DU-Net: Dual-Level Uncertainty-Aware Few-Shot Medical Image Segmentation via Evidential Deep Learning

Lanrong Bian, Long Ying*
School of Computer Science, Nanjing University of Information Science and Technology
Nanjing 211800, China
202312490565@nuist.edu.cn, lying@nuist.edu.cn

Abstract

Few-shot Medical Image Segmentation (FSMIS) aims to achieve accurate segmentation with extremely limited labeled data. However, medical images are characterized by blurred boundaries, incomplete information, and high noise, leading to high uncertainty in model predictions. Most existing FSMIS methods, including recent descriptor-based state-of-the-art (SOTA) methods, generate segmentation results without explicitly modeling uncertainty at both the descriptor and prediction levels, which limits their reliability in complex regions. To address these problems, we propose DU-Net, a novel Dual-Level Uncertainty-Aware Network designed to achieve reliable segmentation. Evidence extraction networks are incorporated to estimate the uncertainty of foreground and background descriptors. To suppress the interference of high-uncertainty descriptors, the Descriptor-Level Uncertainty-Aware Similarity Maps Fusion (DUSMF) module dynamically modulates the contribution of each descriptor during fusion. To enhance the prediction accuracy, the Evidential Dual-Branch Prediction (EDP) module is designed, which enables joint optimization between segmentation and evidential uncertainty estimation in a multi-task learning fashion. We impose consistency constraints on the predictions of the above two branches in low-uncertainty regions to conduct mutual supervision. Experiments on three public medical image datasets demonstrate that our method outperforms current SOTA methods.

Keywords: Few-shot medical image segmentation, Evidential deep learning, Multiple representative descriptors, Uncertainty-aware.

1. Introduction

In recent years, deep learning [28] has achieved significant breakthroughs in medical image segmentation [33, 1], improving the accuracy of automatic segmentation [39, 37, 47, 6, 21]. Efficient training of these methods heavily relies on large amounts of accurately labeled data, which require time-consuming and labor-intensive manual annotation by experienced clinicians. Moreover, access to medical images is constrained by ethical, privacy and legal considerations, leading to a scarcity of high-quality labeled data that hinders model performance and generalization in clinical practice. To address this challenge, few-shot medical image segmentation (FSMIS) [45, 13, 40, 36] has been proposed, enabling accurate segmentation with a limited number of labeled samples.

Most existing FSMIS models [45, 35, 12, 46] employ masked average pooling (MAP) [30] to extract support features from support images, generating the corresponding prototypes. The segmentation of novel classes is then achieved by computing similarity between these prototypes and the query image, inevitably leading to the loss of class-discriminative information. Furthermore, most models construct only a single prototype for each class, making it difficult to fully capture the complex class distribution. To address these limitations, the GMRD [8] method generates multiple representative descriptors for both the foreground and background of each class, based on which foreground and background prediction maps are calculated. This strategy not only enables a more comprehensive characterization of class features and overall distribution but also mitigates the adverse effects caused by the imbalance between foreground and background proportions in medical images.

Due to the limited number of training samples in FSMIS and the inherent challenges of medical images, such as blurred tissue boundaries and diverse lesion morphologies, the representational capabilities of foreground and background descriptors are divergent, and the model is prone to producing unstable and low-confidence predictions. Specifically, 1) at the descriptor level, some descriptors can reliably capture the features of a class, while others exhibit high uncertainty, influenced by noise or highly similar anatomical structures. The descriptors with high uncertainty are difficult to accurately represent the categorical attributes and introduce noise in the calculation. 2) At the predic-

*Corresponding author.

tion level, complex anatomical structures and limited training samples lead to insufficient model generalization, which consequently has a negative impact on segmentation accuracy and prediction reliability, particularly when the model encounters unseen classes or morphologies in query images.

Although current studies [20] have begun to leverage uncertainty to guide feature refinement, they do not elaborately model and collaboratively process uncertainty at different levels. Recently, Evidential Deep Learning (EDL)[41, 34, 14] has attracted attention for its solid theoretical foundation and uncertainty quantification in a single forward pass. EDL has been successfully applied in medical image analysis, particularly in medical image classification [15, 16, 17, 9] and segmentation [57, 22, 29, 51, 7]. However, its application to FSMIS remains largely unexplored.

In this work, we propose a novel Dual-Level Uncertainty-Aware Network (DU-Net), which integrates EDL into the GMRD [8] method to model uncertainty at both the descriptor and prediction levels. Evidence extraction networks are constructed to estimate the uncertainty of the generated foreground and background descriptors. The Descriptor-Level Uncertainty-Aware Similarity Maps Fusion (DUSMF) module utilizes this uncertainty to enhance fine-grained semantic alignment and improve the stability of similarity estimation by dynamically adjusting the weights of similarity maps calculated between the representative descriptors and query features. Maps obtained through high-uncertainty descriptors are downweighted to suppress unreliable information, while those through low-uncertainty descriptors receive more attention for reliable and discriminative cues. To improve the segmentation accuracy and prediction interpretability, the Evidential Dual-Branch Prediction (EDP) module is designed, which enables joint optimization between segmentation and evidential uncertainty estimation in a multi-task learning fashion. The consistency constraints are imposed on the predictions of the previous two branches in low-uncertainty regions to conduct mutual supervision.

In summary, our main contributions are as follows:

1) A Dual-Level Uncertainty-Aware Network (DU-Net) is proposed to incorporate EDL to explicitly model uncertainty on both the descriptor and prediction levels.

2) At the descriptor level, the Descriptor-Level Uncertainty-Aware Similarity Maps Fusion (DUSMF) module explores the estimated uncertainty of the foreground and background descriptors to suppress the interference of high-uncertainty descriptors, enhancing the robustness of similarity calculation between descriptors and query features.

3) At the prediction level, the Evidential Dual-Branch Prediction (EDP) module integrates segmentation with evidential uncertainty estimation to jointly train. The consistency constraint loss is introduced to conduct mutual supervision between the two branches in low-uncertainty regions.

4) Extensive experiments and ablation studies on three popular medical image datasets demonstrate the effectiveness of our proposed method.

2. Related Work

2.1. Few-Shot Medical Image Segmentation

Although deep learning methods have made significant progress in medical image segmentation [2], their further advancement is limited by the scarcity of high-quality labeled data in practice. In recent years, FSMIS has gained increasing attention for its ability to achieve competitive performance with a limited number of labeled samples.

As one of the pioneers in FSMIS, SENet[40] achieves strong interactions between the conditioner and segmenter arms with a 'Squeeze & Excitation' module. SSL-ALPNet[35] proposed an adaptive local prototype pooling network and incorporated pseudo labels generated from superpixels as self-supervision signals. ADNet[18] adopts an anomaly detection perspective, computing an anomaly score for query pixels based on a single foreground prototype and predicting the mask with a learnable fixed threshold. Building on ADNet[18], Q-Net[43] adopts a dual-path feature extraction module to capture multi-scale features. It introduces query-informed threshold adaptation and prototype refinement modules to leverage query-specific information for segmentation optimization. To better characterize class distributions, GMRD [8] generates multiple representative descriptors and introduces a dual-path Multiple Affinity Maps-based Prediction (MAMP) module to fuse affinity maps from these descriptors. PSMNet[46] decomposes support masks into fine-grained submasks in a self-guided manner and introduces a Multi-Level Cross Attention Module (MCAM) to enhance query features by leveraging multi-level support information.

In FSMIS tasks, the scarcity of labeled samples often results in high prediction uncertainty. Recent studies have begun to focus on its negative impact on segmentation accuracy and stability. For instance, ADNet++[20] estimates voxel-level prediction uncertainty via masked randomized average pooling (MRAP) in a prototype-based framework, and leverages this uncertainty to guide superpixel feature refinement. In this work, we conduct evidential uncertainty estimation at both the descriptor and prediction levels within the FSMIS framework, enabling reliable and interpretable segmentation.

2.2. Uncertainty Estimation in Medical Image Analysis

Uncertainty modeling serves as the cornerstone for achieving reliable medical image analysis. With the continuous advancement of deep learning [28] in medical imaging, various uncertainty quantification paradigms have been

proposed, broadly categorized into probabilistic and non-probabilistic methods [23]. Probabilistic methods rely on probability distributions to quantify prediction uncertainty, typically using prediction entropy or variance to estimate output distributions. Representative methods include Bayesian inference[48, 49, 4], Monte Carlo dropout [26], and model ensembles[11]. Non-probabilistic methods do not require strong assumptions about the prior distribution of data and are suitable for scenarios where precise probabilistic information is unavailable. These methods mainly include interval analysis [38], fuzzy sets and fuzzy logic theory [53], and Dempster-Shafer theory[10, 42].

