Concentrate on Weakness: Mining Hard Prototypes for Few-Shot Medical Image Segmentation

Jianchao Jiang and **Haofeng Zhang**[*]

School of Computer Science and Engineering, Nanjing University of Science and Technology, China

{jiangjianchao, zhanghf}@njust.edu.cn

Abstract

Few-Shot Medical Image Segmentation (**FSMIS**) has been widely used to train a model that can perform segmentation from only a few annotated images. However, most existing prototype-based FSMIS methods generate multiple prototypes from the support image solely by random sampling or local averaging, which can cause particularly severe boundary blurring due to the tendency for normal features accounting for the majority of features of a specific category. Consequently, we propose to focus more attention to those weaker features that are crucial for clear segmentation boundary. Specifically, we design a Support Self-Prediction (**SSP**) module to identify such weak features by comparing true support mask with one predicted by global support prototype. Then, a Hard Prototypes Generation (**HPG**) module is employed to generate multiple hard prototypes based on these weak features. Subsequently, a Multiple Similarity Maps Fusion (**MSMF**) module is devised to generate final segmenting mask in a dual-path fashion to mitigate the imbalance between foreground and background in medical images. Furthermore, we introduce a boundary loss to further constrain the edge of segmentation. Extensive experiments on three publicly available medical image datasets demonstrate that our method achieves state-of-the-art performance. Code is available at https://github.com/jcjiang99/CoW.

1 Introduction

Automatic segmentation of organs, tissues and lesions appearing in medical images is broadly applied in clinical research, including disease diagnosis [Tsochatzidis *et al.*, 2021], treatment planning [El Naqa *et al.*, 2007; Zaidi and El Naqa, 2010; Chen *et al.*, 2021] and surgical guidance [Wang *et al.*, 2018; Hesamian *et al.*, 2019]. Recently, fully supervised deep learning segmentation frameworks [Fang and Yan, 2020; Shi *et al.*, 2021; Isensee *et al.*, 2021] exhibit superior performance, but they heavily depend on large amounts
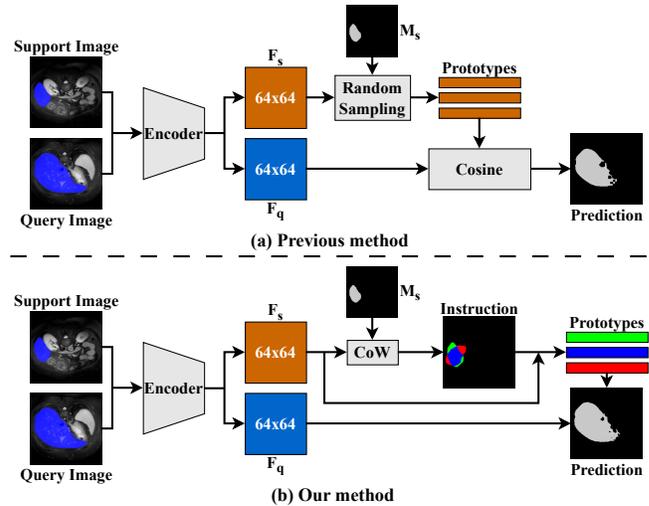
---

[*]Corresponding author.

Figure 1: Comparison between previous methods and ours. (a) The previous methods generate multiple prototypes only by random sampling or local averaging, giving insufficient attention to weak boundary details. (b) Our proposed CoW instead concentrates more on these weak features of a specific category, thus acquiring a more comprehensive representation of the category distribution and achieving clearer segmentation boundary.

of densely annotated training data, which is impractical for medical scenarios for the reason that labeling medical images requires considerable clinical expertise, cost and time [Luo *et al.*, 2021].

To solve this challenge, Few-Shot Learning (FSL) [Snell *et al.*, 2017; Sung *et al.*, 2018; Garcia and Bruna, 2017; Finn *et al.*, 2017] has been proposed as a potential learning paradigm to train models that can efficiently adapt to novel classes when given extremely scarce annotated data [Shaban *et al.*, 2017], based on which corresponding segmentation method is called Few-Shot Medical Image Segmentation (FSMIS). In particular, FSMIS typically trains models with base classes and segment unseen classes in the inference stage. Consequently, models trained in this manner are capable of segmenting novel organ or lesion leveraging solely a few annotated data, evading the need for retraining, which holds significant importance for medical image segmentation.

As one of dense image prediction tasks, FSMIS has an in-

dispensable demand for features with strong category information and precise spatial boundary details. In order to leverage the annotated samples, the majority of current FSMIS models distill prior knowledge based on prototype network [Ouyang *et al.*, 2020; Shen *et al.*, 2023; Wu *et al.*, 2022], among which multiple prototypes-based network is the predominant method. However, these methods generate multiple prototypes from support images either by a random selecting strategy or simple local averaging, which incline to normal features in the central area of a specific organ or involve background information in foreground prototypes as illustrated in Fig. 1(a). Furthermore, those weaker features mainly distributed along the boundary of the organ are neglected to a considerable extent, which can lead to a poorly segmentation boundary. As a result, some foreground and background points will be misclassified due to the excessive focus on normal features.

In this paper, we propose a Concentrate on Weakness (CoW) approach that learns to mine hard prototypes for each category to represent its feature distribution more comprehensively as illustrated in Fig. 1(b), to tackle the aforementioned problem. Specifically, we devise a Support Self-Prediction (SSP) module to provide knowledge for weak features carrying critical boundary information, which are subsequently utilized by the Hard Prototypes Generation (HPG) module to generate hard prototypes to improve segmentation accuracy and robustness. Right after is the Multiple Similarity Maps Fusion (MSMF) module that fuses multiple similarity maps to incorporate comprehensive information included in them to get the final predicted mask. To acquire a more precise segmenting boundary, we also devise a boundary loss, which effectively suppress over-segmentation whether for the foreground or the background. In conclusion, our main contributions are summarized as follows:

- We propose a novel method to identify weak features by comparing predicted support mask and its true label. Besides, multiple hard prototypes are generated by further processing those weak features, ensuring more comprehensive representation of overall distribution.

