Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(2): 2212–2232.
DOI: 10.3934/mbe.2024097
Received: 24 November 2023
Revised: 05 January 2024
Accepted: 08 January 2024
Published: 11 January 2024

Research article

Dual uncertainty-guided multi-model pseudo-label learning for semi-supervised medical image segmentation

Zhanhong Qiu, Weiyan Gan, Zhi Yang, Ran Zhou and Haitao Gan*

School of Computer Science, Hubei University of Technology, Wuhan 430068, China

* **Correspondence:** E-mail: htgan01@hbut.edu.cn.

Abstract: Semi-supervised medical image segmentation is currently a highly researched area. Pseudo-label learning is a traditional semi-supervised learning method aimed at acquiring additional knowledge by generating pseudo-labels for unlabeled data. However, this method relies on the quality of pseudo-labels and can lead to an unstable training process due to differences between samples. Additionally, directly generating pseudo-labels from the model itself accelerates noise accumulation, resulting in low-confidence pseudo-labels. To address these issues, we proposed a dual uncertainty-guided multi-model pseudo-label learning framework (DUMM) for semi-supervised medical image segmentation. The framework consisted of two main parts: The first part is a sample selection module based on sample-level uncertainty (SUS), intended to achieve a more stable and smooth training process. The second part is a multi-model pseudo-label generation module based on pixel-level uncertainty (PUM), intended to obtain high-quality pseudo-labels. We conducted a series of experiments on two public medical datasets, ACDC2017 and ISIC2018. Compared to the baseline, we improved the Dice scores by 6.5% and 4.0% over the two datasets, respectively. Furthermore, our results showed a clear advantage over the comparative methods. This validates the feasibility and applicability of our approach.

Keywords: medical image segmentation; semi-supervised learning; pseudo-labeling; uncertainty estimation

1. Introduction

Medical image analysis has found extensive applications in clinical research, enhancing healthcare quality by facilitating precise lesion detection and disease categorization. Medical image segmentation is a crucial task in medical image analysis, involving the labeling of each pixel as a specific class to obtain a segmentation map of the target region. It plays a vital role in subsequent diagnosis, surgical planning, and postoperative analysis, attracting significant attention from researchers. In recent years,

deep learning-based methods have achieved state of the art performance in various medical image segmentation tasks [1, 2]. Beyond single segmentation tasks, Qiu et al. [3] proposed a joint learning framework to accomplish parallel deformable registration and segmentation tasks by minimizing an integrated loss function. However, these methods often demand a large amount of annotated data for training, posing challenges in terms of both acquisition and time consumption. Given the insufficient amount of labeled data, some studies have attempted to address this problem from different perspectives. For instance, Kim et al. [4] proposed an end-to-end unsupervised image segmentation network that enables image segmentation in a completely unsupervised manner. The proposed CNN architecture comprises convolutional filters for feature extraction and a differentiable process for feature clustering, enabling end-to-end network training. Lei et al. [5] introduced a novel one-time medical image segmentation framework that combines one-time localization and weakly supervised segmentation. Using a labeled image and a large number of unlabeled volumes, a noisy training algorithm is employed to supervise the segmentation model. Furthermore, semi-supervised learning methods have found broad application in medical image segmentation, enabling the training of high-precision models with limited labeled data and a substantial amount of unlabeled data.

Traditional semi-supervised learning (SSL) methods use limited labeled data and compensate for the information gap in unlabeled data by generating pseudo-labels [6]. Bai et al. [7] proposed a self-training method for semi-supervised cardiac image segmentation, iteratively generating and refining pseudo-labels using conditional random fields (CRF) to enhance segmentation accuracy. The quality of pseudo-labels is crucial due to inevitable model noise. Subsequent efforts focus on estimating pseudo-label uncertainty to generate high-quality labels and reduce noise during training. For example, Dropout [8] simulates model uncertainty by stochastically dropping out neurons during training and estimates pseudo-label uncertainty by calculating differences between outputs from multiple dropout operations. Additionally, many semi-supervised methods use SoftMax confidence [9, 10] to estimate pseudo-label uncertainty by setting a predefined threshold in the SoftMax layer. If the predicted probability exceeds this threshold, the corresponding pixel is considered part of the foreground. While these methods enhance pseudo-label quality to some extent, they often introduce additional complexity and computational overhead to the model.