In recent years, EDL[41, 34, 14] has emerged as a mainstream approach for uncertainty modeling. Unlike traditional methods, which are characterized by computational complexity and high costs, EDL constructs a theoretical framework based on Dempster-Shafer Evidence Theory (DST) [52] and Subjective Logic (SL) [24]. Its core mechanism involves a neural network that outputs non-negative evidence for each class, where evidence reflects the degree of support for the corresponding class hypothesis. By parameterizing the Dirichlet distribution based on the evidence, EDL simultaneously obtains class prediction and uncertainty estimates during a single forward pass. This not only avoids computationally intensive operations such as repeated sampling but also enhances the interpretability and reliability of model predictions. Owing to its unique advantages in uncertainty quantification, EDL has been applied to several computer vision tasks[3, 5, 54, 44]. For example, UDEL[5] extends the traditional paradigm of EDL to adapt to the weakly-supervised multi-label classification goal by leveraging uncertainty at both video and snippet levels.

However, related research on FSMIS remains limited. Our work integrates EDL into the FSMIS method for uncertainty estimation at different levels.

3. Methodology

3.1. Problem Definition

The FSMIS aims to train a model on the training dataset $\mathcal{D}_{tr}$ with known classes $\mathcal{C}_{tr}$ to accurately segment novel classes $\mathcal{C}_{te}$ in the testing dataset $\mathcal{D}_{te}$, where $\mathcal{C}_{tr} \cap \mathcal{C}_{te} = \emptyset$. In this study, we adopt the commonly used meta-learning paradigm for FSMIS. $\mathcal{D}_{tr} = \{(S^i_{tr}, Q^i_{tr})\}^{N_{tr}}_{i=1}$ and $\mathcal{D}_{te} = \{(S^i_{te}, Q^i_{te})\}^{N_{te}}_{i=1}$ are constructed as several randomly sampled episodes, where N_{tr} and N_{te} denote the number of episodes in the training and testing phases, respectively. Each episode follows the N-Way K-shot setting, consisting of K annotated support images and several query images, all sampled from the same N classes. Specifically, during training, the support set is defined as $S^i_{tr} = \{(I^k_s, M^k_s)\}^K_{k=1}$, where I^k_s and M^k_s represent the k-th support image and its corresponding ground-truth segmentation mask, respectively. The query set $Q^i_{tr} = \{(I^k_q, M^k_q)\}^{N_q}_{k=1}$ contains N_q query image-mask pairs sampled from the same classes as the support set. The model leverages knowledge learned from support images to guide the segmentation of query images. During the testing phase, the model is evaluated under the same N-Way K-shot setting as used during training. In this experiment, we adopt GMRD [8] as the baseline and follow the experimental settings of SSL-ALPNet[35] and PANet [50], setting $N = K = 1$.

3.2. Overview

The proposed DU-Net is illustrated in Fig. 1. The support and query images are first input into a parameter-shared feature encoder to extract corresponding feature vectors. The support features and the support mask are then fed into the GMRD module [8] to generate multiple representative descriptors for the foreground and background classes. To facilitate evidence modeling, the evidence extraction network utilizes linear layers to project descriptor features into a lower-dimensional feature space and then estimate the uncertainty at the descriptor level. Similarity maps are calculated between the descriptors and query features, which are further aggregated through two lightweight decoders. Descriptor-level uncertainty is introduced as a weighting factor to dynamically adjust the contribution of each similarity map, reducing the influence of high-uncertainty descriptors. Finally, the foreground and background prediction maps obtained from the decoders are fed into two parallel branches. The segmentation branch employs the softmax function to produce pixel-wise prediction maps, while the evidential uncertainty estimation branch parameterizes a Dirichlet distribution for uncertainty quantification at the prediction level and evidential probability calculation. Consistency constraints are imposed on the predictions of the two branches to conduct mutual supervision in low-uncertainty regions.

3.3. Uncertainty Modeling via Evidential Deep Learning

To model the uncertainty at both the descriptor and prediction levels, we incorporate the EDL framework based on DST[52] and SL[24] into the proposed FSMIS model. For a segmentation task with C classes, given an input sample x, the network generates a non-negative evidence vector $\mathbf{e} = [e_1, \ldots, e_C]$ by applying the *softplus* activation function to the output of the final layer. To represent the intensity of activation of each class, evidence e_c is defined as a measure of the support collected from the data indicating the sample belongs to class c. Based on the evidence, SL[24] provides a belief mass b_c for each class $c = 1, \ldots, C$ and the overall uncertainty mass u, then we have:

$$u + \sum_{c=1}^{C} b_c = 1, \tag{1}$$

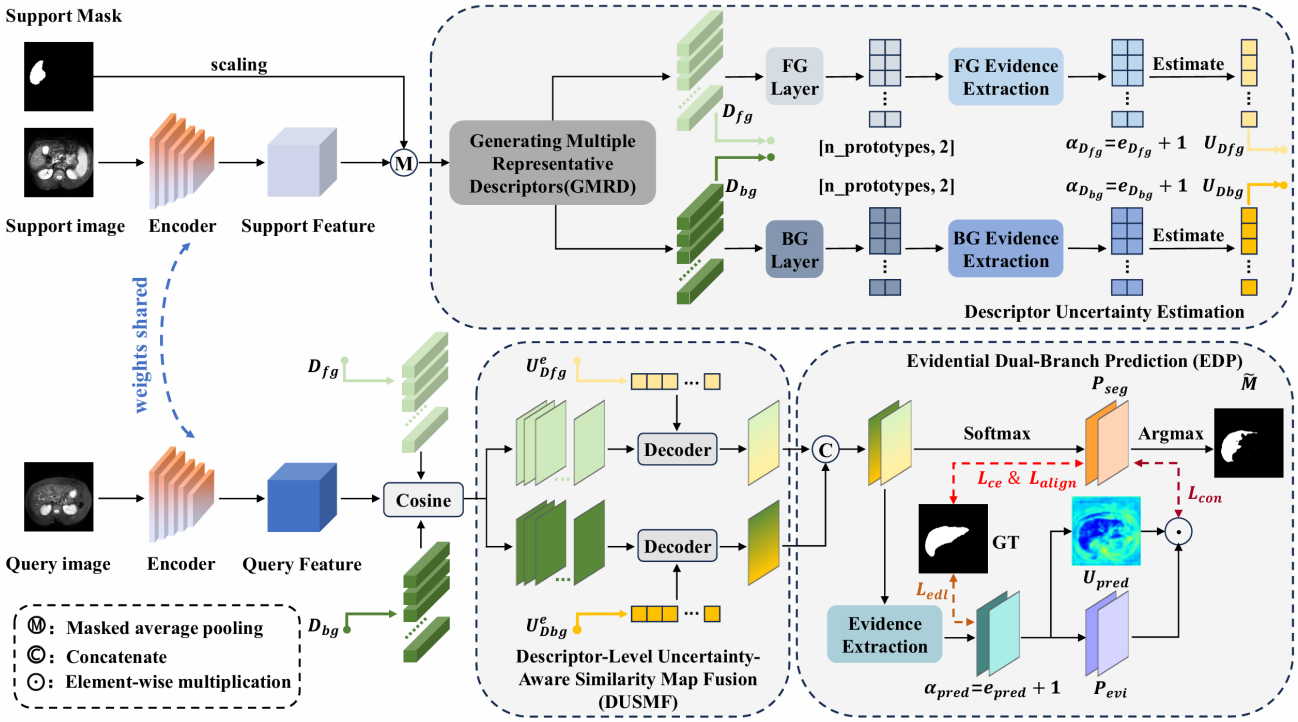

Figure 1: Overview of the proposed DU-net architecture. The shared-weight feature encoder extracts the support and query features. The support features and masks are then processed by the GMRD module to generate representative descriptors for the foreground and background classes. The uncertainty of the generated foreground and background descriptors is estimated through evidence extraction networks and fed into the DUSMF module to dynamically modulate the contributions of the similarity maps. Finally, the EDP module integrates segmentation with evidential uncertainty estimation to jointly train. The consistency constraint loss is introduced to conduct mutual supervision between the two branches in low-uncertainty regions.

where $u \geq 0$ and $b_c \geq 0$. Then, b_c and u can be calculated as follows:

$$b_c = \frac{e_c}{S}, \quad u = \frac{C}{S}, \tag{2}$$

where $S = \sum_{c=1}^{C}(e_c + 1)$. The total evidence $\sum_{c=1}^{C} e_c$ is inversely related to uncertainty: a larger total evidence indicates lower uncertainty, while the uncertainty reaches its maximum value $u = 1$ when there is no evidence for all classes. Building on the evidence collected from all classes, the Dirichlet distribution is parameterized by $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_C]$, where $\alpha_c = e_c + 1$. Consequently, the belief mass can also be derived from the Dirichlet parameters as $b_c = (\alpha_c - 1)/S$, where $S = \sum_{c=1}^{C} \alpha_c$ denotes the Dirichlet strength. The Dirichlet distribution density function is defined as:

