SC denotes the function to adjusts the frequency signal to center the zero-frequency component.

Subsequently, we apply a band-pass filter to decompose the frequency-domain signal into three bands, namely high, medium, and low frequencies. Finally, we revert the frequency signals back to the spatial domain using the Inverse Fast Fourier Transform (IFFT):

$$\mathbf{F}_s(i) = \rho^{-1}(\mathcal{F}^{-1}(\mathcal{B}_p(\phi_s, I(i)))) = \begin{cases} \mathbf{F}_s^l, & i = 1 \\ \mathbf{F}_s^m, & i = 2 \\ \mathbf{F}_s^h, & i = 3 \end{cases}, \quad (5)$$

where $\mathbf{F}_s(i)$ denotes the support foreground feature in a specific frequency band, $\mathcal{F}^{-1}$ denotes IFFT, $I(i)$ is a binary mask vector where values are set to 1 for preserved components and 0 for discarded components, and $\mathcal{B}_p(\phi, I)$ is the band-pass filter that can be defined by:

$$\phi'(u,v) = \begin{cases} 0, & \text{if } I(u,v) = 0 \\ \phi(u,v), & \text{otherwise} \end{cases}. \qquad (6)$$

We perform the same operation on the query foreground feature to obtain $\mathbf{F}_q^l, \mathbf{F}_q^m, \mathbf{F}_q^h$.

Multi-Spectrum Attention-based Matching. As shown in Figure 1, high-frequency and low-frequency signals vary significantly across different domains. During typical training, models often rely on these DSFBs to better adapt to the tasks in the current scenario. However, this reliance leads to over-fitting to certain prominent features. For example, a model may achieve precise segmentation on CT images by focusing on the contained information of DSFBs, such as the high-frequency edge information. Yet, this approach often suffers from degradation when transferred to an unseen domain where edge information is less distinct, such as MRI. Our proposed method aims to enhance the model's generalization capabilities by matching support and query features while simultaneously reducing the model's reliance on DSFBs.

Specifically, this module consists of three processes, as illustrated in the top-right part of Figure 2. First, given the support and query foreground feature pair $(\mathbf{f}_s, \mathbf{f}_q)$ belonging to the same frequency band $\mathcal{B}$, we apply linear transformations using two learnable matrices $\mathbf{W}_s, \mathbf{W}_q \in \mathbb{R}^{N \times N}$ to map the features into a joint space. This further reduces the intra-domain differences between the support and query, enhancing the stability of matching. This operation can be formalized as:

$$\begin{cases} \mathbf{F}_s' = \mathbf{F}_s \mathbf{W}_s \\ \mathbf{F}_q' = \mathbf{F}_q \mathbf{W}_q \end{cases}, \qquad (7)$$

where $\mathbf{F}_s', \mathbf{F}_q' \in \mathbb{R}^{C \times N}$ denote the transformed support and query foreground features, respectively.

Secondly, based on the transformed features, we compute the attention-based similarity score matrix between the features using cosine similarity:

$$\mathbf{A}(\mathbf{F}_s', \mathbf{F}_q') = \sigma(\frac{\mathbf{F}_s' \cdot \mathbf{F}_q'}{\|\mathbf{F}_s'\|\|\mathbf{F}_q'\|}), \qquad (8)$$

where $\mathbf{A} \in \mathbb{R}^{1 \times N}$ denotes the computed attention matrix, and σ denotes the sigmoid activation function. During training, our module updates the attention scores between features via the learnable transformation matrices $\mathbf{W}_s$ and $\mathbf{W}_q$, enabling the model to better learn domain-agnostic similarities between features.

Thirdly, based on the notion that DSFBs and domain-agnostic frequency bands (DAFBs) should be treated differently, we apply distinct attention weightings to feature pairs in three frequency bands. To match the size of $\mathbf{F}_s'$ and $\mathbf{F}_q'$, we first obtain $\mathbf{A}' \in \mathbb{R}^{C \times N}$ by repeating $\mathbf{A}$ along the channel dimension. For feature pairs in DAFBs, We directly multiply the attention matrix with the features element-wise to

highlight the similar components while suppressing the dissimilar ones:

$$\begin{cases} \mathbf{F}''_{s,0} = \mathbf{A}' \odot \mathbf{F}'_s \\ \mathbf{F}''_{q,0} = \mathbf{A}' \odot \mathbf{F}'_q \end{cases}, \qquad (9)$$

where $\mathbf{F}''_{s,0}$ and $\mathbf{F}''_{q,0} \in \mathbb{R}^{C \times N}$ are the weighted features belonging to the DAFBs.

While for feature pairs in DSFBs, an inverse attention weighting is performed to suppress the similar components between features, thereby reducing the model's reliance on these components:

$$\begin{cases} \mathbf{F}''_{s,1} = (1 - \mathbf{A}') \odot \mathbf{F}'_s \\ \mathbf{F}''_{q,1} = (1 - \mathbf{A}') \odot \mathbf{F}'_q \end{cases}, \qquad (10)$$

where $\mathbf{F}''_{s,1}$ and $\mathbf{F}''_{q,1} \in \mathbb{R}^{C \times N}$ are the weighted features belonging to the DSFBs.

Thus, the FAM module completes the enhancement or suppression of full-spectrum similarity between features through $\mathbf{A}$, which is derived from full-spectrum attention weights. We pass the concatenated feature pair through an MLP afterward to fuse the features belonging to the support and query, obtaining a new fused feature that acts as the final feature representation of $\mathcal{B}$:

$$\mathbf{F}_b = \mathrm{MLP}\left(\mathrm{Cat}\left(\mathbf{F}''_{s,fb}, \mathbf{F}''_{q,fb}\right), \varphi\right), \quad fb \in \{0,1\} \quad (11)$$

where $\mathbf{F}_b \in \mathbb{R}^{C \times N}$ denotes the fused feature in a specific frequency band, $\mathrm{MLP}(\cdot, \varphi)$ is a function of a MLP with parameters φ. Notably, the MLP consists of two fully connected layers with a ReLU activation function in between, which can better learn the fusion patterns and reduce the parameters of the MLP, $\mathrm{Cat}(a, b)$ denotes the function that concatenates a and b along the last dimension, and fb serves as an indicator specifying DAFBs or DSFBs.