Recent research has explored estimating model uncertainty by calculating the Kullback-Leibler (KL) divergence between two sets of predictions. KL divergence, also known as relative entropy, measures the similarity between two probability distributions. Luo et al. [11] calculated KL divergence between predictions at different scales within the network decoder and the average prediction to quantify uncertainty for rectifying consistency loss. Wu et al. [12] expanded the network with additional decoders to obtain diverse predictions, applying a sharpening function and imposing consistency loss between the sharpened predictions and soft pseudo-labels to constrain the model. This method enables uncertainty estimation in a single forward pass without added computational burden. However, [11, 12] solely used this uncertainty for consistency loss and did not apply it to pseudo-labels. Additionally, these efforts overlook differences between samples. Notably, variations in sample data among different patients are significant, and treating these samples equally is not advisable.

Consistency regularization is a widely studied approach in semi-supervised medical image segmentation. It involves perturbing the same image in various ways and using the predictions from perturbed data to enforce consistent segmentation results from the model. For instance, Xie et al. [13] introduced a method that learns features from unlabeled data through intra- and inter-pair consistency. It estab-

lishes pixel-level relationships between pairs of images in the feature space, creating attention maps. Consistency constraints are then applied to these attention maps from multiple image pairs, resulting in refined attention maps. Chen et al. [14] proposed the generative consistency-based semi-supervised (GCS) model, leveraging reconstruction consistency to enhance the model's texture representation. Additionally, they introduced TRSF-Net to establish spatial correlations in the graph space of the data. One of the most well-known methods is the mean teacher (MT) model [15], which employs the moving average of the student model to update the teacher model and introduces stronger perturbations to the inputs of both models. Consistency loss is then applied to their predictions to encourage more consistent outputs. This approach effectively enhances the model's robustness and has demonstrated significant success in SSL. Consequently, numerous semi-supervised medical image segmentation methods based on the MT model have achieved segmentation accuracy close to that of fully supervised approaches. For instance, Zhang et al. [16] proposed a mutual consistency learning framework based on the MT model. Their backbone network simultaneously produces segmentation probability maps and signed distance maps. The student-teacher models enforce both inter-task and intra-task consistency constraints on these two outputs. Wang et al. [17] extended the MT model to enforce consistency constraints between segmentation tasks, information reconstruction tasks, and signed distance map tasks. These methods have showcased excellent performance in various medical image segmentation tasks, underscoring the high adaptability of the MT model in reducing model uncertainty.

After reviewing previous research, we have identified several potential areas for improvement: 1) Differences between samples may lead to unstable training processes. 2) Uncertainty in model estimation can be used to rectify pseudo-labels. 3) Generating pseudo-labels by averaging teacher models rather than student models can reduce noise accumulation. Recent research [18] introduces both sample-level uncertainty and pixel-level uncertainty to enhance the training process, building upon traditional self-training methods. Sample-level uncertainty is employed to differentiate the priority of unlabeled samples, while pixel-level uncertainty is utilized to correct pseudo-labels. While this work addresses some of the mentioned issues, it has a simplistic approach to dividing unlabeled samples. Additionally, their method employs a single model for pseudo-label generation, leading directly to noise accumulation. In this paper, our focus is on distinguishing unlabeled samples and obtaining high-confidence pseudo-labels. To achieve this, we propose a dual uncertainty-guided multi-model pseudo-label learning framework (DUMM) (Figure 1). Specifically, DUMM comprises two main modules: 1) sample selection module based on sample-level uncertainty (SUS), and 2) multi-model pseudo-label generation module based on pixel-level uncertainty (PUM). The SUS module estimates the sample-level uncertainty of unlabeled samples by performing dropout operations on a pretrained model. It ranks the samples based on their uncertainty and introduces Softmax confidence for filtering. The PUM module introduces a teacher model to generate pseudo-labels, supervises the student model, and utilizes multiple model outputs to estimate pixel-level uncertainty for rectifying pseudo-labels. We validated our method on the publicly available ACDC2017 dataset and the ISIC2018 dataset. Experimental results demonstrate that our model outperforms existing semi-supervised models. In general, our contribution to this work includes the following points:

1) We design a method to evaluate the segmentation stability of unlabeled samples and introduce softmax confidence to guide the division process of unlabeled samples. Reasonably distinguishing and screening unlabeled samples can effectively improve the stability and smoothness of the training process.