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^{C} p_c^{\alpha_c - 1}, & \text{for } \mathbf{p} \in \mathcal{S}_C, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{S}_C = \{\mathbf{p} \mid \sum_{c=1}^{C} p_c = 1,\ 0 \leq p_c \leq 1\}$ is the C-dimensional unit simplex, $\mathbf{p} = [p_1, \ldots, p_C]$ is a random probability vector defined on $\mathcal{S}_C$ following the Dirichlet distribution, and $B(\boldsymbol{\alpha})$ denotes the C-dimensional Beta function with parameter $\boldsymbol{\alpha}$. Under the EDL framework, the predicted probability for class c is computed as the expected value of the corresponding Dirichlet distribution:

$$\hat{p}_c = \frac{\alpha_c}{S}. \tag{4}$$

To encourage the model to accumulate more informative evidence for the target class, a negative log-likelihood (NLL) loss is typically employed in the EDL framework. Specifically, treating the Dirichlet distribution $D(\mathbf{p} \mid \boldsymbol{\alpha})$ as a prior over the likelihood function $L(\mathbf{y} \mid \mathbf{p})$, the negative log of the marginal likelihood can be defined as:

$$\begin{aligned} \mathcal{L}_{\text{nll-edl}} &= -\log\Big(\int L(\mathbf{y} \mid \mathbf{p}) D(\mathbf{p} \mid \boldsymbol{\alpha})\, d\mathbf{p}\Big) \\ &= -\log\Big(\int \prod_{c=1}^{C} p_c^{y_c} \frac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^{C} p_c^{\alpha_c - 1}\, d\mathbf{p}\Big) \\ &= \sum_{c=1}^{C} y_c \big(\log(S) - \log(\alpha_c)\big), \end{aligned} \tag{5}$$

where $\mathbf{y}$ is the one-hot label vector. Furthermore, to prevent the model from making overconfident predictions on

samples with high uncertainty and to maintain high uncertainty when evidence is insufficient, EDL introduces an additional Kullback-Leibler (KL) divergence regularization term to suppress the generation of evidence for non-target classes. This regularization term is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{kl-edl}} &= \mathrm{KL}\big[D(\mathbf{p} \mid \tilde{\boldsymbol{\alpha}}) \,\|\, D(\mathbf{p} \mid \mathbf{1})\big] \\ &= \log \frac{\Gamma\Big(\sum_{c=1}^{C} \tilde{\alpha}_c\Big)}{\Gamma(C) \prod_{c=1}^{C} \Gamma(\tilde{\alpha}_c)} \\ &+ \sum_{c=1}^{C} (\tilde{\alpha}_c - 1)\Big[\psi(\tilde{\alpha}_j) - \psi\Big(\sum_{c=1}^{C} \tilde{\alpha}_c\Big)\Big],\end{aligned} \tag{6}$$

where $\Gamma(\cdot)$ denotes the gamma function and $\psi(\cdot)$ denotes the digamma function, $\tilde{\boldsymbol{\alpha}} = \mathbf{y} + (\mathbf{1} - \mathbf{y}) \odot \boldsymbol{\alpha}$ represents the Dirichlet parameters obtained after removing non-misleading evidence for the target class: the parameter of the target class is reset to 1 while retaining the original evidence for non-target classes. $D(\mathbf{p} \mid \mathbf{1})$ denotes a uniform Dirichlet distribution with all parameters equal to 1.

The final optimization objective of EDL is defined as:

$$\mathcal{L}_{\mathrm{edl}} = \mathcal{L}_{\mathrm{nll-edl}} + \mu_t \, \mathcal{L}_{\mathrm{kl-edl}}, \tag{7}$$

where $\mu_t = \min\left(1, \frac{t}{0.5T}\right)$ is a warm-up annealing coefficient used to avoid imposing excessive regularization on unreliable evidence predictions during early training stages, with t denoting the current epoch index and T the total number of epochs.

3.4. Quantification of Descriptor-Level Uncertainty

Similar to the baseline method [50], given a support image I_s and a query image I_q, we first extract the feature $F_s = f_\theta(I_s) \in \mathbb{R}^{H \times W \times F}$ and $F_q = f_\theta(I_q) \in \mathbb{R}^{H \times W \times F}$ using a feature encoder $f_\theta(\cdot)$ pre-trained on the MS-COCO dataset [31], where H and W denote the height and width of the feature map, respectively, and F represents the number of feature channels. The encoder adopts a ResNet [19] network structure, and its parameters θ are shared during the extraction of support and query features to ensure consistency in the feature space. The extracted support features together with the mask M_s are then fed into the GMRD [8] module to generate representative descriptors. This module employs a dual-path architecture, focusing separately on the feature representations of the foreground and background.

1) Generating multiple representative descriptors: We first perform morphological operations on the support mask to generate the inner-boundary mask M_s^{ib} and the outer-boundary mask M_s^{ob} for the foreground. Then, we conduct MAP on the support features by leveraging the foreground mask M_s, the background mask $1 - M_s$, and boundary masks to obtain the foreground prototype P_s^{fg}, the background prototype P_s^{bg}, and the boundary prototypes P_s^{ib}, P_s^{ob}. Subsequently, in the foreground and background paths, the support features are element-wise multiplied with the foreground and background masks to obtain purified features. For zero-valued regions in the features, a random sampling and populating strategy is employed to reconstruct features where all pixels belong to the respective class. Then, the reconstructed foreground and background feature are fed into two lightweight multi-layer perceptrons (MLP) to generate N_d^{fg} representative foreground descriptors $d_{\mathrm{fg}} \in \mathbb{R}^{N_d^{\mathrm{fg}} \times F}$ and N_d^{bg} background descriptors $d_{\mathrm{bg}} \in \mathbb{R}^{N_d^{\mathrm{bg}} \times F}$. Finally, the foreground descriptor set $D_{\mathrm{fg}} = d_{\mathrm{fg}} \oplus P_s^{\mathrm{fg}} \oplus P_s^{\mathrm{ib}}$, $D_{\mathrm{fg}} \in \mathbb{R}^{(N_d^{\mathrm{fg}}+2) \times F}$ is formed by concatenating the representative foreground descriptors with the foreground and inner-boundary prototypes, while the background descriptor set $D_{\mathrm{bg}} = d_{\mathrm{bg}} \oplus P_s^{\mathrm{bg}} \oplus P_s^{\mathrm{ob}}$, $D_{\mathrm{bg}} \in \mathbb{R}^{(N_d^{\mathrm{bg}}+2) \times F}$ is formed by concatenating the representative background descriptors with the background and outer-boundary prototypes. $\oplus$ denotes concatenation.

2) Uncertainty modeling at the descriptor level: To enable uncertainty-aware contribution modulation in similarity fusion, we adopt EDL to model descriptor-level uncertainty. As direct estimation in the high-dimensional descriptor space can be noisy and inefficient, two class-specific linear layers are introduced to project the high-dimensional descriptor features into a two-dimensional evidence space corresponding to the foreground and background classes. For $\forall d_{\mathrm{fg}}^i \in D_{\mathrm{fg}}$ and $\forall d_{\mathrm{bg}}^j \in D_{\mathrm{bg}}$, the projected two-dimensional vectors are calculated as

$$z_{\mathrm{fg}}^i = W_{\mathrm{fg}} d_{\mathrm{fg}}^i + b_{\mathrm{fg}}, \tag{8}$$

$$z_{\mathrm{bg}}^j = W_{\mathrm{bg}} d_{\mathrm{bg}}^j + b_{\mathrm{bg}}, \tag{9}$$

where $W_{\mathrm{fg}}, W_{\mathrm{bg}} \in \mathbb{R}^{2 \times F}$ are the weight matrices, $b_{\mathrm{fg}}, b_{\mathrm{bg}} \in \mathbb{R}^2$ are the bias vectors, and i, j denote the indices of the foreground and background descriptors, respectively. Subsequently, non-negative evidence at the descriptor level is generated through evidence extraction networks:

$$e_{D_{\mathrm{fg}}}^i = g(z_{\mathrm{fg}}^i) = \ln(1 + \exp(z_{\mathrm{fg}}^i)), \tag{10}$$

$$e_{D_{\mathrm{bg}}}^j = g(z_{\mathrm{bg}}^j) = \ln(1 + \exp(z_{\mathrm{bg}}^j)), \tag{11}$$

where $e_{D_{\mathrm{fg}}}^i, e_{D_{\mathrm{bg}}}^j \in \mathbb{R}^2$, and $g(\cdot)$ denotes the softplus activation function. Finally, the uncertainty of each foreground descriptor $u_{D_{\mathrm{fg}}}^i$ and each background descriptor $u_{D_{\mathrm{bg}}}^j$ can be calculated according to Eq. (2).