Multi-Spectral Fusion (MSF) Module

Recent research and our experiments in the Supplementary Materials have revealed the subtle relationship between frequency domain signals and image feature information: 1) Different frequency bands contain different information. Low and high frequencies contain color and style information, while the middle-frequency band contains more structural and shape information (Huang et al. 2021). 2) In cross-domain tasks, low and high frequencies exhibit significant differences across different domains. 3) Directly discarding high and low-frequency features is unreasonable, as frequency domain decomposition fails to completely decouple domain-variant information (DVI) and domain-invariant information (DII).

These insights lead us to a question: *Is there a mechanism that can extract the DII remaining in the DSFBs guided by the DII in the DAFBs?* We thus considered the cross-attention mechanism, which is widely used in multi-modal feature fusion: When feature $\mathbf{\Phi}$ is used as the key ($\mathbf{K}$) and value ($\mathbf{V}$), and another feature $\mathbf{\Psi}$ is used as the query ($\mathbf{Q}$), the resulting $\mathbf{V}$ is the representation of feature $\mathbf{\Phi}$ weighted by the similarity between feature $\mathbf{\Psi}$ and feature $\mathbf{\Phi}$.

Propelled by this knowledge, we propose the MSF module, a cross-attention-based feature fusion module. This module retains high and low-frequency information while using mid-frequency information to extract DII from high and low-frequency features and suppress DVI.

To be specific, for each feature triplet $(\mathbf{F}_f^l, \mathbf{F}_f^m, \mathbf{F}_f^h)$, the three fused features do not overlap in the frequency domain, *i.e.*, $\phi(\mathbf{F}_f^l) \cap \phi(\mathbf{F}_f^m) = \varnothing$, $\phi(\mathbf{F}_f^l) \cap \phi(\mathbf{F}_f^h) = \varnothing$, and $\phi(\mathbf{F}_f^m) \cap \phi(\mathbf{F}_f^h) = \varnothing$, where $\phi(\mathbf{F})$ represents the spectrum of $\mathbf{F}$. However, $\zeta(\mathbf{F}_f^l) \cap \zeta(\mathbf{F}_f^m) \neq \varnothing$ and $\zeta(\mathbf{F}_f^m) \cap \zeta(\mathbf{F}_f^h) \neq \varnothing$, where $\zeta(\mathbf{F})$ denotes the contained information of $\mathbf{F}$. We use $\mathbf{F}_f^l$ or $\mathbf{F}_f^h$, along with $\mathbf{F}_f^m$ to compute the attention between the given features. The output matrix of cross-attention (CA) can be represented as:

$$f_{\mathrm{CA}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} = \mathbf{S}\mathbf{V}, \qquad (12)$$

where d denotes a scaling factor, $\mathbf{S} \in \mathbb{R}^{N \times N}$ denotes the attention weight matrix. Consequently, the process of obtaining the refined features for the low and high-frequency components can be formalized by:

$$\begin{cases} \mathbf{F}_f^{l'} = f_{\mathrm{CA}}((\mathbf{F}_f^m)^T \mathbf{W}_Q, (\mathbf{F}_f^l)^T \mathbf{W}_K, (\mathbf{F}_f^l)^T \mathbf{W}_V)^T \\ \mathbf{F}_f^{h'} = f_{\mathrm{CA}}((\mathbf{F}_f^m)^T \mathbf{W}_Q, (\mathbf{F}_f^h)^T \mathbf{W}_K, (\mathbf{F}_f^h)^T \mathbf{W}_V)^T \end{cases}, \qquad (13)$$

where $\mathbf{F}_f^{l'}, \mathbf{F}_f^{h'}$ denote the refined fused foreground features, and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C \times C}$ are the learnable linear transformation matrices.

Finally, a straightforward addition is performed to integrate the features from the three frequency bands, followed by a ReLU activation function, as the module output $\mathbf{F}_f$:

$$\mathbf{F}_f = \mathrm{ReLU}(\mathbf{F}_f^{l'} + \mathbf{F}_f^m + \mathbf{F}_f^{h'}) \in \mathbb{R}^{C \times N}. \qquad (14)$$

We use the fused foreground features to compute the final query mask. A global average pooling (GAP) operation is performed to obtain the frequency-aware and query-informed foreground prototype:

$$\mathbf{p}_f^{fg}(c) = \frac{1}{N} \sum_{i=1}^N \mathbf{F}_f(c, i), \qquad (15)$$

where c denotes the channel index.

Hence, the final query foreground prediction of our proposed model can be calculated in a similar way in Eq. 2:

$$\widetilde{\mathbf{M}}_q^{fg} = 1.0 - \sigma\left(d(\mathbf{F}_q^{ini}, \mathbf{p}_f^{fg}) - \tau\right), \qquad (16)$$

while the background prediction can be obtained by $\widetilde{\mathbf{M}}_q^{bg} = 1 - \widetilde{\mathbf{M}}_q^{fg}$ accordingly.