- To make full use of the information provided by generated multiple prototypes, we design an MSMF module to fuse multiple similarity maps predicted by these prototypes to get the final mask.

- Extensive experiments conducted on three popular medical image datasets reveal that our proposed model outperforms other existing state-of-the-art (SOTA) models and exhibits significant performance improvements.

2 Related Work

2.1 Medical Image Segmentation

With the development of Convolutional Neural Networks (CNNs), deep learning methods have become the dominated approach to MIS on various tumor [Peiris *et al.*, 2022], anatomical structures [Dolz *et al.*, 2018] and lesions [Li *et al.*, 2018] recently. Fully Convolutional Networks (FCNs), which replace the fully connected layers of standard CNNs with fully convolutional layers, prove to be a powerful net

architecture for semantic segmentation [Long *et al.*, 2015]. Afterwards, the encoder-decoder networks gradually become the prevailing architecture for semantic segmentation, among which U-Net is a widely recognized method for MIS with remarkable results [Ronneberger *et al.*, 2015]. Besides, most recent MIS networks are designed based on visual transformer [Cao *et al.*, 2022], multi-scale fusion [Wu *et al.*, 2019], feature pyramid [Chen *et al.*, 2018] and self-attention mechanism [Gu *et al.*, 2020]. Unfortunately, all these models require abundant densely annotated data.

2.2 Few-Shot Medical Image Segmentation

FSMIS is a promising approach since it can work with a few labeled samples. The prototypical network [Wang *et al.*, 2019; Ouyang *et al.*, 2020], which extracts representative prototypes of semantic classes from the support, is one of prevalent FSMIS methods featured by its strong interpretability. PANet [Wang *et al.*, 2019] argues that parametric segmentation generalizes poorly and proposes a non-parametric prototype matching method with a prototype alignment regularization. ADNet [Hansen *et al.*, 2022] builds further on the branch initiated by PANet [Wang *et al.*, 2019] and does not explicitly model the complex background class namely relying solely on one foreground prototype. Q-Net [Shen *et al.*, 2023] replaces fixed thresholds used in ADNet with adaptive ones learned from the query to mitigate distribution shifts between the query image and the support set, and simultaneously predicts the query segmentation masks with dual-scale features. Most of these methods solely employ masked average pooling to extract a single prototype, which can not well represent the distribution of a specific category.

Unlike these single prototype-based methods, SSL-ALPNet [Ouyang *et al.*, 2020] devises an adaptive local prototype pooling module where additional local prototypes are computed on a regular grid to preserve local information lost in global average operation. AAS-DCL [Wu *et al.*, 2022] constructs prototypical contrastive learning and contextual contrastive learning among the labeled support guidance information and unlabeled tissue knowledge, to strengthen the discriminability of the prototypes. GMRD [Cheng *et al.*, 2024] randomly selects multiple prototypes from both foreground and background to comprehensively represent the commonality within the corresponding class distribution. However, these methods pay no particular attention to weak boundary features even though they have constructed relatively comprehensive representation of each class with multiple prototypes.

3 Methodology

3.1 Problem Definition

The FSMIS task aims at training a model on a dataset D_{train} that contains labeled samples belonging to seen classes C_{base}, allowing a quick adaptation to segmentation of previously unseen novel classes C_{novel} when exposed to just a few annotated samples without retraining, where C_{base} and C_{novel} are disjoint, namely $C_{base} \cap C_{novel} = \varnothing$. In general, the episodic [Vinyals *et al.*, 2016; Dong and Xing, 2018] meta-learning setting is commonly utilized in FSS tasks. Specifically, by means of random sampling, the training set and the testing
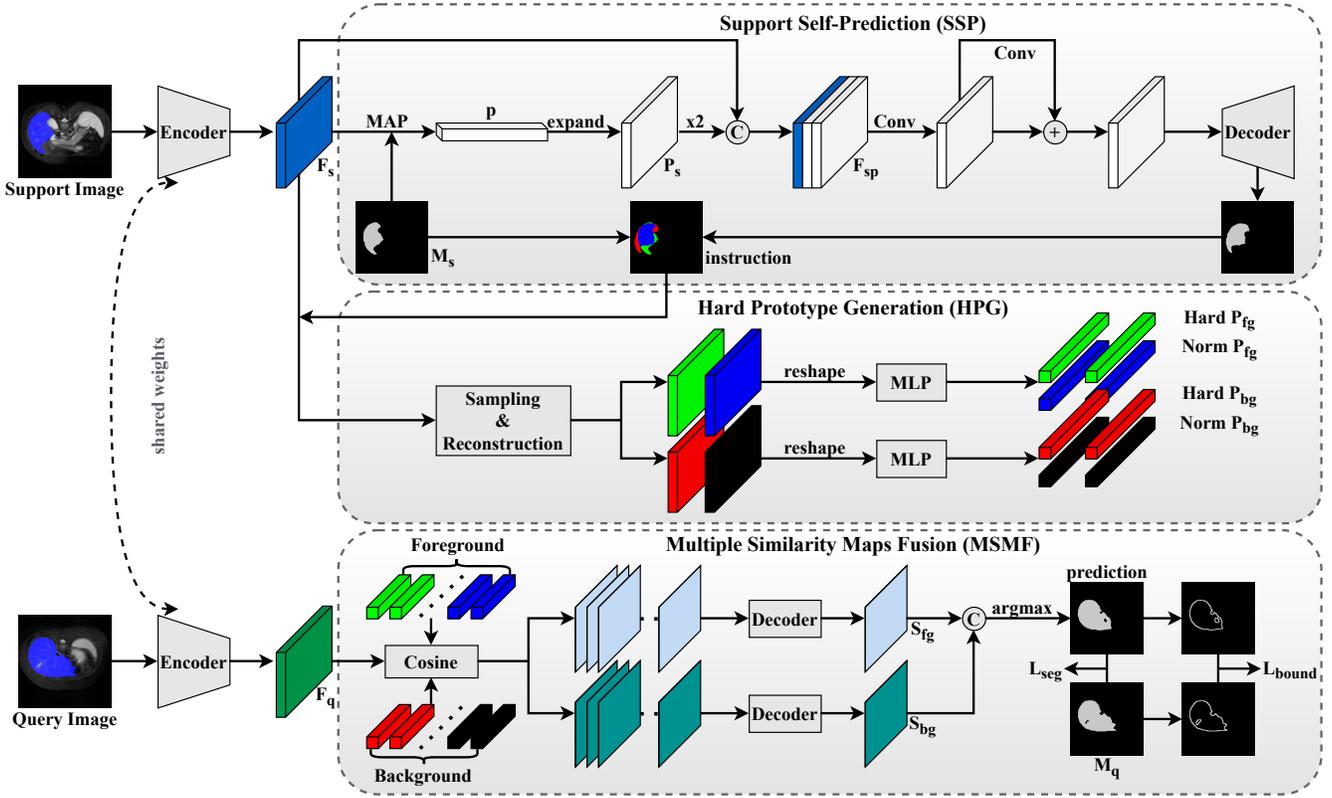
Figure 2: Illustration of CoW. We employ a shared feature encoder to learn deep features for both support and query images. In each episode, we firstly conduct SSP with support prototype **p** obtained via MAP. Next we compare this prediction with the ground truth M_s to acquire the instruction for hard and normal feature points. Then we perform random sampling and reconstruct support feature map, which is used for generating multiple hard and normal prototypes for foreground and background, respectively. At last, these prototypes are applied to generate multiple prediction similarity maps that are fused by a lightweight decoder to get final segmentation mask.