3.5. Descriptor-Level Uncertainty-Aware Similarity Maps Fusion

Existing FSMIS methods typically assume that all descriptors contribute equally when constructing query-support matching relationships, while overlooking the vari-

ability in descriptor-level uncertainty arising from imaging noise, annotation ambiguity, or anatomical variations, causing high-uncertainty descriptors to negatively impact segmentation accuracy. To address this, we propose the DUSMF module that dynamically adjusts the contribution weights of similarity maps generated from individual descriptors during fusion, prioritizing descriptors with higher confidence. The similarity between the representative descriptors and query features is computed as follows:

$$S_{\text{fg}} = \cos(F_q, D_{\text{fg}}), \tag{12}$$

$$S_{\text{bg}} = \cos(F_q, D_{\text{bg}}), \tag{13}$$

where $S_{\text{fg}} \in \mathbb{R}^{(N_d^{\text{fg}}+2)\times H\times W}$ and $S_{\text{bg}} \in \mathbb{R}^{(N_d^{\text{bg}}+2)\times H\times W}$ denote the foreground and background similarity maps, respectively. Subsequently, the foreground and background similarity maps are separately fed into two decoders for fusion, generating the foreground prediction map $\hat{P}_{\text{fg}}$ and background prediction map $\hat{P}_{\text{bg}}$. During the fusion of similarity maps, we introduce descriptor-level uncertainty as a modulation factor for the weights:

$$\hat{P}_{\text{fg}} = \text{Decoder}\big(S_{\text{fg}} \odot \phi(u_{D_{\text{fg}}})\big), \;\; \phi(u_{D_{\text{fg}}}) = \frac{1}{1+\delta_1\, u_{D_{\text{fg}}}}, \tag{14}$$

$$\hat{P}_{\text{bg}} = \text{Decoder}\big(S_{\text{bg}} \odot \phi(u_{D_{\text{bg}}})\big), \phi(u_{D_{\text{bg}}}) = \frac{1}{1+\delta_2\, u_{D_{\text{bg}}}}, \tag{15}$$

where $\text{Decoder}(\cdot)$ denotes a lightweight convolutional neural network, and $\odot$ denotes element-wise multiplication along the channel dimension. δ_1 and δ_2 are fixed scaling coefficients controlling the effect of descriptor-level uncertainty on fusion weights, both set to 0.1.

3.6. Evidential Dual-Branch Prediction

Most existing FSMIS methods directly output prediction results without modeling prediction uncertainty, compromising segmentation reliability. Therefore, an EDP module is designed to generate segmentation results and model uncertainty at the prediction level. This module takes the foreground and background prediction maps as input, concatenates them into a dual-channel tensor, and feeds it into two parallel branches: the segmentation branch and the evidential uncertainty estimation branch.

1) Segmentation branch: The segmentation branch aims to generate pixel-wise probability maps and is optimized with a supervised loss to ensure segmentation accuracy. First, we employ the softmax function along the channel dimension of the prediction maps to obtain the pixel-wise probability map P_{seg}, which consists of two channels corresponding to the foreground and background probability maps, denoted as $\tilde{P}_{\text{fg}}$ and $\tilde{P}_{\text{bg}}$, respectively. Furthermore, we adopt the cross-entropy loss to measure the discrepancy between the prediction and the ground-truth mask, defined as:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{HW}\sum_{h,w}\Big[(1-M_q)\log(\tilde{P}_{\text{bg}}) + M_q \log(\tilde{P}_{\text{fg}})\Big], \tag{16}$$

where M_q denotes the ground-truth mask of the query image. In addition, to enhance the model's generalization ability across images, we draw inspiration from prior research [50, 36] and introduce a prototype alignment regularization loss into the segmentation branch. The core idea is to combine the predicted segmentation mask with the query image to form a new support set, which is then used to inversely predict the segmentation mask of the support image. By substituting the predicted and ground-truth masks of the query image in Eq. (16) with the support image and ground-truth mask, we obtain the alignment loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{HW}\sum_{h,w}\Big[(1-M_s)\log(\bar{P}_{\text{bg}}) + M_s \log(\bar{P}_{\text{fg}})\Big], \tag{17}$$

where M_s denotes the ground-truth mask of the support image, and $\bar{P}_{\text{fg}}$ and $\bar{P}_{\text{bg}}$ represent the predicted foreground and background probability maps generated when the support image is treated as the query input. Finally, the segmentation mask $\tilde{M}$ is derived from the predicted probability maps via the argmax function:

$$\tilde{M} = \arg\max_{c\in\{\text{fg},\text{bg}\}} \tilde{P}_c, \tag{18}$$

where $\tilde{P}_c$ represents the pixel-wise probability map for class c.

2) Evidential uncertainty estimation branch: The evidential uncertainty estimation branch aims to parameterize a Dirichlet distribution under the EDL framework to explicitly model the uncertainty at the prediction level. Specifically, we apply the softplus activation function to the prediction maps to obtain non-negative evidence at the prediction level, denoted as e_{pred}. Subsequently, by adding a unit vector to the evidence vector, we derive the Dirichlet distribution parameters as $\alpha_{\text{pred}} = e_{\text{pred}} + 1$, which defines the Dirichlet distribution at the prediction level. Based on this distribution, the evidential probability map P_{evi} and the prediction-level uncertainty U_{pred} can be further estimated according to Eqs. (4) and (2). Finally, the evidential uncertainty estimation branch is optimized through the evidence loss defined in Eq. (7).

3) Collaborative optimization: To ensure consistent predictions between the segmentation branch and the evidential uncertainty estimation branch in low-uncertainty regions, a consistency constraint loss is constructed.

A dynamic threshold is first calculated according to the overall uncertainty, and a binary mask M_u is generated to

identify low-uncertainty regions:

$$\tau = \beta \,\mathrm{mean}(U_{\text{pred}}), \quad M_u = \mathbf{1}[U_{\text{pred}}(h,w) < \tau], \quad (19)$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, β is a fixed coefficient set to 1 that controls the sensitivity of the threshold, and $U_{\text{pred}}(h,w)$ represents the uncertainty of pixel (h,w).

Within the low-uncertainty regions indicated by M_u, we introduce three complementary losses—KL divergence, Jensen-Shannon (JS) divergence, and Dice loss—for dual-branch collaborative optimization. The KL divergence is employed to evaluate the overall discrepancy between the probability distributions of the above two branches:

$$\mathcal{L}_{\text{kl-con}} = \sum_{c=1}^{C} P_{\text{evi}}^{c} \log \frac{P_{\text{evi}}^{c}}{P_{\text{seg}}^{c} + \epsilon}, \quad (20)$$

where P_{evi}^{c} and P_{seg}^{c} denote the probability maps for class c from the segmentation and evidential uncertainty estimation branches, respectively. $\epsilon > 0$ is a small smoothing term for numerical robustness. Minimizing this term encourages the predicted probabilities from the segmentation branch to align with those produced by the evidential uncertainty estimation branch.