Objective Function

We adopt the binary cross-entropy loss $\mathcal{L}_{ce}$ to evaluate the error between the predicted query mask and its corresponding ground truth. Mathematically, our final prediction loss $\mathcal{L}_{final}$ can be expressed as:

$$\mathcal{L}_{final} = \mathcal{L}_{ce}(\mathbf{M}^q, \widetilde{\mathbf{M}}_q^{fg}, \widetilde{\mathbf{M}}_q^{bg})$$
$$= -\frac{1}{HW} \sum_{h,w} \mathbf{M}^q \log(\widetilde{\mathbf{M}}_q^{fg}) + (1 - \mathbf{M}^q) \log(\widetilde{\mathbf{M}}_q^{bg}).$$
$$(17)$$

Method	Ref.	Abdominal CT → MRI					Abdominal MRI → CT				
		Liver	LK	RK	Spleen	Mean	Liver	LK	RK	Spleen	Mean
PANet	ICCV'19	39.24	26.47	37.35	26.79	32.46	40.29	30.61	26.66	30.21	31.94
SSL-ALP	TMI'22	70.74	55.49	67.43	**58.39**	63.01	71.38	34.48	32.32	51.67	47.46
ADNet	MIA'22	50.33	39.36	37.88	39.37	41.73	64.25	37.39	25.62	42.94	42.55
QNet	IntelliSys'23	58.82	42.69	51.67	44.58	49.44	70.98	38.64	30.17	43.28	45.77
CATNet	MICCAI'23	44.58	43.67	50.27	46.34	46.21	54.52	41.73	40.24	45.84	45.60
RPT	MICCAI'23	49.22	42.45	47.14	48.84	46.91	65.87	40.07	35.97	51.22	48.28
PATNet	ECCV'22	57.01	50.23	53.01	51.63	52.97	**75.94**	46.62	42.68	63.94	57.29
IFA	CVPR'24	48.81	45.79	51.46	51.42	49.37	50.05	36.45	32.69	43.08	40.57
Ours	—	**73.01**	**57.28**	**74.68**	58.21	**65.79**	73.57	**57.79**	**61.89**	**65.78**	**64.75**

Table 1: Quantitative comparison of different methods Dice score (%) on the Cross-Modality Dataset. The best value is shown in bold font, and the second best is underlined.

Method	Ref.	Cardiac LGE → b-SSFP				Cardiac b-SSFP → LGE			
		LV-BP	LV-MYO	RV	Mean	LV-BP	LV-MYO	RV	Mean
PANet	ICCV'19	51.43	25.75	25.75	36.66	36.24	26.37	23.47	28.69
SSL-ALP	TMI'22	83.47	22.73	66.21	57.47	65.81	25.64	51.24	47.56
ADNet	MIA'22	58.75	36.94	51.37	49.02	40.36	37.22	43.66	40.41
QNet	IntelliSys'23	50.64	37.88	45.24	44.58	31.08	34.03	39.45	34.85
CATNet	MICCAI'23	64.63	42.41	56.13	54.39	45.77	43.51	46.02	45.10
RPT	MICCAI'23	60.84	42.28	57.30	53.47	50.39	40.13	50.50	47.00
PATNet	ECCV'22	65.35	50.63	68.34	61.44	66.82	**53.64**	59.74	60.06
IFA	CVPR'24	64.04	43.22	74.58	62.28	68.07	36.07	**60.42**	54.85
Ours	—	**86.64**	**51.84**	**76.26**	**71.58**	**77.37**	52.05	54.75	**61.39**

Table 2: Quantitative comparison of different methods Dice score (%) on the Cross-Sequence Dataset. The best value is shown in bold font, and the second best is underlined.

To capture more precise and sufficient query foreground features, an accurate coarse prediction of the query foreground is needed. We continue to use the binary cross-entropy loss to quantify the dissimilarity between the coarse prediction and $\mathbf{M}_q$:

$$\mathcal{L}_{coarse} = \mathcal{L}_{ce}(\mathbf{M}_q, \widetilde{\mathbf{M}}_q^{coarse}, 1 - \widetilde{\mathbf{M}}_q^{coarse}). \qquad (18)$$

Overall, the computation of the total loss $\mathcal{L}_{total}$ for our proposed model can be denoted as $\mathcal{L}_{total} = \mathcal{L}_{final} + \mathcal{L}_{coarse}$.

Experiments

Datasets

We detail our proposed task into three cross-domain settings and evaluate our method on the following three datasets:

The Cross-Modality dataset comprises two abdominal datasets. The first is Abdominal MRI obtained from (Kavur et al. 2021), which includes 20 3D T2-SPIR MRI scans. The second is Abdominal CT, which comprises 20 3D abdominal CT scans from (Landman et al. 2015). We select four common categories from the two datasets: the left kidney (LK), right kidney (RK), liver, and spleen, for assessment.

The Cross-Sequence dataset is a cardiac dataset from (Zhuang et al. 2022), which includes 45 3D LGE MRI scans

and 45 b-SSFP MRI scans, both comprising 3 distinct labels: the blood pool (LV-BP), the left ventricle myocardium (LV-MYO), and the right ventricle myocardium (RV).

The Cross-Institution dataset consists of 321 3D prostate T2-weighted MRI scans collected by the University College London hospitals (UCLH) and 82 3D prostate MRI scans from the National Cancer Institute (NCI), Bethesda, Maryland, USA. The data from UCLH are collected from 4 studies: INDEX (Dickinson et al. 2013), the SmartTarget Biopsy Trial (Hamid et al. 2019), PICTURE (Simmons et al. 2014), Promise12 (Simmons et al. 2014), and organized by (Li et al. 2023). The data from NCI are provided in (Choyke et al. 2016). All the data are annotated by (Li et al. 2023). We select three common categories, bladder, central gland (CG) and rectum, for assessment.

Implementation Details

Our method is implemented on an NVIDIA GeForce RTX 4080S GPU. Initially, we employ the 3D supervoxel clustering method (Hansen et al. 2022) to generate pseudo-masks as the supervision in the episode-based meta-learning task, and we follow the same pre-processing techniques as (Hansen et al. 2022). The experiments are conducted under the 1-way 1-shot condition. During inference, we randomly

Baseline	CPG	FAM	MSF	CT → MRI				
				Liver	LK	RK	Spleen	Mean
✓				39.24	26.47	37.35	26.79	32.46
✓	✓			69.22	49.52	45.73	51.41	53.97
✓	✓	✓		71.68	55.45	67.20	53.75	62.02
✓	✓	✓	✓	**73.01**	**57.28**	**74.68**	**58.21**	**65.79**

Table 3: Ablation studies for the effect of each component in Dice score (%).