set are both divided into a support set and a query set (e.g., $D_{train} = \{(S_i, Q_i)\}_{i=1}^{N_{train}}$ and $D_{test} = \{(S_i, Q_i)\}_{i=1}^{N_{test}}$), where N_{train} and N_{test} are the number of episodes for training and testing respectively. When considering a N-way K-shot task, the support set $S_i = \{(I_s^k, M_s^k(c_j))\}_{k=1}^{K}$ contains K image-mask pairs of an image $I \in \mathbb{R}^{H \times W}$ and its corresponding binary mask $M \in \{0,1\}^{H \times W}$ for class $c_j \in C_{base}, j = 1, 2, ..., N$ and the query set $Q_i = \{(I_q^k, M_q^k(c_j))\}_{k=1}^{N_{qry}}$ contains N_{qry} image-mask pairs from the same class as the support set. Notably, the background class denoted as c_0 does not count toward C_{base} or C_{novel}. Given a support image-mask pair and a query image as input, the model for FSS task outputs predicted segmentation mask $\widetilde{M}_q$ for the query image.

3.2 Network Overview

The proposed CoW is a dual-path architecture where similar operations are performed, one for segmenting foreground regions and the other for background. The framework comprises following primary components: (1) a generic feature encoder shared by the support and query images for extracting features; (2) an **SSP** module for acquiring a coarse predicted mask of the support image using global support prototype; (3) an **HPG** module for generating a specified number

of hard prototypes for both foreground and background; (4) an **MSMF** module for fusing multiple similarity maps to get final segmenting mask of the query image. Figure 2 shows a more detailed pipeline of our method.

3.3 Support Self-Prediction

As shown in Figure 2, our proposed CoW starts by a parameter-sharing feature encoder denoted as f_θ, which extracts features from both support and query images. We denote the support and the query features as $F_s = f_\theta(I_s)$ and $F_q = f_\theta(I_q)$, $F_s, F_q \in \mathbb{R}^{C' \times H' \times W'}$, where C' is the channel depth of the feature, and H' and W' denote the height and width of the feature respectively. Following the practice in Q-Net [Shen *et al.*, 2023], we adopt ResNet-101 pre-trained on MS-COCO [Lin *et al.*, 2014] as the backbone, and utilize the output feature of the third residual block.

In order to find out the weak feature points easily misclassified, which lead to over-segmentation and boundary-blurring, we devise a Support Self-Prediction (SSP) module to provide clues for them. Specifically, we firstly employ masked average pooling to generate a global support prototype $\mathbf{p}_s \in \mathbb{R}^{1 \times C'}$ from the support feature map F_s:

$$\mathbf{p}_s = \frac{\sum_{h,w} F_s(h,w) \odot M_s(h,w)}{\sum_{h,w} M_s(h,w)}, \quad (1)$$

where F_s is resized to the size of corresponding mask M_s. $\odot$ denotes the Hadamard product, and (h, w) is the pixel position. Then we duplicate and expand the support prototype $\mathbf{p}_s$ to the same size as F_s, represented as F_p, which is subsequently concatenated with F_s to generate a new feature map:

$$F_{sp} = Concat([F_s, F_p, F_p]), \tag{2}$$

where $Concat(\cdot)$ denotes the concatenation operator.

Next, the similarity map for the support image is generated after passing through the feature processing module and the decoder that adopts single-scale residual layers following an ASPP module:

$$S_{support} = \mathcal{D}(FPM(F_{sp}); \psi), \tag{3}$$

where $S_{support} \in \mathbb{R}^{1 \times H' \times W'}$ is the predicted support similarity map. $FPM(\cdot)$ is the feature processing operation and $\mathcal{D}(\cdot; \psi)$ means the support decoder parameterized with ψ. After that, we generate the predicted support mask $\hat{M}_s$ through a concatenation operation followed by an argmax operation:

$$\hat{M}_s = argmax(Concat([1 - S_{support}, S_{support}])), \tag{4}$$

where $S_{support}$ is upsampled to (H, W) in advance and $1 - S_{support}$ represents the background similarity map.