2) We propose a multi-model pseudo-label learning module to generate high-confidence pseudo-labels. Generating pseudo-labels through a teacher model can effectively reduce the accumulation of noise, and two different sets of pseudo-labels to supervise the student model can enhance the robustness of the model. At the same time, the pixel-level uncertainty obtained from the output of multiple models can rectify the pseudo-labels.

3) We conducted a series of experiments on two publicly available 2D medical image segmentation datasets to investigate the effectiveness of our method. The comprehensive experimental results demonstrate the effectiveness of our proposed two modules and the superiority of DUMM.

2. Related work

2.1. SSL

SSL is a machine learning approach designed to train models using a small amount of labeled data and a large quantity of unlabeled data. It is typically employed in situations where labeling data at scale is challenging. Current mainstream SSL methods include pseudo-labeling [6] and consistency regularization [19], among others. Pseudo-labeling methods are based on training a model using the labeled data and then generating pseudo-labels for the unlabeled data, which are used as targets in the cross-entropy loss. This process is also referred to as self-training. The segmentation performance of self-training relies heavily on the quality of the generated pseudo-labels. MixMatch [20] applied various data augmentations to unlabeled samples to obtain an average prediction, which is then sharpened to produce low-entropy pseudo-labels. ReMixMatch [21] built upon previous work by making improvements. On one hand, it encouraged the edge distribution of predictions on unlabeled data to match the ground truth edge distribution more closely. On the other hand, it used predictions from weakly augmented data as pseudo-labels for strongly augmented data. The main idea of consistency regularization is to enforce consistency in predictions for different perturbed versions of the same sample, effectively enhancing the model's robustness. For instance, the Π-model [22] used the mean squared difference between the prediction probability distributions of two different transformed views of the same sample as a loss function. During training, this loss is minimized to achieve the goal of maintaining prediction consistency. Building upon this foundation, various SSL methods have emerged, with a primary focus on investigating different data augmentations or transformations to make prediction probability distributions more reliable. Currently, state of the art SSL methods combine pseudo-labeling and consistency regularization. For example, CoMatch [23] introduced two classification heads, one to generate image-level predictions and another to create a connectivity graph structure. Specifically, for the same image, both weak and strong augmentations are applied, with predictions from the weakly augmented image serving as pseudo-labels for the strongly augmented image. Consistency constraints are then imposed on the connectivity graph generated from the predictions of the weakly augmented image and the embedding graph generated from the strongly augmented image. SimMatch [24], on the other hand, uses weakly augmented images to generate semantic pseudo-labels and instance pseudo-labels. It then separately estimates the semantic and instance similarities between samples and fuses these similarities to obtain the final pseudo-labels. Both of these approaches make effective use of both pseudo-labeling and consistency constraints in training, resulting in the current state of the art performance.

In this work, we concurrently employ both consistency regularization and pseudo-labeling. The

utilization of two sets of teacher models for generating pseudo-labels not only enhances the robustness of the student model but also mitigates noise accumulation in the pseudo-labels.

2.2. Uncertainty estimation

Due to the inherent uncertainty in models, disparities often exist between predictions and true labels. Estimating prediction uncertainty can effectively enhance the model's performance [25]. In the semi-supervised domain, Rizve et al. [26] used traditional self-training methods to generate pseudo-labels. Subsequently, they iteratively select a subset of pseudo-labels with higher confidence from the entire pseudo-label set to include in the training dataset. This process continues until the number of selected pseudo-labels converges. In the realm of consistency-based methods, UA-MT [27], building upon the mean-teacher model, introduces uncertainty awareness. It estimates uncertainty in teacher-side predictions and, during the computation of consistency loss, filters out unreliable predictions, retaining only the reliable ones. Recently, UCC [28] introduced a novel uncertainty estimation approach to guide collaborative training. They apply both weak and strong augmentations to the same sample and obtain corresponding predictions. Uncertainty is addressed using softmax voting. For a given pixel, it is considered as foreground only when the prediction probability from the weak augmentation data is greater than or equal to the prediction probability from the strong augmentation data; otherwise, it is classified as background. Similarly, in the field of semi-supervised medical image segmentation, UCMT [29] introduced uncertainty estimation into collaborative training. They apply consistency constraints to the two predictions generated during collaborative training. Additionally, they transform the prediction map into an uncertainty map. Based on this uncertainty map, the K pixels with the lowest confidence in the original image are replaced with the K pixels with the highest confidence. The updated original image undergoes a second round of collaborative training. Xu et al. [30] employed uncertainty estimation to dynamically adjust the weights of consistency regularization, avoiding the use of fixed weights. Specifically, the framework included an additional term for temporal consistency regularization, encouraging coherence between the predictions of the student model and the time-integrated predictions of the teacher model. Simultaneously, the self-integrated teacher model heuristically adjusted regularization weights for each image pair using transformed uncertainty and derived appearance uncertainty. In a different context, Zhang et al. [31] introduced an uncertainty-based approach for reliable imagined trajectory generation. If the accumulated uncertainty along a trajectory reaches a predefined threshold, the trajectory is truncated. This adaptive truncation ensures conservative execution.