Frequency band			CT → MRI				
Low	Mid	High	Liver	LK	RK	Spleen	Mean
−	+	−	**73.01**	**57.28**	**74.68**	58.21	**65.79**
−	+	+	66.28	<u>55.68</u>	62.41	**60.79**	<u>61.29</u>
+	+	−	<u>68.14</u>	53.47	<u>64.09</u>	54.36	60.02
+	+	+	64.46	48.77	62.21	<u>59.28</u>	58.68

Table 4: Ablation study (in Dice score %) for the distinct attention weightings in DSFBs & DAFBs. Given attention matrix **A**, + indicates using **A** for attention weighting, and − indicates using 1-**A** for attention weighting.

sample a scan from the source domain and select a middle slice containing the foreground as the support image, with the remaining slices as the query images. For all datasets, we train the model for 39K iterations, comprising 3000 iterations per epoch with the batch size set to 1. To comprehensively test the performance of our proposed model, we conduct bidirectional evaluations within each dataset. For instance, in the Cross-Modality dataset, we evaluate performance both on CT → MRI and MRI → CT directions.

Additionally, for the training of our model, the output size N for the adaptive average pooling in FAM is set to 30^2. In the multi-spectral decoupling of foreground features, we divide the frequency band into low, mid, and high frequencies with a ratio of 3:4:3. We chose the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001, a momentum of 0.9 and a decay factor of 0.95 every 1K iterations.

Evaluation metric

In order to evaluate the model under a uniform standard, we adopt the Sorensen-Dice coefficient (Ouyang et al. 2022) that is commonly used in FSMIS tasks, as the evaluation metric. The Dice score is used to evaluate the overlap between the segmentation results and the ground truth, which can be denoted as

$$\text{DSC}(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{19}$$

where X and Y denote the two masks respectively, and the DSC denotes the Dice score ranges from 0 to 1, with 1 indicating complete overlap and 0 indicating no overlap.

Quantitative and Qualitative Results

To demonstrate the effectiveness of our proposed method, we compare its performance with various FSMIS models, including PANet (Wang et al. 2019), SSL-ALPNet (Ouyang et al. 2022), ADNet (Hansen et al. 2022), QNet (Shen et al. 2023), CATNet (Lin et al. 2023), and RPT (Zhu et al. 2023). Additionally, two CD-FSS models, PATNet (Lei et al. 2022) and IFA (Nie et al. 2024) are also used for comparison. Note that we evaluate CD-FSS models without fine-tuning.

As shown in Table 1, our proposed method significantly outperforms all existing FSMIS and CD-FSS models under both CT → MRI and MRI → CT directions. Specifically, in the CT → MRI direction, the Dice score reached 65.79%, which is 2.78% higher than the second-best method. More significantly, the proposed model exhibited an overall 7.46%

higher performance compared to the highest corresponding method in the MRI → CT direction. While PATNet performs 2.37% better than our model in the liver category in the MRI → CT direction, it underperforms in smaller categories like RK, LK, and spleen. This discrepancy arises from the imbalanced foreground and background in medical images, which the CD-FSS models do not adequately address. In contrast, our model is better adapted to medical scenarios, resulting in a higher overall Dice score.

As depicted in Table 2, Our method consistently performs exceptionally well on the Cross-Sequence dataset compared to other methods, achieving the highest Dice scores of 71.58% and 61.39% in both directions, surpassing the second-best method by 10.14% and 1.33%, respectively. In the LV-BP category under the LGE → b-SSFP scenario, our model even surpasses the second-best model by 21.29%, reaching 86.64%, which is comparable to FSMIS models' segmentation accuracy in non-cross-sequence conditions.

For quantitative and qualitative results on the Cross-Institution dataset and visual segmentation results on three cross-domain datasets, please refer to the Supplementary Materials. All experiments demonstrate that our FAMNet is a medical image segmentation model with excellent generalization capabilities and minimal data dependency.

Ablation Studies

Effect of each component. In this section, we discuss the effect of each component. Table 3 shows the contribution of each module to the overall model performance. Combined with CPG, the proposed FAM module significantly enhances baseline (PANet) performance by 29.56%, primarily by mitigating overfitting to DSFBs and implementing inter-domain debiasing for the support and query features. Additionally, the MSF module aids in integrating the frequency-decoupled features from the FAM module, further suppressing DVI, and contributing an additional 3.77% improvement in model performance.

Distinct attention weightings in DAFBs & DSFBs. In the FAM module, we apply distinct attention weighting methods to assign weights to features belonging to DSFBs and DAFBs. This approach reduces the model's dependency on the support-query correlation within DSFBs while enhancing its focus on DAFBs. Table 4 illustrates the ablation study for distinct attention weightings. It is evident that using the uniform attention weighting method throughout leads to a

significant drop in the Dice score, with a reduction of 7.11%, and applying positive attention weighting to any DSFB results in a decrease in model performance. This decline is attributed to substantial overfitting caused by the attention mechanism to the source domain's support-query correlation, which weakens the model's ability to generalize to novel target domains.

For further discussion, we present extensive ablation experiments in the Supplementary Materials, which include more method comparisons and hyperparameter analysis.

Conclusion

In this paper, we have addressed a novel task: cross-domain few-shot medical image segmentation (CD-FSMIS). We proposed a Frequency-aware Matching Network (FAMNet), which comprises a Frequency-aware Matching (FAM) module to enhance the model's generalization capabilities and reduce support-query bias by performing attention-based matching of the foreground features for specific frequency bands, which simultaneously handle intra-domain and inter-domain variations. Furthermore, we introduced a Multi-Spectral Fusion (MSF) module to integrate features decoupled by the FAM module and further suppress the detrimental impact of domain-variant information on the model's robustness. Extensive experiments on three cross-domain datasets demonstrated the excellent generalization ability and data independence of the proposed method.

Acknowledgment

References

Chen, H.; Dong, Y.; Lu, Z.; Yu, Y.; and Han, J. 2024. Pixel Matching Network for Cross-Domain Few-Shot Segmentation. In *WACV*, 978–987.