For the purpose of finding weaker feature points in the support feature map, we adopt a cross-entropy loss on SSP module, which is written as:

$$\mathcal{L}_{ssp} = -\frac{1}{HW} \sum_{h,w} M_s^{bg} \log(\hat{M}_s^{bg}) + M_s^{fg} \log(\hat{M}_s^{fg}), \tag{5}$$

where $\hat{M}_s^{bg} = 1 - \hat{M}_s^{fg}$, and the superscript fg and bg is the indicator for the foreground and background respectively.

Finally, we calculate the position of weak feature points and normal ones borrowing information from both predicted support mask $\hat{M}_s$ and support ground truth M_s. This process can be mathematically expressed as:

$$\begin{cases} M_s^{hf} = [M_s(h, w) = 1]\&[\hat{M}_s(h, w) = 0], \\ M_s^{hb} = [M_s(h, w) = 0]\&[\hat{M}_s(h, w) = 1], \\ M_s^{nf} = [M_s(h, w) = 1]\&[\hat{M}_s(h, w) = 1], \\ M_s^{nb} = [M_s(h, w) = 0]\&[\hat{M}_s(h, w) = 0], \end{cases} \tag{6}$$

where the superscript hf, hb, nf, nb refers to hard foreground, hard background, normal foreground and normal background respectively.

3.4 Hard Prototypes Generation

As opposed to previous works that employ multiple normal prototypes, which inevitably loses part of the detailed information and inadequately represents the overall feature distribution within a class, we devise a novel Hard Prototypes Generation (HPG) module to generate multiple hard prototypes for both foreground and background to guarantee full use of the support information. We will take the example of generating hard foreground prototypes as described below.

Building on the weak feature points instruction provided by the SSP module, we firstly remove all support feature points except for those belonging to hard foreground, which can be described as:

$$F_s^{hf} = F_s \odot M_s^{hf}, \tag{7}$$

where $\odot$ denotes element-wise multiplication.

Next, we reshape the features of hard foreground region, denoted as $D_{hf} = [d_{hf}^1, d_{hf}^2, ..., d_{hf}^{N_{hf}}] \in \mathbb{R}^{N_{hf} \times C'}$. Then, a random sampling operation is performed on D_{hf} to extract some hard foreground points D'_{hf}. This process can be expressed as:

$$D'_{hf} = \mathcal{F}(D_{hf}, N), \tag{8}$$

where $\mathcal{F}(\cdot)$ is a random sampling function and N is the number of D'_{hf}.

Subsequently, we populate the hard foreground feature map F_s^{hf} with D'_{hf} where $M_s^{hf} = 0$ and retain original weak foreground feature points, which can be summarized mathematically in the following formulation:

$$F_s^{hf}(h, w) = \begin{cases} F_s^{hf}(h, w) & , M_s^{hf}(h, w) = 1, \\ D'_{hf} & , M_s^{hf}(h, w) = 0, \end{cases} \tag{9}$$

where (h, w) denotes the spatial position of the pixel.

After that, we acquire a reconstructed hard foreground feature map where all feature points exclusively pertain to the hard foreground. In order to generate multiple hard foreground prototypes that fully capture the feature information, we pass the reconstructed F_s^{hf} through a lightweight MLP which is composed of a fully connected layer, a ReLU activation layer, and another fully connected layer. This process can be denoted as:

$$\mathbf{P}_{hf} = MLP(F_s^{hf}; \phi), \tag{10}$$

where $MLP(\cdot; \phi)$ represents the MLP module parameterized with ϕ. The hard foreground feature map is reshaped to the dimension of $HW \times C'$ before being fed into the MLP and $\mathbf{P}_{hf} \in \mathbb{R}^{N_{hf} \times C'}$ denotes generated N_{hf} hard foreground prototypes.

Notably, the aforementioned procedure for generating hard prototypes also applies to hard background, normal foreground and background, namely $\mathbf{P}_{hb}, \mathbf{P}_{nf}, \mathbf{P}_{nb}$. Eventually, we concatenate the generated hard prototypes with normal ones and corresponding global support prototype for foreground and background respectively, which can be formulated as:

$$\begin{cases} \mathbf{P}_{fg} = \mathbf{P}_{hf} \oplus \mathbf{P}_{nf} \oplus \mathbf{p}_s^{fg}, \\ \mathbf{P}_{bg} = \mathbf{P}_{hb} \oplus \mathbf{P}_{nb} \oplus \mathbf{p}_s^{bg}, \end{cases} \tag{11}$$

where $\oplus$ denotes the concatenation operation. $\mathbf{p}_s^{fg}$ and $\mathbf{p}_s^{bg}$ are support foreground and background prototypes generated via MAP. $\mathbf{P}_{fg} \in \mathbb{R}^{N_{fg} \times C'}$ and $\mathbf{P}_{bg} \in \mathbb{R}^{N_{bg} \times C'}$ are final representative prototypes used for predicting the foreground and background similarity maps of the query image respectively.

3.5 Multiple Similarity Maps Fusion

In previous FSMIS tasks, researchers commonly leverage a single prototype or multiple normal prototypes to make predictions for the query image directly based on cosine-similarity. However, there exist extreme imbalance between

foreground and background in medical image and lost information in the masked feature, which easily lead to misclassified foreground and background points. To tackle this challenge, we design a Multiple Similarity Maps Fusion (MSMF) module, which fuses similarity maps generated by multiple prototypes obtained from HPG module, thus making more accurate segmentation.