We utilize the KL divergence between predictions from multiple models to estimate an uncertainty map. Pixels corresponding to high uncertainty in the pseudo-labels are assigned lower weights, whereas pixels with low uncertainty receive higher weights. This weighting mechanism adapts as self-training progresses, allowing the model to acquire increasingly accurate pseudo-labels.

2.3. Semi-supervised medical image segmentation

In recent years, significant advancements have been made in the field of semi-supervised medical image segmentation. Analogous to traditional SSL, two pivotal methods, namely, pseudo-labeling and consistency regularization, have found widespread application. Additionally, within the realm of semi-supervised medical image segmentation, various studies have extended the MT framework in diverse ways [15]. Pseudo-labeling methods primarily focus on improving the quality of pseudo-labels. For

instance, Wang et al. [32] introduced a trust module to reevaluate pseudo-labels within the model's outputs, setting a threshold to select those with high confidence. Another approach by Shi et al. [33] proposed a conservative-aggressive network, where the conservative setting tends to predict pixels as background, and the aggressive setting tends to predict pixels as foreground. pseudo-labels are then derived from specific regions within the predictions on unlabeled data that overlap between the conservative and aggressive settings. Conversely, methods based on consistency regularization aim to ensure consistent predictions for different perturbed versions of the same data. Zhang et al. [34], for example, randomly transformed unlabeled data multiple times during training, incorporating various transformations such as affine, Euclidean, and similarity. They optimized the supervised loss on labeled data along with a consistency regularization loss on multiple transformed versions of unlabeled data. Basak et al. [35] introduced a novel consistency regularization strategy that encourages the interpolation segmentation of two unlabeled data points to be consistent with the interpolation of their segmentation mappings, thereby minimizing overfitting on labeled data with high confidence values.

Simultaneously, several extensions of the MT framework in semi-supervised medical image segmentation have yielded noteworthy success. For instance, Basak et al. [36] introduced a method that leverages pseudo-labels to guide contrastive learning for more effective semi-supervised medical image segmentation. The pseudo-labels generated through self-training provide additional guidance for contrastive learning, facilitating the learning of discriminative class information and leading to accurate multi-class segmentation. Furthermore, they introduced a novel loss function that collaboratively encourages inter-class separability and intra-class compactness among the learned representations. Another notable contribution comes from Bai et al. [37], who proposed a bidirectional copy-paste approach integrated into the MT framework. In this approach, they performed copy-paste operations by placing randomly cropped labeled images (foreground) onto unlabeled images (background) and vice versa. These composite images were then fed into a student network and supervised by a combination of pseudo-labels and ground truth signals.

Furthermore, recent research in semi-supervised medical image segmentation has moved beyond a singular focus on individual models or datasets. For instance, Xu et al. [38] introduced a non-parametric unlabeled-to-labeled learning scheme that explicitly constrains the predictions of prototype propagation using scarce expert labels. This approach enables the model to leverage discriminative and domain-insensitive features from heterogeneous multi-site data, supporting local centrality. Another notable example is the work by Pan et al. [39], who employed a human-computer interactive tissue prototype learning pipeline to establish a connection between whole-slide image (WSI) patch pretraining and pixel-level tissue segmentation through contrastive learning. Pathologists selected centroids of clusters to construct a tissue prototype dictionary. The original WSI underwent mapping to an embedding space using an encoder, and a pseudo tissue mask was generated by querying the nearest prototype to the current local region.

Our work integrates the MT model into the conventional self-training framework. Two distinct teacher models generate two sets of pseudo-labels to supervise the student model, utilizing different parameter exponential moving averages (EMA) of the student model to update the teacher models. Simultaneously, to enhance the smoothness of the self-training process, we estimate the Softmax probabilities of the predictions for unlabeled data, providing a more rational prioritization of all unlabeled data.