Cheng, Z.; Wang, S.; Xin, T.; Zhou, T.; Zhang, H.; and Shao, L. 2024. Few-Shot Medical Image Segmentation via Generating Multiple Representative Descriptors. *IEEE Transactions on Medical Imaging*, 43(6): 2202 – 2214.

Choyke, P.; Turkbey, B.; Pinto, P.; Merino, M.; and Wood, B. 2016. Data from PROSTATE-MRI. The Cancer Imaging Archive.

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *CVPR*, 172–17209.

Dickinson, L.; Ahmed, H.; Kirkham, A.; Allen, C.; Freeman, A.; Barber, J.; Hindley, R.; Leslie, T.; Ogden, C.; Persad, R.; Winkler, M.; and Emberton, M. 2013. A multi-centre prospective development study evaluating focal therapy using high intensity focused ultrasound for localised prostate cancer: The INDEX study. *Contemporary Clinical Trials*, 36(1): 68–80.

Ding, H.; Sun, C.; Tang, H.; Cai, D.; and Yan, Y. 2023. Few-shot Medical Image Segmentation with Cycle-resemblance Attention. In *WACV*, 2487–2496.

Feng, R.; Zheng, X.; Gao, T.; Chen, J.; Wang, W.; Chen, D. Z.; and Wu, J. 2021. Interactive Few-Shot Learning: Limited Supervision, Better Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10): 2575–2588.

Guha Roy, A.; Siddiqui, S.; Pölsterl, S.; Navab, N.; and Wachinger, C. 2019. 'Squeeze & Excite' Guided Few-Shot Segmentation of Volumetric Images. *Medical Image Analysis*, 59: 101587.

Hamid, S.; Donaldson, I. A.; Hu, Y.; Rodell, R.; Villarini, B.; Bonmati, E.; Tranter, P.; Punwani, S.; Sidhu, H. S.; Willis, S.; van der Meulen, J.; Hawkes, D.; McCartan, N.; Potyka, I.; Williams, N. R.; Brew-Graves, C.; Freeman, A.; Moore, C. M.; Barratt, D.; Emberton, M.; and Ahmed, H. U. 2019. The SmartTarget Biopsy Trial: A Prospective, Within-person Randomised, Blinded Trial Comparing the Accuracy of Visual-registration and Magnetic Resonance Imaging/Ultrasound Image-fusion Targeted Biopsies for Prostate Cancer Risk Stratification. *European Urology*, 75(5): 733–740.

Hamming, R. W. 1977. *Digital Filters*. Signal Processing Series. Englewood Cliffs: Prentice–Hall.

Hansen, S.; Gautam, S.; Jenssen, R.; and Kampffmeyer, M. 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78: 102385.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

He, W.; Zhang, Y.; Zhuo, W.; Shen, L.; Yang, J.; Deng, S.; and Sun, L. 2024. APSeg: Auto-Prompt Network for Cross-Domain Few-Shot Semantic Segmentation. In *CVPR*, 23762–23772.

Herzog, J. 2024. Adapt Before Comparison: A New Perspective on Cross-Domain Few-Shot Segmentation. In *CVPR*, 23605–23615.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021. FSDR: Frequency Space Domain Randomization for Domain Generalization. In *CVPR*, 6891–6902.

Kavur, A. E.; Gezer, N. S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D. D.; Chatterjee, S.; Ernst, P.; Özkan, S.; Baydar, B.; Lachinov, D.; Han, S.; Pauli, J.; Isensee, F.; Perkonigg, M.; Sathish, R.; Rajan, R.; Sheet, D.; Dovletov, G.; Speck, O.; Nürnberger, A.; Maier-Hein, K. H.; Bozdağı Akar, G.; Ünal, G.; Dicle, O.; and Selver, M. A. 2021. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69: 101950.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Workshop*, 12.

Lei, S.; Zhang, X.; He, J.; Chen, F.; Du, B.; and Lu, C.-T. 2022. Cross-Domain Few-Shot Semantic Segmentation. In *ECCV*, 73–90.

Li, Y.; Fu, Y.; Gayo, I. J.; Yang, Q.; Min, Z.; Saeed, S. U.; Yan, W.; Wang, Y.; Noble, J. A.; Emberton, M.; et al. 2023. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *Medical Image Analysis*, 90: 102935.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Lin, Y.; Chen, Y.; Cheng, K.-T.; and Chen, H. 2023. Few Shot Medical Image Segmentation with Cross Attention Transformer. In *MICCAI*, 233–243.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path Aggregation Network for Instance Segmentation. In *CVPR*, 8759–8768.

Nie, J.; Xing, Y.; Zhang, G.; Yan, P.; Xiao, A.; Tan, Y.-P.; Kot, A. C.; and Lu, S. 2024. Cross-Domain Few-Shot Segmentation via Iterative Support-Query Correspondence Mining. In *CVPR*, 3380–3390.

Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; and Rueckert, D. 2022. Self-Supervised Learning for Few-Shot Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 41(7): 1837–1848.

Ouyang, C.; Chen, C.; Li, S.; Li, Z.; Qin, C.; Bai, W.; and Rueckert, D. 2021. Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 42: 1095–1106.

Shen, Q.; Li, Y.; Jin, J.; and Liu, B. 2023. Q-Net: Query-Informed Few-Shot Medical Image Segmentation. In Arai, K., ed., *Intelligent Systems and Applications*, 610–628.

Simmons, L. A.; Ahmed, H. U.; Moore, C. M.; Punwani, S.; Freeman, A.; Hu, Y.; Barratt, D.; Charman, S. C.; Van der Meulen, J.; and Emberton, M. 2014. The PICTURE study — Prostate Imaging (multi-parametric MRI and Prostate HistoScanning™) Compared to Transperineal Ultrasound guided biopsy for significant prostate cancer Risk Evaluation. *Contemporary Clinical Trials*, 37(1): 69–83.