First of all, we calculate the similarity maps between the multiple foreground and background prototypes, and the query image respectively:

$$
\begin{cases}
S_{fg}(h,w) = \dfrac{\mathbf{P}_{fg}^T \cdot F_q(h,w)}{\|\mathbf{P}_{fg}\| \cdot \|F_q(h,w)\|}, \\
S_{bg}(h,w) = \dfrac{\mathbf{P}_{bg}^T \cdot F_q(h,w)}{\|\mathbf{P}_{bg}\| \cdot \|F_q(h,w)\|},
\end{cases}
\tag{12}
$$

where $S_{fg} \in \mathbb{R}^{N_{fg} \times H' \times W'}$ and $S_{bg} \in \mathbb{R}^{N_{bg} \times H' \times W'}$ denote the generated similarity maps for foreground and background of the query respectively.

Then, both S_{fg} and S_{bg} will be fed into a lightweight decoder to fuse multiple similarity maps, thereby generating the foreground prediction map and background prediction map. This process can be denoted as:

$$
\begin{cases}
\hat{M}_q^{fg} = \mathcal{D}_1(S_{fg}; \psi_1), \\
\hat{M}_q^{bg} = \mathcal{D}_2(S_{bg}; \psi_2),
\end{cases}
\tag{13}
$$

where $\mathcal{D}_1(\cdot; \psi_1)$ and $\mathcal{D}_2(\cdot; \psi_2)$ denote a straightforward decoder based on convolutional neural network with parameter ψ_1 and ψ_2. $\hat{M}_q^{fg} \in \mathbb{R}^{1 \times H' \times W'}$ and $\hat{M}_q^{bg} \in \mathbb{R}^{1 \times H' \times W'}$ are the final single predicted similarity map for query foreground and background respectively.

Finally, we upsample $\hat{M}_q^{fg}$ and $\hat{M}_q^{bg}$ to the size of (H, W) and integrate them by utilizing the concatenation operation and a softmax layer to get the final prediction, as follows:

$$
\hat{M}_q = softmax(\hat{M}_q^{bg} \oplus \hat{M}_q^{fg}),
\tag{14}
$$

where $softmax(\cdot)$ executes along the channel dimension and $\hat{M}_q \in \mathbb{R}^{2 \times H \times W}$ will eventually be processed by an $argmax$ operation to get the final segmentation mask of the query image.

3.6 Loss Function

In order to quantify the dissimilarity between the predicted and ground truth, we employ the binary cross-entropy loss as the segmentation loss for each training episode:

$$
\mathcal{L}_{seg} = -\frac{1}{HW} \sum_{h,w} M_q^{bg} \log(\hat{M}_q^{bg}) + M_q^{fg} \log(\hat{M}_q^{fg}),
\tag{15}
$$

where $\mathcal{L}_{seg}$ works as the primary loss of our proposed CoW.

Following common practice [Wang et al., 2019; Ouyang et al., 2020], we inversely predict labels of the support images by using query images as the support set, then build a prototype alignment regularization loss, which is calculated as follows:

$$
\mathcal{L}_{align} = -\frac{1}{HW} \sum_{h,w} M_s^{bg} \log(\hat{M}_s^{bg}) + M_s^{fg} \log(\hat{M}_s^{fg}).
\tag{16}
$$

Additionally, we employ the intra-class loss to help diminish intra-class variation of the support and query and the inter-class loss to improve discrimination among different classes, both enhancing the representativeness of our generated hard and normal prototypes for foreground and background, which is formulated as:

$$
\mathcal{L}_{intra} = 2 - \sum_{* \in \{bg,fg\}} \frac{1}{N_*} \sum_{k=1}^{N_*} \max_{\mathbf{p}_k^s \in \mathbf{P}_*^s} (\cos(\mathbf{p}_k^s, \mathbf{P}_*^q)),
\tag{17}
$$

$$
\mathcal{L}_{inter} = \sum_{* \in \{s,q\}} \frac{1}{N_{fg} N_{bg}} \sum_{\mathbf{p}_i^* \in \mathbf{P}_{fg}^*} \sum_{\mathbf{p}_j^* \in \mathbf{P}_{bg}^*} \cos(\mathbf{p}_i^*, \mathbf{p}_j^*),
\tag{18}
$$

where the superscripts s and q denote the support and query image, $\cos$ is the cosine similarity, and $\max(\cdot)$ means taking the maximum.

Furthermore, to get a more precise predicted mask, we devise a boundary loss to refine the segmentation boundary. Specifically, we generate the boundary map of ground truth and predicted mask through a $maxpool$ operation, denoted as M_q^b and $\hat{M}_q^b$. Then, we calculate f1-score between them and get final boundary loss $\mathcal{L}_{bound}$ as follows:

$$
\mathcal{L}_{bound} = 1 - BF1(M_q^b, \hat{M}_q^b),
\tag{19}
$$

where $BF1(\cdot, \cdot)$ computes the f1-score between input pair.

Overall, the loss function for each training episode is defined to be $\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{align} + \lambda_0 \mathcal{L}_{bound} + \lambda_1(\mathcal{L}_{intra} + \mathcal{L}_{inter} + \mathcal{L}_{ssp})$, where λ_0, λ_1 are the balancing coefficients.

4 Experiment

4.1 Datasets, Settings and Evaluation Metrics

Datasets. To demonstrate the superiority of our method under different scenarios, we perform evaluations on three representative publicly available datasets: abdominal organs segmentation for MRI and CT (Abd-MRI [Kavur et al., 2021] and Abd-CT [Landman et al., 2015]) and cardiac segmentation for MRI (CMR [Zhuang, 2018]). The detailed introduction and split strategy is depicted in [Zhu et al., 2023].

Settings. The majority of FSMIS models follow the settings in the work [Ouyang et al., 2022] to evaluate their performance. In order to make a fair comparison, we also adopt the two experimental settings outlined in SSL-ALPNet [Ouyang et al., 2020] and ADNet [Hansen et al., 2022].