However, the KL divergence is asymmetric. When the two distributions differ significantly—particularly in high-uncertainty regions, it may produce unbalanced gradients, leading to biased optimization or unstable convergence. To mitigate this issue, the JS divergence is employed to measure the divergence of two distributions relative to their mean distribution, preventing the model from overfitting to a single branch. It is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{js-con}} = \frac{1}{2}\sum_{c=1}^{C}\Bigg(& P_{\text{seg}}^{c} \log \frac{2P_{\text{seg}}^{c}}{(P_{\text{seg}}^{c} + P_{\text{evi}}^{c}) + \epsilon} \\ & + P_{\text{evi}}^{c} \log \frac{2P_{\text{evi}}^{c}}{(P_{\text{seg}}^{c} + P_{\text{evi}}^{c}) + \epsilon}\Bigg).\end{aligned} \quad (21)$$

To enhance the spatial and structural consistency between the predictions of the two branches, we further introduce the Dice loss to quantify the class-wise overlap between the predicted probability maps:

$$\mathcal{L}_{\text{dice-con}} = 1 - \frac{2\sum_{c=1}^{C} P_{\text{seg}}^{c} P_{\text{evi}}^{c}}{\sum_{c=1}^{C}\left((P_{\text{seg}}^{c})^2 + (P_{\text{evi}}^{c})^2\right) + \epsilon}. \quad (22)$$

For balanced optimization, the three losses are adaptively weighted according to the mean predicted uncertainty. Specifically, the KL divergence is assigned a larger weight when uncertainty is low to improve probabilistic calibration, whereas the Dice loss is prioritized under high uncertainty due to its robustness to outliers. The JS divergence serves as a stable regularization term throughout training. The adaptive weighting scheme is formulated as follows:

$$\hat{\omega}_{\text{kl-con}} = \varepsilon(1 - \mathrm{mean}(U_{\text{pred}})) + \theta, \quad (23)$$

$$\hat{\omega}_{\text{js-con}} = \tfrac{1}{2}\varepsilon(1 - \mathrm{mean}(U_{\text{pred}})), \quad (24)$$

$$\hat{\omega}_{\text{dice-con}} = \varepsilon\,\mathrm{mean}(U_{\text{pred}}) + \theta, \quad (25)$$

where ε and θ denote the weight balancing coefficient and the base weight, respectively, which are set to 0.8 and 0.2 in our experiments. The weights are then normalized to ensure a unit sum:

$$\hat{\omega} = \hat{\omega}_{\text{kl-con}} + \hat{\omega}_{\text{js-con}} + \hat{\omega}_{\text{dice-con}}, \quad (26)$$

$$\omega_{\text{kl-con}} = \frac{\hat{\omega}_{\text{kl-con}}}{\hat{\omega}}, \omega_{\text{js-con}} = \frac{\hat{\omega}_{\text{js-con}}}{\hat{\omega}}, \omega_{\text{dice-con}} = \frac{\hat{\omega}_{\text{dice-con}}}{\hat{\omega}}, \quad (27)$$

Finally, the consistency constraint loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{con}} = \frac{1}{\|M_u\|_1}\sum_{h,w} M_u \odot \big(&\omega_{\text{kl-con}}\mathcal{L}_{\text{kl-con}} \\ &+ \omega_{\text{js-con}}\mathcal{L}_{\text{js-con}} + \omega_{\text{dice-con}}\mathcal{L}_{\text{dice-con}}\big),\end{aligned} \quad (28)$$

where $\|\cdot\|_1$ denotes the L_1-norm and $\odot$ represents element-wise multiplication.

3.7. Loss Function

In summary, the final loss function is defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{ce}} + \lambda_2\mathcal{L}_{\text{align}} + \lambda_3\mathcal{L}_{\text{edl}} + \lambda_4\mathcal{L}_{\text{con}}, \quad (29)$$

where λ_1 and λ_2 are balance coefficients fixed to 1.0. To mitigate interference from unreliable uncertainty modeling during the early training stage, we apply an annealing strategy to the evidential loss weight λ_3, defined as $\lambda_3 = \min\left(0.1, 0.1 \cdot \frac{t}{0.5T}\right)$, where t denotes the index of the current training epoch and T represents the total number of training epochs. Additionally, to prevent unreliable uncertainty estimates from providing misleading supervision for the consistency constraint, we adopt a delayed activation strategy. The consistency constraint loss is gradually activated only after the training process exceeds the halfway point ($t > 0.5T$), with its weight defined as: $\lambda_4 = \min\left(0.1, 0.1 \cdot \frac{t-0.5T}{0.5T}\right)$. This mechanism promotes the model to focus on uncertainty modeling during the early training phase while progressively strengthening consistency constraints as reliable uncertainty estimation has been established.

4. Experiments

4.1. Datasets

We evaluate the proposed model on three publicly available datasets, as detailed below:

Abdominal CT dataset(ABD-CT) [27]: This dataset is obtained from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, comprising 30 3D abdominal CT scans. In this study, four key abdominal organs are selected as segmentation targets, including the left kidney (LK), right kidney (RK), liver, and spleen.

Abdominal MRI dataset(ABD-MRI) [25]: This dataset is obtained from the ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge, comprising 20 3D T2-SPIR MRI scans. It shares the same anatomical annotation labels as the abdominal CT dataset, facilitating cross-modality analysis and comparison.

Cardiac MRI dataset(CMR) [56]: This dataset is obtained from the MICCAI 2019 Multi-Sequence Cardiac MRI Segmentation Challenge, comprising 35 clinical 3D cardiac MRI scans. This study focuses on the segmentation of three key cardiac structures: the blood pool (LV-BP), the left ventricular myocardium (LV-MYO), and the right ventricular myocardium (RV).

4.2. Experimental Settings

To systematically evaluate the performance differences between the proposed model and existing FSMIS methods, our study adopts the same experimental settings as SSL-ALPNet [35, 36], considering two experimental settings: (1) **Setting 1**: the target classes in the test set may appear as unlabeled background regions in training images; (2) **Setting 2**: test classes are completely unseen during training, meaning all slices containing test classes are strictly excluded from the training process.

Due to the high spatial continuity of target structures across slices in the CMR dataset, complete exclusion of slices containing test classes is impractical. Therefore, only Setting 1 is used for this dataset.

4.3. Implementation Details

To ensure a unified evaluation standard, we adopt the same implementation details as in [35, 36]. Specifically, 3D scanned images are decomposed into 2D slices and resampled to 256×256. For single-channel MRI or CT slices, we replicate them three times along the channel dimension to match the input format of the proposed model.

Model evaluation is conducted using 5-fold cross-validation, where each dataset is evenly partitioned into five subsets. In each fold, one subset is used for testing while the remaining four are used for training. The final performance reported is the average across all folds. During each fold, a single 3D volume is randomly selected from the patient samples to form the support set, while the remaining samples of the same class constitute the query set. The slice partitioning between the support and query sets follows the region matching strategy proposed in [40]: the slices of each 3D volume (in both support and query sets) are divided into three equally sized subregions. For any slice in a particular subregion of a query volume, its corresponding support slice is defined as the central slice of the same subregion in the support volume.

The proposed method is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU. Following the experimental setup in GMRD [8], the number of generated foreground and background descriptors is fixed at 100 and 600, respectively. The experiment adopts a 1-way 1-shot few-shot learning setting. Model parameters are optimized using stochastic gradient descent (SGD) with an initial learning rate of 0.001, which decays exponentially by a rate of 0.98. The total number of training iterations is 36K, with 3000 iterations per epoch and a batch size of 1.

4.4. Evaluation Metric

The Sorensen-Dice coefficient is adopted as the evaluation metric, defined as:

$$\text{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \times 100\% \tag{30}$$

where X and Y denote the predicted mask and the ground truth, respectively. $|X \cap Y|$ represents the number of pixels in the intersection of the predicted and ground truth regions, while $|X|$ and $|Y|$ denote the total number of pixels in each region. The final reported results are obtained by averaging the Dice scores across the five-fold cross-validation.

4.5. Performance Comparisons

Following the experimental setup of SSL-ALPNet [35, 36], we compare the proposed method with eight representative FSMIS methods, including SENet [40], PANet [50], SSL-ALPNet [35], ADNet [18], Q-Net [43], CAT-Net [32], RPT [55], and GMRD [8]. The experimental results are summarized in Table 1 and 2, which provides a comprehensive evaluation of the proposed FSMIS method on three publicly available datasets.

As shown in Table 1, the proposed method achieves superior segmentation performance compared to existing methods under both experimental settings. Under Setting 1, our 5-fold cross-validated method attains average Dice scores of 80.67% and 83.83% on the ABD-CT and ABD-MRI datasets, respectively, surpassing the current highest result (GMRD) by 2.15% and 0.93%. Under Setting 2, our method achieves an average Dice score of 78.98% on the ABD-CT dataset, representing a 1.66% improvement over GMRD. Notably, for the spleen class, our method exhibits a significant gain of 5.77% compared to the best existing result (GMRD). On the ABD-MRI dataset, our method also demonstrates consistent superiority, achieving a 2.54% improvement in average Dice score over GMRD.