Su, J.; Fan, Q.; Pei, W.; Lu, G.; and Chen, F. 2024. Domain-Rectifying Adapter for Cross-Domain Few-Shot Segmentation. In *CVPR*, 24036–24045.

Su, Z.; Yao, K.; Yang, X.; Huang, K.; Wang, Q.; and Sun, J. 2023. Rethinking Data Augmentation for Single-Source Domain Generalization in Medical Image Segmentation. In *AAAI*, 2366–2374.

Sun, L.; Li, C.; Ding, X.; Huang, Y.; Chen, Z.; Wang, G.; Yu, Y.; and Paisley, J. 2022. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in Biology and Medicine*, 140: 105067.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In *ICCV*, 9196–9205.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wei, T.; Li, X.; Chen, Y. P.; Tai, Y.-W.; and Tang, C.-K. 2019. FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation. In *CVPR*, 2866–2875.

Xu, Y.; Xie, S.; Reynolds, M.; Ragoza, M.; Gong, M.; and Batmanghelich, K. 2022. Adversarial Consistency for Single Domain Generalization in Medical Image Segmentation. In *MICCAI*, 671–681.

Zhou, Z.; Qi, L.; Yang, X.; Ni, D.; and Shi, Y. 2022. Generalizable Cross-modality Medical Image Segmentation via Style Augmentation and Dual Normalization. In *CVPR*, 20856–20865.

Zhu, Y.; Wang, S.; Xin, T.; and Zhang, H. 2023. Few-Shot Medical Image Segmentation via a Region-Enhanced Prototypical Transformer. In *MICCAI*, 271–280. Springer.

Zhuang, X.; Xu, J.; Luo, X.; Chen, C.; Ouyang, C.; Rueckert, D.; Campello, V. M.; Lekadir, K.; Vesal, S.; RaviKumar, N.; Liu, Y.; Luo, G.; Chen, J.; Li, H.; Ly, B.; Sermesant, M.; Roth, H.; Zhu, W.; Wang, J.; Ding, X.; Wang, X.; Yang, S.; and Li, L. 2022. Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge. *Medical Image Analysis*, 81: 102528.

Appendix for "FAMNet: Frequency-aware Matching Network for Cross-domain Few-shot Medical Image Segmentation"

Yuntian Bo, Yazhou Zhu, Lunbo Li, Haofeng Zhang*

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
{yuntian.bo, zyz_nj, lunboli, zhanghf}@njust.edu.cn

Ablation Studies

Unless otherwise specified, all experiments are conducted using the same training setting and model configuration, consistent with the implementation details.

Attention Mechanisms	Affected Fg Pixels	CT → MRI				
		Liver	LK	RK	Spleen	Mean
Hard Attention	20%	74.28	56.49	_69.44_	55.25	63.87
	50%	_74.29_	**58.47**	66.45	_57.00_	_64.05_
	80%	**76.44**	56.06	64.66	53.54	62.68
Soft Attention	100%	73.01	_57.28_	**74.68**	**58.21**	**65.79**

Table 1: Comparison of hard attention at different dropout rates and soft attention in Dice score (%).

Attention Mechanism in FAM

In this section, we discuss the implementation of attention in the FAM module. In our experiments, the FAM module employs soft attention to weight the support and query foreground features, thereby mitigating the model's reliance on DSFBs. An alternative approach involves using hard attention to discard similar points directly. We conducted analytical experiments to analyze the feasibility of this method and compared the results with our soft attention approach. The comparative results in Table 1 demonstrated the superiority of our method. Specifically, we selected the top 20%, 50%, 80% of points based on the computed similarity matrix and discarded them accordingly.

Feasibility of Directly Discarding Frequency Bands

In a previous section "Multi-Spectral Fusion (MSF) Module", we concluded that directly discarding DSFBs is unreasonable. Table 2 presents experimental evidence supporting this conclusion. The experiments were conducted using the FAM module but without the inclusion of the MSF module. We observed that discarding any one or more frequency bands adversely affects the model's performance compared to using all frequency bands, with a reduction of up to 3.90%.

Frequency band			CT → MRI				
Low	Mid	High	Liver	LK	RK	Spleen	Mean
✓	✓	✓	_71.68_	_55.45_	**67.20**	53.75	**62.02**
	✓	✓	71.62	54.89	64.21	_55.54_	_61.57_
✓	✓		70.77	**55.61**	_66.64_	53.12	61.54
		✓	**71.71**	50.05	55.06	**55.66**	58.12

Table 2: The impact of directly discarding a specific frequency band in Dice score (%).

Based on these results, we conjecture that this outcome is related to the content-irrelevance of frequency decoupling. For instance, MRI images contain a large amount of fine textures that are absent in CT images, causing a domain shift. When discarding the high-frequency band, the fine texture information should be discarded as well. However, the high-frequency band also includes edge information, which plays a significant positive role in segmentation (Cheng et al. 2024), and this information remains consistent between CT and MRI (edge information of organs is similar in both modalities). Directly discarding the high-frequency band leads to the loss of both DVI and DII, resulting in the model learning in an information-deficient environment and thus decreasing segmentation performance. Furthermore, directly discarding a specific frequency band prevents the attention weights in that band from being trained, leading to over-fitting in the remaining frequency bands.