Evaluation Metrics. Following the common practice [Roy et al., 2020] in FSMIS, we employ the mean Sorensen-Dice coefficient (DSC) to compare the prediction to the ground truth. It measures the effectiveness of the same area between the predicted mask and the ground-truth.

4.2 Implementation Details

Following the practice in SSL-ALPNet [Ouyang et al., 2020], the experiments are conducted based on a 1-way 1-shot setting with 5-fold cross-validation. Initially, pseudo labels are generated for training by leveraging 3D superpixel clustering. After that, we reshape the size of the 2D slice from the 3D scanned image to 256×256 for Abd-MRI and CMR and 257×257 for Abd-CT. The total number of generated hard

Setting	Method	Reference	Abd-MRI					Abd-CT				
			Liver	RK	LK	Spleen	Mean	Spleen	RK	LK	Liver	Mean
1	PA-Net	CVPR'19	50.40	32.19	30.99	40.58	38.53	36.04	21.19	20.67	49.55	31.86
	SSL-ALPNet	ECCV'20	76.10	85.18	81.92	72.18	78.84	70.96	71.81	72.36	78.29	73.35
	SR&CL	MICCAI'22	80.23	87.42	79.34	76.01	80.77	73.41	71.22	73.45	76.06	73.53
	Q-Net	IntelliSys'23	81.74	87.98	78.36	75.99	81.02	77.81	70.98	73.83	76.37	74.75
	CAT-Net	MICCAI'23	75.02	83.23	75.31	67.31	75.22	66.02	64.56	68.82	80.51	70.88
	RPT	MICCAI'23	**82.86**	89.82	80.72	76.37	82.44	79.13	72.58	77.05	**82.57**	77.83
	GMRD	TMI'24	81.42	90.12	83.96	76.09	82.90	78.31	74.46	81.70	79.60	78.52
	DSPNet	MIA'24	75.06	85.37	81.88	70.93	78.31	69.31	74.54	78.01	69.32	72.79
	MSFSeg	MICCAI'24	76.11	88.10	84.18	77.12	81.38	73.64	78.41	81.11	78.91	78.02
	Ours (CoW)	–	81.55	**90.27**	**85.58**	**79.00**	**84.10**	**83.56**	**80.46**	**83.60**	82.35	**82.49**
2	PA-Net	CVPR'19	42.26	38.64	53.45	50.90	46.33	29.59	17.37	32.34	38.42	29.43
	SSL-ALPNet	ECCV'20	73.05	78.39	73.63	67.02	73.02	60.25	54.82	63.34	73.65	63.02
	SR&CL	MICCAI'22	75.55	84.24	77.07	73.73	77.65	67.36	63.37	67.39	73.63	67.94
	Q-Net	IntelliSys'23	78.25	65.94	64.81	65.37	68.59	68.10	61.12	63.32	80.49	68.26
	CAT-Net	MICCAI'23	78.98	78.90	74.01	68.83	75.18	67.65	60.05	63.36	75.31	66.59
	RPT	MICCAI'23	76.37	86.01	78.33	75.46	79.04	70.80	67.73	72.99	75.24	71.69
	GMRD	TMI'24	80.25	86.66	78.65	73.25	79.70	75.30	76.17	77.40	80.39	77.32
	DSPNet	MIA'24	78.56	82.01	76.47	68.27	76.33	66.48	63.55	68.46	69.16	66.17
	MSFSeg	MICCAI'24	76.14	86.98	**82.83**	78.07	81.01	75.21	77.36	79.24	76.73	77.14
	Ours (CoW)	–	**80.46**	**89.15**	82.40	74.55	**81.64**	**79.92**	**79.44**	**83.36**	**82.26**	**81.25**

Table 1: Quantitative comparison (in DSC score %) of different methods under Setting 1 and Setting 2 on Abd-MRI and Abd-CT datasets. Bold and underlined numbers denote the best and second best results, respectively.

Method	Reference	CMR			
		LV-MYO	LV-BP	RV	Mean
PA-Net	CVPR'19	25.18	58.04	12.86	32.02
SSL-ALPNet	ECCV'20	66.74	83.99	79.96	76.90
SR&CL	MICCAI'22	65.83	84.74	78.41	76.32
Q-Net	IntelliSys'23	65.92	90.25	78.19	78.15
CAT-Net	MICCAI'23	66.85	**90.54**	79.71	79.03
RPT	MICCAI'23	66.91	89.90	80.78	79.19
GMRD	TMI'24	67.04	90.00	80.29	79.11
DSPNet	MIA'24	64.91	87.75	79.73	77.46
Ours (CoW)	–	**68.41**	89.25	**82.33**	**80.00**

Table 2: Quantitative comparison (in DSC score %) of different methods under Setting 1 on CMR dataset. Bold and underlined numbers denote the best and second best results, respectively.

and normal prototypes for foreground and background are set to 100 and 600 respectively. We set the initial learning rate to 0.001, the decay rate to 0.9, and the total number of required iterations to $50K$, containing 5000 iterations each epoch. The balancing coefficients λ_0 and λ_1 are set to 0.5 and 0.3 respectively in our experiment.

4.3 Comparison with SOTA Methods

For the purpose of convincingly demonstrating the superiority and effectiveness of our proposed CoW, we compare our method with several existing state-of-the-art methods, including PA-Net [Wang *et al.*, 2019], SSL-ALPNet [Ouyang *et al.*, 2020], SR&CL [Wang *et al.*, 2022], CAT-Net [Lin *et al.*, 2023], Q-Net [Shen *et al.*, 2023], RPT [Zhu *et al.*, 2023], GMRD [Cheng *et al.*, 2024], DSPNet [Tang *et al.*, 2024] and MSFSeg [Zheng *et al.*, 2024], from the perspective of segmentation performance under two experimental settings. In this paper, considering that most researchers commonly adopt the PA-Net [Wang *et al.*, 2019] as the baseline and the settings in SSL-ALPNet [Ouyang *et al.*, 2020] to evaluate their performance, we also follow them to ensure a fair comparison.