As shown in Table 2, the proposed method yields the best performance among all compared methods on the CMR

Setting	Method	ABD-CT					ABD-MRI				
		Spleen	Liver	LK	RK	Mean	Spleen	Liver	LK	RK	Mean
1	PA-Net [50]	36.04	49.55	20.67	21.19	32.86	40.58	50.40	30.99	32.19	38.53
	SE-Net [40]	43.66	35.42	24.42	12.51	29.00	47.30	29.02	45.78	47.96	42.51
	SSL-ALPNet [35]	70.96	78.29	72.36	71.81	73.35	72.18	76.10	81.92	85.18	78.84
	ADNet [18]	63.48	77.24	72.13	**79.06**	72.97	72.29	82.11	73.86	85.80	78.51
	Q-Net [43]	78.06	73.36	78.26	77.63	76.83	75.99	81.74	78.36	87.98	81.02
	CAT-Net [32]	67.65	75.31	63.36	60.05	66.59	68.83	78.98	74.01	78.90	75.18
	RPT [55]	79.13	**82.57**	77.05	72.58	77.83	76.37	**82.86**	80.72	89.82	82.44
	GMRD [8]	78.31	79.60	**81.70**	74.46	78.52	76.09	81.42	83.96	90.12	82.90
	Ours (DU-Net)	**81.00**	81.98	80.88	78.81	**80.67**	**77.67**	82.37	**84.63**	**90.65**	**83.83**
2	PA-Net [50]	29.59	38.42	32.34	17.37	29.43	50.90	42.26	53.45	38.64	46.33
	SE-Net [40]	0.23	0.27	32.83	14.34	11.91	51.80	27.43	62.11	61.32	50.66
	SSL-ALPNet [35]	60.25	73.65	63.34	54.82	63.02	67.02	73.05	73.63	78.39	73.02
	ADNet [18]	50.97	70.63	48.41	40.52	52.63	59.44	77.03	59.64	56.68	63.20
	Q-Net [43]	74.69	70.73	70.16	71.71	71.82	65.37	78.25	64.81	65.94	68.59
	CAT-Net [32]	66.02	**80.51**	68.82	64.56	70.88	67.31	75.02	75.31	83.23	75.22
	RPT [55]	70.80	75.24	72.99	67.73	71.69	**75.46**	76.37	78.33	86.01	79.04
	GMRD [8]	75.30	80.39	77.40	76.17	77.32	73.25	80.25	78.65	86.66	79.70
	Ours (DU-Net)	**81.07**	79.91	**78.38**	**76.57**	**78.98**	73.49	**82.97**	**82.63**	**89.87**	**82.24**

Table 1: Comparison of qualitative results with other FSMIS methods on ABD-CT and ABD-MRI under setting 1 and setting 2. Dice score (%) is used as the metric, the best value is shown in bold, the second-best value is underlined, and '–' indicates not reported.

Method	CMR			
	LV-BP	LV-MYO	RV	Mean
PA-Net [50]	58.04	25.18	12.86	32.02
SE-Net [40]	72.77	44.76	57.13	58.20
SSL-ALPNet [35]	83.99	66.74	79.96	76.90
ADNet [18]	87.53	62.43	77.31	75.76
Q-Net [43]	90.25	65.92	78.19	78.15
CAT-Net [32]	90.54	66.85	79.71	79.03
RPT [55]	89.90	66.91	80.78	79.19
GMRD [8]	90.00	67.04	80.29	79.11
Ours (DU-Net)	**90.71**	**69.10**	**81.76**	**80.52**

Table 2: Comparison of qualitative results with other FSMIS methods on CMR under Settings 1. Dice score (%) is used as the metric, the best value is shown in bold, the second-best value is underlined.

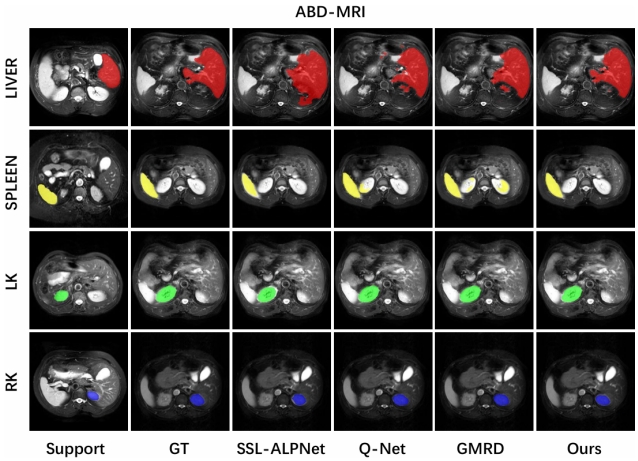

Figure 2: Comparison of qualitative results with other FSMIS methods on ABD-MRI.

dataset, achieving a 1.41% improvement in average Dice score over the previous best-performing method (RPT) and attaining the highest Dice scores across all three cardiac structures: LV-BP, LV-MYO, and RV.

In summary, our method exhibits outstanding segmentation performance across diverse experimental settings and multi-modal datasets, demonstrating its significant advantages in FSMIS tasks.

4.6. Visualization

To facilitate intuitive analysis and comparison, Fig. 2, 3 and 4 visualize segmentation results of the proposed method and several baseline methods on the ABD-CT, ABD-MRI, and CMR datasets, respectively. As shown in Fig. 2 and 3, our method exhibits superior segmentation performance compared to baseline methods, particularly in accurately delineating organ boundaries and complex regions on the

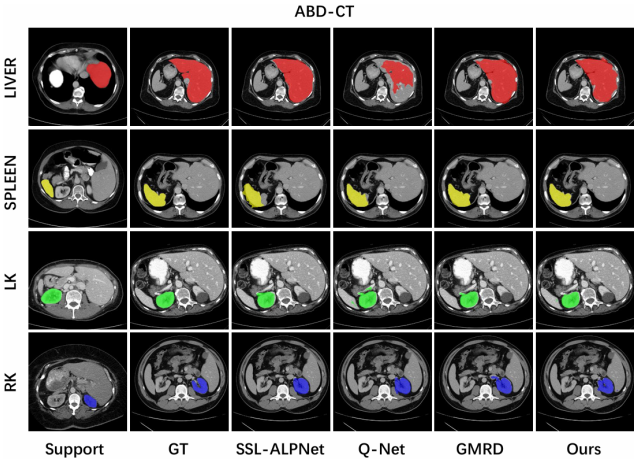

Figure 3: Comparison of qualitative results with other FS-MIS methods on ABD-CT.

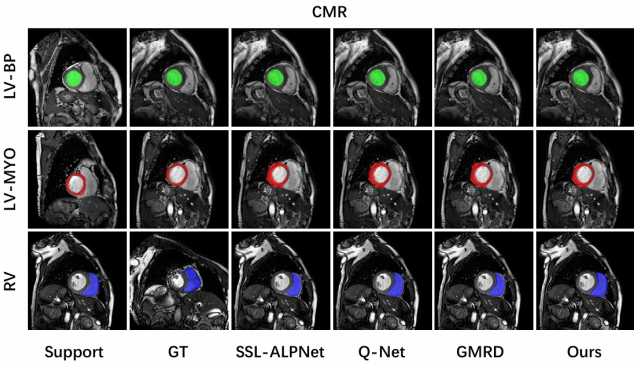

Figure 4: Comparison of qualitative results with other FS-MIS methods on CMR.

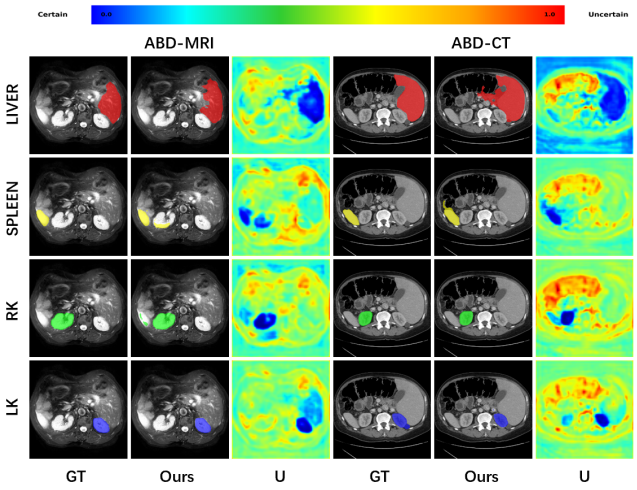

Figure 5: Visualization results of the untertainty maps generated by our model on ABD-CT and ABD-MRI.

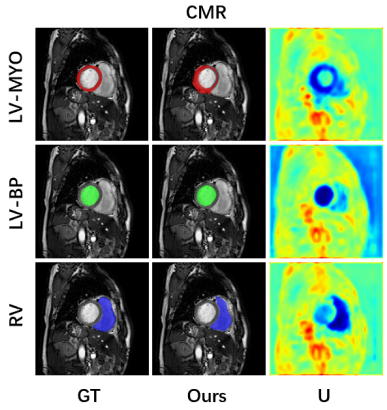

Figure 6: Visualization results of the untertainty maps generated by our model on CMR.

ABD-CT and ABD-MRI datasets. As shown in Fig. 4, our method demonstrates better generalization across the three cardiac structures on the CMR dataset, effectively alleviating the over-segmentation issues commonly observed in baseline methods.