Performing Matching Exclusive to Specific Frequency Bands

Instead of matching features across all frequency bands, this section focuses on matching exclusive to specific frequency bands. At least one DAFB is used to train the attention weights for support-query correlation to prevent overfitting issues, as detailed in the section "Distinct Attention Weightings in DAFBs & DSFBs". The results of the study are shown in Table 3. When the low or high-frequency bands are not matched, the Dice score decreases by 2.59% and 2.01%, respectively, compared to matching across all frequency bands. Excluding both low and high-frequency bands from matching results in a minimum Dice score of 61.63%. This decline occurs because the query sample may

| Frequency band | | | CT → MRI | | | | |
Low	Mid	High	Liver	LK	RK	Spleen	Mean
✓	✓	✓	<u>73.01</u>	**57.28**	**74.68**	**58.21**	**65.79**
	✓	✓	72.62	55.68	<u>67.25</u>	57.26	63.20
✓	✓		**74.01**	<u>57.08</u>	66.62	<u>57.40</u>	<u>63.78</u>
	✓		72.90	56.00	64.26	53.05	61.63

Table 3: Ablation study (in Dice score %) for the impact of matching is exclusive to specific frequency bands. ✓ signifies feature matching in the corresponding frequency band.

| Division Ratio (Low:Mid:High) | | | CT → MRI | | | | |
Low	Mid	High	Liver	LK	RK	Spleen	Mean
2.0	6.0	2.0	**75.11**	56.49	66.35	57.27	63.81
2.5	5.0	2.5	73.19	57.06	70.75	56.06	64.27
3.0	4.0	3.0	73.01	57.28	**74.68**	<u>58.21</u>	65.79
3.5	3.0	3.5	72.75	**60.52**	<u>74.44</u>	57.44	**66.29**
4.0	2.0	4.0	<u>74.01</u>	<u>59.35</u>	72.77	57.37	<u>65.88</u>
4.5	1.0	4.5	68.68	55.77	68.42	**60.31**	63.30

Table 4: Ablation study (in Dice score %) for the impact of different frequency band division ratios.

be out of distribution or the support prototype may not accurately represent the mean of a category due to potential intra-domain shifts between support and query samples, a phenomenon particularly evident in the 1-shot setting. These inter-domain variations have detrimental effects on the segmentation performance of the CD-FSMIS model. Figure 3 illustrates the debiasing effect achieved by our proposed FAM module.

Impact of Different Frequency Band Division Ratios

In this section, we discuss the impact of different frequency band division ratios on model performance. The division ratio determines our segmentation of DSFBs and DAFBs. In the FAM module, we apply different attention weightings to the image features belonging to DSFBs and DAFBs based on this segmentation. In the MSF module, we extract DII from features in DAFBs to extract residual DII from DSFBs, while simultaneously suppressing DVI in the final module output. Table 4 shows the impact of different division ratios. From Table 4, we observe that our model's performance peaks with a Dice score of 66.29% when the low:mid:high ratio is 3.5:3:3.5. When the proportion of the mid-frequency band exceeds 30%, the model's performance shows a decreasing trend as the mid-frequency band increases. We attribute this to the blurred boundary between DSFBs and DAFBs. During frequency band division, DAFBs inevitably include some frequency domain information from DSFBs. As the bandwidth of the mid-frequency band increases, DAFBs contain more domain-specific frequency signals. This deviates from the design intent of the FAM and MSF modules. In the FAM module, certain support-query matching relationships from DSFBs are also reinforced, which could be crucial for optimization during source domain training, leading to overfitting of source domain information. In the MSF module, since the mid-frequency band contains DVI, its guiding significance in the cross-attention extraction process decreases, resulting in certain DVI in DSFBs being enhanced rather than suppressed. When the proportion of the mid-frequency band falls below 30%, the model's performance declines sharply due to insufficient information in the mid-frequency band, rendering both the FAM and MSF modules inadequately trained.

Impact of Foreground Pixel Number N

In this section, we analyze the fixed number N of foreground pixels, which standardizes the number of foreground pixels through the adaptive average pooling. In our experiment, we selected N as 30^2. We further vary N and evaluate the model's performance in CT → MRI direction, as illustrated in Figure 1. It can be observed that as N increases, the Dice score initially rises steadily and then fluctuates around a certain value. When N is 60^2, the model achieves the highest Dice score, but the difference compared to the reported results is minimal. Meanwhile, since the value of N is directly related to the subsequent network design, a large N introduces a significant number of unnecessary parameters and computational load. Hence, it is crucial to choose a moderate value of N to maintain our model's efficiency.
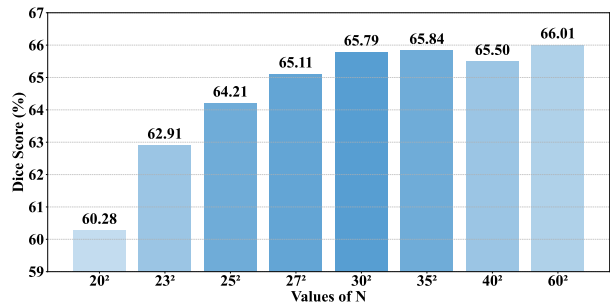

Figure 1: Ablation study (in Dice score %) for the impact of different N values.

Quantative and Qualitative Result on the Cross-Institution Dataset

As shown in Table 5, our FAMNet also excels on the Cross-Institution dataset, achieving a 2.27% and 3.18% higher Dice score in the UCLH → NCI and NCI → UCLH directions, respectively, compared to the second-best method. The performance improvement is attributed to FAMNet's ability to better address the variations arising from the use of instruments of different models or different manufacturers across institutions. These discrepancies typically result in differences in image contrast, signal-to-noise ratio, and other factors, which are particularly pronounced in the fre-

Method	Ref.	Prostate UCLH → NCI				Prostate NCI → UCLH			
		Bladder	CG	Rectum	Mean	Bladder	CG	Rectum	Mean
PANet	ICCV'19	51.85	38.89	36.28	42.34	49.90	36.27	43.58	43.25
SSL-ALP	TMI'22	**54.86**	42.59	41.65	46.37	66.6	43.12	56.94	55.55
ADNet	MIA'22	_53.92_	46.11	42.26	47.43	62.84	51.34	60.45	58.21
QNet	IntelliSys'23	39.66	35.71	34.25	36.54	41.22	43.21	39.19	41.21
CATNet	MICCAI'23	47.61	45.29	42.51	45.14	51.98	48.81	55.01	51.93
RPT	MICCAI'23	52.17	41.22	_46.93_	46.77	62.35	_51.66_	_64.63_	59.55
PATNet	ECCV'22	50.04	**52.71**	44.57	_49.10_	_69.91_	**53.02**	62.79	_61.91_
IFA	CVPR'24	41.02	39.15	36.45	38.87	52.54	39.85	34.27	42.22
Ours	—	53.06	_48.48_	**52.57**	**51.37**	**77.64**	48.73	**68.91**	**65.09**