Baseline	SSP	HPG	MSMF	Abd-MRI				
				Liver	RK	LK	Spleen	Mean
✓	✗	✗	✗	50.40	32.19	30.99	40.58	38.54
✓	✓	✓	✗	80.29	88.57	81.38	75.52	81.44
✓	✓	✓	✓	**81.55**	**90.27**	**85.58**	**79.00**	**84.10**

Table 3: Ablation study on Abd-MRI dataset under Setting 1 for the effect of each component on the performance of CoW in terms of the DICE score.

$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{bound}$	$\mathcal{L}_{ssp}$	Abd-MRI				
				Liver	RK	LK	Spleen	Mean
✗	✗	✗	✗	80.77	89.12	77.68	73.57	80.29
✓	✗	✗	✗	80.88	89.91	78.22	75.13	81.04
✓	✓	✗	✗	80.91	89.95	79.62	75.74	81.56
✓	✓	✓	✗	81.16	90.22	83.91	78.27	83.39
✓	✓	✓	✓	**81.55**	**90.27**	**85.58**	**79.00**	**84.10**

Table 4: Ablation study on Abd-MRI dataset under Setting 1 for the effect of auxiliary losses on the performance of CoW.

As shown in Table 1 and Table 2, our CoW significantly outperforms all listed methods in terms of the mean dice score on three different datasets.

4.4 Ablation Studies

Effect of each component. To evaluate separate contributions of each component within our model, we present their positive contributions in Table 3. Our SSP module and HPG module significantly improve its performance on top of the baseline model [Wang *et al.*, 2019], achieving a mean dice score of 81.44%, which is attributed to the combination of normal and hard prototypes carrying rich subject and boundary information. In addition, the MSMF module with a dual-path architecture that aims at fusing multiple similarity maps further improves the performance by 2.66%.

N_{hf}	N_{nf}	Abd-MRI				
		Liver	RK	LK	Spleen	Mean
0	100	81.49	89.22	84.09	77.44	83.06
25	75	81.55	90.03	84.39	78.66	83.66
50	50	81.55	**90.27**	**85.58**	**79.00**	**84.10**
75	25	81.22	89.85	82.99	77.83	82.97
100	0	**82.04**	90.15	83.75	72.43	82.09

Table 5: Ablation study on Abd-MRI dataset under Setting 1 for the effect of number of $\mathbf{P}_{hf}$ and $\mathbf{P}_{nf}$ on the performance of CoW, where $N_{hf} + N_{nf} = 100$, $N_{hb} = 100$ and $N_{nb} = 500$.

Method	Abd-CT					
	Fold0	Fold1	Fold2	Fold3	Fold4	Mean
SSL-ALPNet	75.37	65.65	68.68	77.68	79.37	73.35
PMM	74.70	63.81	69.33	78.26	76.15	72.45
K-means	77.06	67.02	75.42	79.76	80.79	76.01
GMRD	79.30	72.62	76.92	80.20	83.55	78.52
Ours (CoW)	**83.46**	**77.72**	**81.14**	**84.32**	**85.80**	**82.49**

Table 6: Ablation study on Abd-CT dataset under Setting 1 about different methods for generating multiple prototypes.

Effect of each loss. As can be seen from the last column in Table 4, each loss indeed makes a separate contribution to the final performance. $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ achieve a total gain of 1.27% by improving discriminability of generated multiple prototypes. Particularly, $\mathcal{L}_{bound}$ gives a 1.83% increase due to its constraint on the segmentation boundary. Moreover, $\mathcal{L}_{ssp}$ brings an additional gain of 0.71% for the reason that we can acquire more robust normal and hard prototypes making support self-prediction more precise.

Number of normal and hard foreground prototypes. We test the effect of the number for $\mathbf{P}_{hf}$ and $\mathbf{P}_{nf}$ on the final performance of CoW. As shown in Table 5, the optimal number setting is half for each. By contrast, employing $\mathbf{P}_{hf}$ or $\mathbf{P}_{nf}$ alone leads to a performance degradation of 2.01% and 1.04% respectively because they solely leverage singular subject or boundary information, thus causing a less comprehensive representation for each category.

Generating multiple prototypes. We compare our method with several previous works [Ouyang *et al.*, 2020; Yang *et al.*, 2020; Hartigan and Wong, 1979; Cheng *et al.*, 2024] in Table 6. It is apparent that our CoW achieves the best performance, surpassing the suboptimal GMRD by 3.97%, for the reason of focusing more attention to weak boundary details.

4.5 Visualization

Qualitative Results. To better demonstrate the superiority of our method, qualitative results of our model and other models, including Q-Net [Shen *et al.*, 2023] and GMRD [Cheng *et al.*, 2024], on Abd-MRI are shown in Fig. 3 (More qualitative results on the CMR dataset and the Abd-CT dataset can be found in Appendix).