Furthermore, to evaluate the reliability of model predictions, we visualize prediction-level uncertainty maps in Fig. 5 and 6. Uncertainty maps exhibit uniform, low-brightness distributions in regions with high prediction confidence, such as the interior of organ structures. Conversely, regions with high uncertainty display bright responses, such as organ boundaries or regions affected by imaging artifacts. Compared to baseline methods, the uncertainty maps generated by our method provide an intuitive depiction of regions where the model is "uncertain", thereby demonstrating the effectiveness of the proposed method in identifying potentially unreliable predictions.

4.7. Ablation Studies

We conducted an ablation study on the ABD-MRI dataset under setting 2 to systematically evaluate the contribution of each key component to model performance.

1) **Component performance:** We present the contribution of each key component of the proposed DU-Net in Table 3. We progressively add individual components to validate their effectiveness. First, incorporating the DUSMF module into the baseline framework improved the average Dice score by 0.8%, indicating that this mechanism effectively suppresses the interference from high-uncertainty descriptors. Subsequently, after introducing the evidential uncertainty estimation branch, the average Dice score rises to 81.18%, demonstrating that uncertainty modeling at the prediction level based on EDL provides more reliable decision support for segmentation. Finally, imposing a consistency constraint loss between the segmentation and eviden-

Baseline	DUSMF	Evidential-Branch	L_{con}	Spleen	Liver	LK	RK	Mean
✓				73.25	80.25	78.65	86.66	79.70
✓	✓			72.61	80.50	81.15	87.73	80.50
✓	✓	✓		73.06	81.63	81.44	88.59	81.18
✓	✓	✓	✓	**73.49**	**82.97**	**82.63**	**89.87**	**82.24**

Table 3: Ablation study for the Component performance. Dice score(%) is used as the metric.

$L_{\text{kl-con}}$	$L_{\text{js-con}}$	$L_{\text{dice-con}}$	Spleen	Liver	LK	RK	Mean
			73.06	81.63	81.44	88.59	81.18
✓			73.00	82.53	80.26	89.35	81.29
	✓		72.62	82.45	79.83	88.53	80.86
		✓	72.61	82.01	79.31	88.97	80.73
✓	✓		73.27	81.93	81.38	89.02	81.40
✓		✓	73.22	82.37	81.07	89.46	81.53
	✓	✓	73.38	82.76	80.37	88.63	81.29
✓	✓	✓	**73.49**	**82.97**	**82.63**	**89.87**	**82.24**

Table 4: Ablation study for the consistency constraint loss term. Dice score(%) is used as the metric.

Region	Spleen	Liver	LK	RK	Mean
$R_{\text{low-u}}$	72.66	82.24	81.14	89.48	81.38
R_{all}	**73.49**	**82.97**	**82.63**	**89.87**	**82.24**

Table 5: Ablation study for the consistency constraint loss application regions. Dice score(%) is used as the metric.

λ_3(max)	Annealing	Spleen	Liver	LK	RK	Mean
0.05	✓	72.85	82.26	79.42	88.07	80.65
0.1	✓	**73.49**	**82.97**	**82.63**	**89.87**	**82.24**
0.1		72.52	82.93	81.25	89.66	81.59
0.15	✓	72.99	82.37	81.41	89.38	81.54

Table 6: Ablation study for the annealing strategy. Dice score(%) is used as the metric.

Schedule	Spleen	Liver	LK	RK	Mean
0T	**74.18**	82.54	81.05	88.31	81.52
0.25T	73.47	82.39	81.48	89.40	81.69
0.5T	73.49	**82.97**	82.63	**89.87**	**82.24**
0.75T	74.00	81.93	**83.07**	89.46	82.12

Table 7: Ablation study for the delayed activation strategy. Dice score(%) is used as the metric.

tial uncertainty estimation branches significantly improves model performance. This improvement can be attributed to the alignment of the predicted probability distributions between the two branches, which facilitates bidirectional supervision and collaborative optimization.

2) **Consistency constraint loss term:** To evaluate the impact of each loss term on the proposed consistency constraint loss, we conduct ablation experiments for both individual and combined losses. Table 4 summarizes the average Dice scores under different loss combinations. $L_{\text{kl-con}}$ measures the divergence between the predicted probability distributions of the segmentation and evidential uncertainty estimation branches, providing a slight improvement in average Dice when used alone. $L_{\text{js-con}}$ promotes distributional smoothness and alignment to ensure training stability, but limits gains in segmentation accuracy. $L_{\text{dice-con}}$ optimizes regional overlap but may induce distributional deviation and degrade performance when used alone. Combining $L_{\text{kl-con}}$ and $L_{\text{dice-con}}$ improves performance, but may cause gradient conflicts without stabilization. Introducing $L_{\text{js-con}}$ alleviates this issue and enables superior segmentation results.

3) **Consistency constraint loss application regions:** As shown in Table 5, we compare applying the consistency constraint loss only in low-uncertainty regions versus all pixels. Applying the loss to all pixels lowers Dice by 0.86% relative to low-uncertainty regions, indicating that enforcing prediction consistency in high-uncertainty regions may excessively propagate erroneous predictions across branches.

4) **Annealing strategy:** As shown in Table 6, we ablate the annealing strategy of the evidential loss weight λ_3 in Eq. (29). The results confirm that adopting an annealing strategy with a maximum weight of 0.1 outperforms the fixed-weight setting, as it suppresses interference during early training and achieves an optimal balance between segmentation performance and uncertainty modeling.

5) **Delayed activation strategy:** Table 7 reports the impact of the proposed delayed activation strategy for the consistency constraint loss. We consider four activation schedules: without delay, and activating the loss at 1/4, 1/2, and 3/4 of the training process (denoted as 0T, 0.25T, 0.5T, and 0.75T). The results show that the model achieves the best performance when the consistency constraint loss is activated after the midpoint of training. This observation suggests that deferring the activation of consistency constraints until the model has developed a preliminary stable foundation for modeling uncertainty leads to a more effective regularization effect.

4.8. Efficiency Analysis

Table 8 compares the efficiency of our DU-Net with baseline methods on the ABD-CT dataset under Setting 1. Our method achieves effective uncertainty modeling with only a marginal increase in computational overhead and inference time over baselines, demonstrating its superiority.

Method	#Params	FLOPs	Time	GPU
SSL-ALPNet [35]	0.5M	11.5G	46.7ms	1.3GB
Q-Net [43]	18.6M	49.8G	18.4ms	4.1GB
GMRD [8]	10.9M	33.0G	27.2ms	5.9GB
Ours (DU-Net)	11.2M	34.5G	27.7ms	6.0GB

Table 8: Efficiency analysis. "#Params" denotes the number of parameters, "FLOPs" denotes the computational cost, "Time" denotes the the inference time per volume and "GPU" denotes the peak GPU memory consumption.

δ_1	Spleen	Liver	LK	RK	Mean
0.05	72.80	**83.33**	79.87	88.93	81.23
0.1	**73.49**	82.97	**82.63**	**89.87**	**82.24**
0.2	72.75	82.51	81.59	89.39	81.56
0.5	72.14	82.06	81.36	89.71	81.32

Table 9: Sensitivity analysis of the hyperparameter δ_1, where $\delta_2 = 0.1$. Dice score(%) is used as the metric.

δ_2	Spleen	Liver	LK	RK	Mean
0.05	73.09	82.36	81.43	88.96	81.46
0.1	**73.49**	**82.97**	**82.63**	**89.87**	**82.24**
0.2	72.35	82.43	80.32	89.33	81.11
0.5	72.99	82.13	79.76	88.95	80.96

Table 10: Sensitivity analysis of the hyperparameter δ_2, where $\delta_1 = 0.1$. Dice score(%) is used as the metric.

β	Spleen	Liver	LK	RK	Mean
0.6	72.91	81.98	80.16	89.55	81.15
0.8	73.41	81.62	**82.69**	89.51	81.81
1.0	73.49	**82.97**	82.63	**89.87**	**82.24**
1.2	**73.64**	82.48	81.24	89.60	81.74

Table 11: Sensitivity analysis of the hyperparameter β. Dice score(%) is used as the metric.

(ε , θ)	Spleen	Liver	LK	RK	Mean
(0.6 , 0.4)	**73.61**	82.17	81.50	88.21	81.37
(0.7 , 0.3)	73.40	82.84	80.85	89.78	81.72
(0.8 , 0.2)	73.49	**82.97**	**82.63**	**89.87**	**82.24**
(0.9 , 0.1)	72.83	82.62	79.95	89.39	81.20

Table 12: Sensitivity analysis of the hyperparameters ε and θ. Dice score(%) is used as the metric.