Table 5: Quantitative comparison of different methods Dice score (%) on the Cross-Institution Dataset. The best value is shown in bold font, and the second best is underlined.
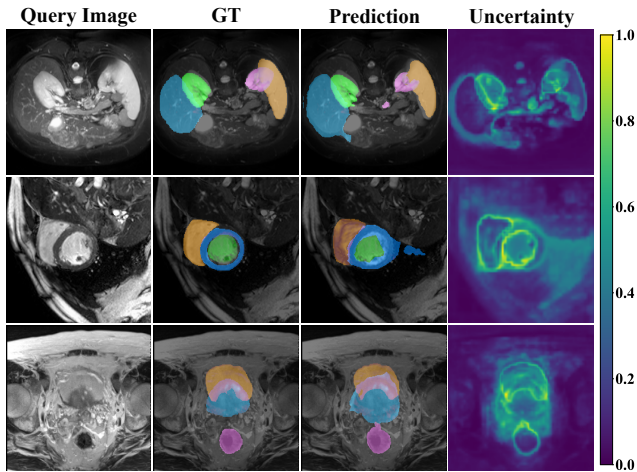

Figure 2: Illustration of query images, ground truths, predictions, and uncertainty maps in the directions of CT → MRI, LGE → b-SSFP, NCI → UCLH by FAMNet.

Figure 3: T-SNE visualization for intra-domain variations, *i.e.*, support-query bias, and debiasing capability of FAM-Net. 'fg' represents the foreground. In (b), the foregrounds in the support and query images exhibit significant differences in brightness and contrast, with notable discrepancies in texture and structural details as well. In (c), there is a pronounced size imbalance in the foregrounds, accompanied by substantial variations in brightness.

quency domain, especially in the high and low-frequency bands (Tang, Peli, and Acton 2003; Bernhardt et al. 2005). FAMNet effectively mitigates the impact of these domain shifts on segmentation performance, thereby enhancing its robustness and accuracy across different institutional data.

Visualization

Uncertainty Maps

Figure 2 illustrates three segmentation examples by FAM-Net from the CD-FSMIS task in the directions of CT → MRI, LGE → b-SSFP, and NCI → UCLH. By observing the predictions and uncertainty maps, it is evident that the edges of the target regions are typically highlighted. This indicates a higher model uncertainty in segmenting these areas, due to the differences in edge depiction across domains, *e.g.*, CT images generally have sharper edges compared to MRI images. Additionally, the model exhibits greater uncertainty when segmenting smaller targets compared to larger organs. This is attributed to the varying focus on detailed re-
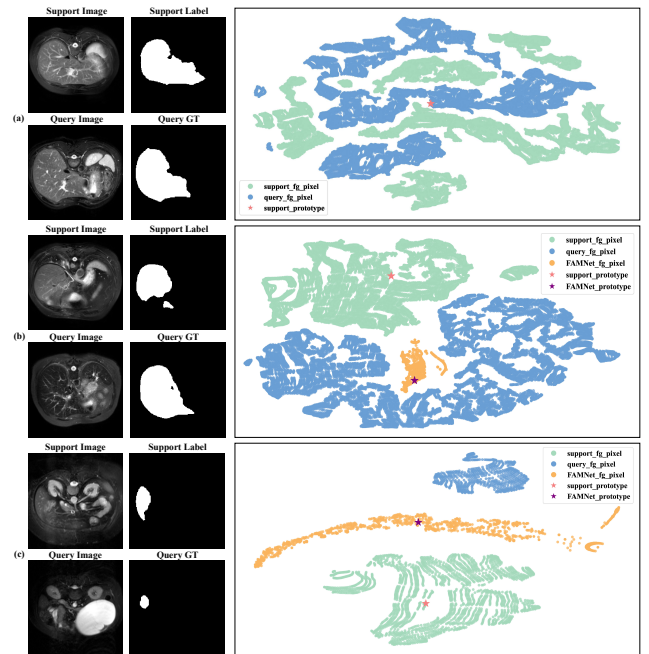
gions by different imaging techniques. Furthermore, when multiple targets are in close proximity or intersect, there is an increased uncertainty, which may lead to false segmentation.

FAMNet's Capability to Mitigate Intra-domain Variations

Figure 3 visualizes the capability of our FAMNet to significantly narrow the support-query bias using t-SNE (Van der Maaten and Hinton 2008). As shown in Figure 3(a), typical support and query samples are closely situated and highly intermixed in the feature space, and the support prototype effectively represents the mean of the query foreground's overall distribution, resulting in excellent segmentation performance. Figure 3(b)(c) illustrates that when there is a moderate or significant bias between support and query samples, the support prototype deviates from the cluster of query foreground pixels in the feature space. This deviation reduces the representational capability of the support prototype, thereby degrading the performance of the cosine similarity-based mask calculation method. FAMNet calculates a new prototype through support-query matching, correcting the bias of the support prototype, thereby reducing or eliminating the detrimental impact of intra-domain variation on segmentation results.

References

Bernhardt, P.; Batz, L.; Ruhrnschopf, E.-P.; and Hoheisel, M. 2005. Spatial frequency-dependent signal-to-noise ratio as a generalized measure of image quality. In *Medical Imaging 2005: Physics of Medical Imaging*, volume 5745, 407–418. SPIE.

Cheng, Z.; Wang, S.; Xin, T.; Zhou, T.; Zhang, H.; and Shao, L. 2024. Few-Shot Medical Image Segmentation via Generating Multiple Representative Descriptors. *IEEE Transactions on Medical Imaging*, 43(6): 2202 – 2214.

Tang, J.; Peli, E.; and Acton, S. 2003. Image enhancement using a contrast measure in the compressed domain. *IEEE Signal Processing Letters*, 10(10): 289–292.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).