Distribution of Hard and Normal Prototypes. As depicted in Fig. 4, we visualize the generated hard and normal prototypes with t-SNE [Van der Maaten and Hinton, 2008]. It can be seen that normal foreground or background prototypes
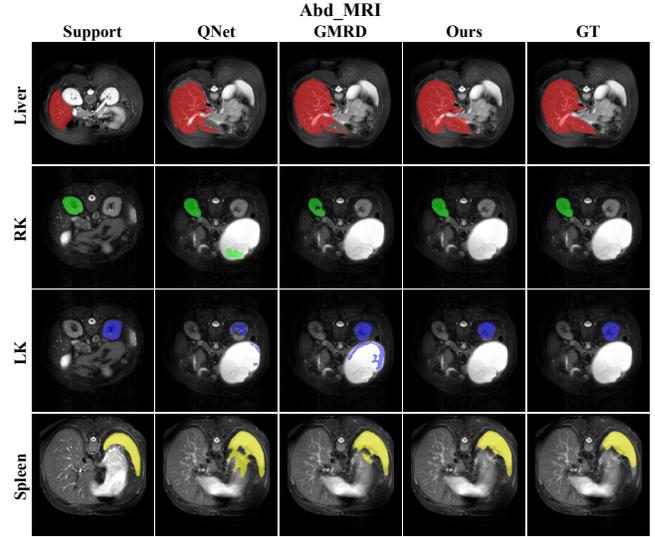

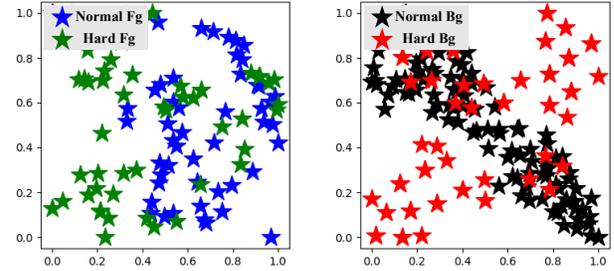Figure 3: Comparison of qualitative results between our method and other on Abd-MRI.

Figure 4: t-SNE visualization for generated normal and hard prototypes. 'Fg' implies foreground and 'Bg' implies background.

alone fail to well represent the entire distribution of their corresponding class. In contrast, the introduction of hard prototypes can compensate for such deficiency.

5 Conclusion

In this paper, we have proposed to mine hard prototypes by concentrating on weak features for few-shot medical image segmentation. It alleviates the challenge of current FSMIS methods that generate multiple prototypes solely through random sampling and local averaging, often failing to give a complete representation of corresponding category distribution. Additionally, our model includes a powerful edge constraint, which can further ensure a clearer segmentation boundary. Numerous experimental results also demonstrate that our method can capture more comprehensive class distribution and precisely segment the intricate organs in medical images. We believe that our model has the potential to assist in the diagnosis of rare disease and improve the efficiency of medical treatment.

Acknowledgments

References

[Cao *et al.*, 2022] 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*, pages 205–218, 2022.

[Chen *et al.*, 2018] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Drinet for medical image segmentation. *IEEE Transactions on Medical Imaging*, 37(11):2453–2462, 2018.

[Chen *et al.*, 2021] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.

[Cheng *et al.*, 2024] 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*, 2024.

[Dolz *et al.*, 2018] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.

[Dong and Xing, 2018] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, pages 1–11, 2018.

[El Naqa *et al.*, 2007] Issam El Naqa, Deshan Yang, Aditya Apte, Divya Khullar, Sasa Mutic, Jie Zheng, Jeffrey D Bradley, Perry Grigsby, and Joseph O Deasy. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning a. *Medical physics*, 34(12):4738–4749, 2007.

[Fang and Yan, 2020] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[Garcia and Bruna, 2017] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

[Gu *et al.*, 2020] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711, 2020.

[Hansen *et al.*, 2022] Stine Hansen, Srishti Gautam, Robert Jenssen, and Michael Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[Hesamian *et al.*, 2019] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, 32:582–596, 2019.

[Isensee *et al.*, 2021] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[Kavur *et al.*, 2021] 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.

[Landman *et al.*, 2015] 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, page 12, 2015.

[Li *et al.*, 2018] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[Lin *et al.*, 2014] 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*, pages 740–755, 2014.

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

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Luo *et al.*, 2021] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image seg-

mentation through dual-task consistency. In *AAAI*, pages 8801–8809, 2021.

[Ouyang *et al.*, 2020] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, pages 762–780, 2020.

[Ouyang *et al.*, 2022] 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.

[Peiris *et al.*, 2022] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *MICCAI*, pages 162–172, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[Roy *et al.*, 2020] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

[Shaban *et al.*, 2017] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[Shen *et al.*, 2023] Qianqian Shen, Yanan Li, Jiyong Jin, and Bin Liu. Q-net: Query-informed few-shot medical image segmentation. In *Proceedings of SAI Intelligent Systems Conference*, pages 610–628, 2023.

[Shi *et al.*, 2021] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, volume 30, 2017.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[Tang *et al.*, 2024] Song Tang, Shaxu Yan, Xiaozhi Qi, Jianxin Gao, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Few-shot medical image segmentation with high-fidelity prototypes. *Medical Image Analysis*, page 103412, 2024.

[Tsochatzidis *et al.*, 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.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3637–364, 2016.

[Wang *et al.*, 2018] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2018.

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

[Wang *et al.*, 2022] Runze Wang, Qin Zhou, and Guoyan Zheng. Few-shot medical image segmentation regularized with self-reference and contrastive learning. In *MICCAI*, pages 514–523, 2022.

[Wu *et al.*, 2019] Yicheng Wu, Yong Xia, Yang Song, Donghao Zhang, Dongnan Liu, Chaoyi Zhang, and Weidong Cai. Vessel-net: Retinal vessel segmentation under multi-path supervision. In *MICCAI*, pages 264–272, 2019.

[Wu *et al.*, 2022] Huisi Wu, Fangyan Xiao, and Chongxin Liang. Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In *ECCV*, pages 417–434, 2022.

[Yang *et al.*, 2020] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *MICCAI*, pages 763–778, 2020.

[Zaidi and El Naqa, 2010] Habib Zaidi and Issam El Naqa. Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *European Journal of Nuclear Medicine and Molecular Imaging*, 37:2165–2187, 2010.

[Zheng *et al.*, 2024] Meng Zheng, Benjamin Planche, Zhongpai Gao, Terrence Chen, Richard J Radke, and Ziyan Wu. Few-shot 3d volumetric segmentation with multi-surrogate fusion. In *MICCAI*, pages 286–296, 2024.

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

[Zhuang, 2018] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2933–2946, 2018.