4.9. Hyperparameter Analysis

As shown in Tables 9–12, we conduct a hyperparameter sensitivity analysis on the ABD-MRI dataset under setting 2. Specifically, δ_1 and δ_2 represent descriptor-level uncertainty scaling coefficient in Eqs. (14)–(15), β represents the threshold coefficient in Eq. (19) and ε and θ represent the weight balancing coefficients in Eqs. (23)–(25).

When $\delta_1 = \delta_2 = 0.1$, the model achieves the optimal performance. Specifically, smaller values inadequately suppress noise of high-uncertainty descriptors, while larger values weaken discriminative features, reducing segmentation accuracy. The threshold coefficient β reaches its best performance at $\beta = 1.0$, while larger or smaller values lead to slight performance degradation, indicating that overly aggressive or conservative uncertainty filtering disrupts the selection of reliable supervision regions under the consistency constraint. For weight balancing coefficients, we evaluate various weight pairs (ε,θ) under the constraint $\varepsilon + \theta = 1$ to focus on the relative balance between uncertainty-aware adaptation and base supervision without altering the optimization dynamics. The configuration (ε,θ) =(0.8,0.2) delivers the best overall performance, suggesting that emphasizing uncertainty-aware adaptive weighting while retaining a moderate base term enhances segmentation reliability.

5. Conclusion

This paper proposes DU-Net, a novel framework that utilizes EDL to model uncertainty at both the descriptor and prediction levels, leveraging dual-level uncertainty to guide segmentation. The DUSMF module is introduced to estimate the uncertainty of the descriptor level, and suppresses noise from high-uncertainty descriptors. The Evidential Dual-Branch Prediction module enables joint optimization between segmentation and evidential uncertainty estimation, with the two branches mutually supervised through a consistency constraint applied to low-uncertainty regions. Extensive experiments on three public medical image datasets demonstrate that DU-Net not only improves segmentation accuracy but also enhances prediction interpretability, offering a promising approach for high-quality segmentation under limited labeled data.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 61902193) and the PAPD fund.

References

[1] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial intelligence review*, 54(1):137–178, 2021. 1

[2] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial intelligence review*, 54(1):137–178, 2021. 2

[3] W. Bao, Q. Yu, and Y. Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 13349–13358, 2021. 3

[4] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 3

[5] M. Chen, J. Gao, and C. Xu. Uncertainty-aware dual-evidential learning for weakly-supervised temporal action localization. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15896–15911, 2023. 3

[6] X. Chen, N. Pawlowski, B. Glocker, and E. Konukoglu. Normative ascent with local gaussians for unsupervised lesion detection. *Medical Image Analysis*, 74:102208, 2021. 1

[7] Y. Chen, Z. Yang, C. Shen, Z. Wang, Z. Zhang, Y. Qin, X. Wei, J. Lu, Y. Liu, and Y. Zhang. Evidence-based uncertainty-aware semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 170:108004, 2024. 2

[8] Z. Cheng, S. Wang, T. Xin, T. Zhou, H. Zhang, and L. Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 43(6):2202–2214, 2024. 1, 2, 3, 5, 8, 9, 12

[9] T. Dawood, E. Chan, R. Razavi, A. P. King, and E. Puyol-Antón. Addressing deep learning model calibration using evidential neural networks and uncertainty-aware training. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 2

[10] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54(3-4):515–528, 1967. 3

[11] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 3

[12] H. Ding, C. Sun, H. Tang, D. Cai, and Y. Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2488–2497, 2023. 1

[13] R. Feng, X. Zheng, T. Gao, J. Chen, W. Wang, D. Z. Chen, and J. Wu. Interactive few-shot learning: Limited supervision, better medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2575–2588, 2021. 1

[14] J. Gao, M. Chen, L. Xiang, and C. Xu. A comprehensive survey on evidential deep learning and its applications. *arXiv preprint arXiv:2409.04720*, 2024. 2, 3

[15] F. C. Ghesu, B. Georgescu, E. Gibson, S. Guendel, M. K. Kalra, R. Singh, S. R. Digumarthy, S. Grbic, and D. Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International conference on medical image computing and computer-assisted intervention*, pages 676–684. Springer, 2019. 2

[16] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, E. Gibson, R. Vishwanath, A. Balachandran, J. M. Balter, Y. Cao, R. Singh, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855, 2021. 2

[17] N. R. Gudhe, M. Sudah, A. Mannermaa, V.-M. Kosma, and H. Behravan. Multi-view deep evidential fusion neural network for assessment of screening mammograms. 2024. 2

[18] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022. 2, 8, 9

[19] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022. 5

[20] S. Hansen, S. Gautam, S. A. Salahuddin, M. Kampffmeyer, and R. Jenssen. Adnet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement. *Medical Image Analysis*, 89:102870, 2023. 2

[21] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 1

[22] L. Huang, S. Ruan, and T. Denoeux. Belief function-based semi-supervised learning for brain tumor segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 160–164. IEEE, 2021. 2

[23] L. Huang, S. Ruan, Y. Xing, and M. Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024. 3

[24] A. Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018. 3

[25] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021. 8

[26] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014. 3

[27] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015. 8

[28] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1, 2

[29] H. Li, Y. Nan, J. Del Ser, and G. Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Computing and Applications*, 35(30):22071–22085, 2023. 2

[30] Z.-P. Li, H.-L. Su, X.-B. Zhu, X.-M. Wei, X.-S. Jiang, V. Gribova, V. F. Filaretov, and D.-S. Huang. Hierarchical graph pooling with self-adaptive cluster aggregation. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):1198–1207, 2021. 1

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[32] Y. Lin, Y. Chen, K.-T. Cheng, and H. Chen. Few shot medical image segmentation with cross attention transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 233–243. Springer, 2023. 8, 9

[33] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[34] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 2, 3

[35] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European conference on computer vision*, pages 762–780. Springer, 2020. 1, 2, 3, 8, 9, 12

[36] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022. 1, 6, 8

[37] L. Pu, N. S. Gezer, S. F. Ashraf, I. Ocak, D. E. Dresser, and R. Dhupar. Automated segmentation of five different body tissues on computed tomography using deep learning. *Medical physics*, 50(1):178–191, 2023. 1

[38] S. S. Rao and L. Berke. Analysis of uncertain structural systems using interval analysis. *AIAA journal*, 35(4):727–735, 1997. 3

[39] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[40] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. 'squeeze & excite' guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020. 1, 2, 8, 9

[41] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 2, 3

[42] G. Shafer. A mathematical theory of evidence. 2020. 3

[43] Q. Shen, Y. Li, J. Jin, and B. Liu. Q-net: Query-informed few-shot medical image segmentation. In *Proceedings of SAI Intelligent Systems Conference*, pages 610–628. Springer, 2023. 2, 8, 9, 12

[44] C. Sun, Y. Jia, and Y. Wu. Evidential reasoning for video anomaly detection. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, pages 2106–2114, 2022. 3

[45] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3918–3928, 2021. 1

[46] P. Teng, W. Liu, X. Wang, D. Wu, C. Yuan, Y. Cheng, and D.-S. Huang. Beyond singular prototype: A prototype splitting strategy for few-shot medical image segmentation. *Neurocomputing*, 597:127990, 2024. 1, 2

[47] T. Vrtovec, D. Močnik, P. Strojan, F. Pernuš, and B. Ibragimov. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Medical physics*, 47(9):e929–e950, 2020. 1

[48] C. Wachinger, P. Golland, M. Reuter, and W. Wells. Gaussian process interpolation for uncertainty estimation in image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 267–274. Springer, 2014. 3

[49] M. Wallman, N. P. Smith, and B. Rodriguez. Computational methods to reduce uncertainty in the estimation of cardiac conduction properties from electroanatomical recordings. *Medical image analysis*, 18(1):228–240, 2014. 3

[50] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 3, 5, 6, 8, 9

[51] Y. Xu, J. Tang, A. Men, and Q. Chen. Eviprompt: A training-free evidential prompt generation method for adapting segment anything model in medical images. *IEEE Transactions on Image Processing*, 2024. 2

[52] R. R. Yager and L. Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008. 3

[53] R. R. Yager and L. A. Zadeh. *An introduction to fuzzy logic applications in intelligent systems*, volume 165. Springer Science & Business Media, 2012. 3

[54] H. Zhang, Y. Liu, Y. Wang, L. Wang, and Y. Qiao. Learning discriminative feature representation for open set action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pages 7696–7705, 2023. 3

[55] Y. Zhu, S. Wang, T. Xin, and H. Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–280. Springer, 2023. 8, 9

[56] X. Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018. 8

[57] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu. Tbrats: Trusted brain tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 503–513. Springer, 2022. 2