3. Methods

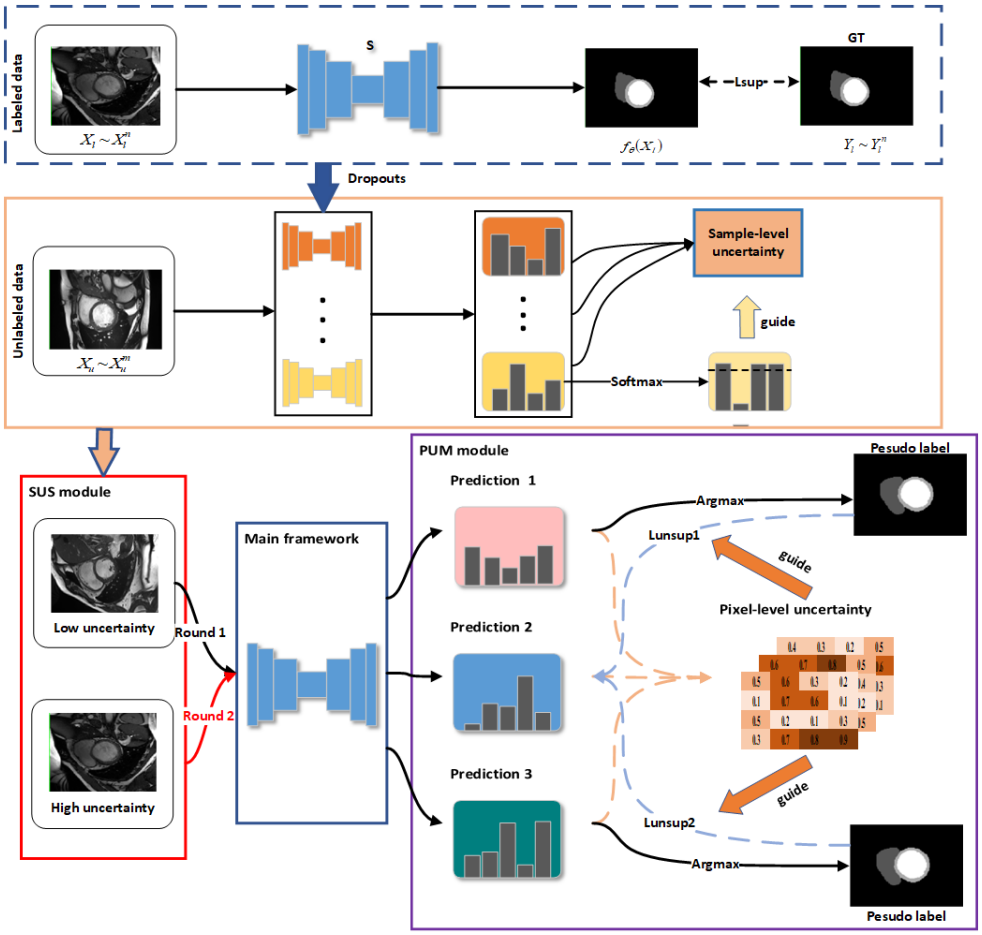

Figure 1. The flowchart of the proposed method (DUMM).

In this section, we will provide a detailed explanation of the proposed method. As shown in Figure 1, our self-training framework (DUMM) consists of two parts. For labeled data, we prioritize and filter unlabeled data through the SUS module, then input them into the PUM module for two rounds of pseudo-label learning based on the prioritization results. Specifically, the SUS module estimates sample-level uncertainty for unlabeled data by performing dropout operations on a pretrained model, considering it as the stability of sample segmentation predictions between different models. Additionally, the Softmax probability map obtained from the predictions is used to guide the prioritization of samples. The PUM module introduces two teacher models to supervise the student model's learning (see Figure 2). The parameters of the two teacher models are updated using different EMAs of the student model. Meanwhile, multiple model predictions are leveraged to estimate pixel-level uncertainty for rectifying pseudo-labels. To precisely describe this work, let's define some mathematical terms.The training dataset is denoted as $D = \{D_l, D_u\}$, where the labeled set is $D_l = \left(X_l \sim X_l^n\right), \left(Y_l \sim Y_l^n\right)$ and the unlabeled set is $D_u = (X_u \sim X_u^m)$. The function $f(\theta)$ represents the parameters of the student model, while $f(\theta'), f(\theta'')$ represent the parameters of the two teacher models.

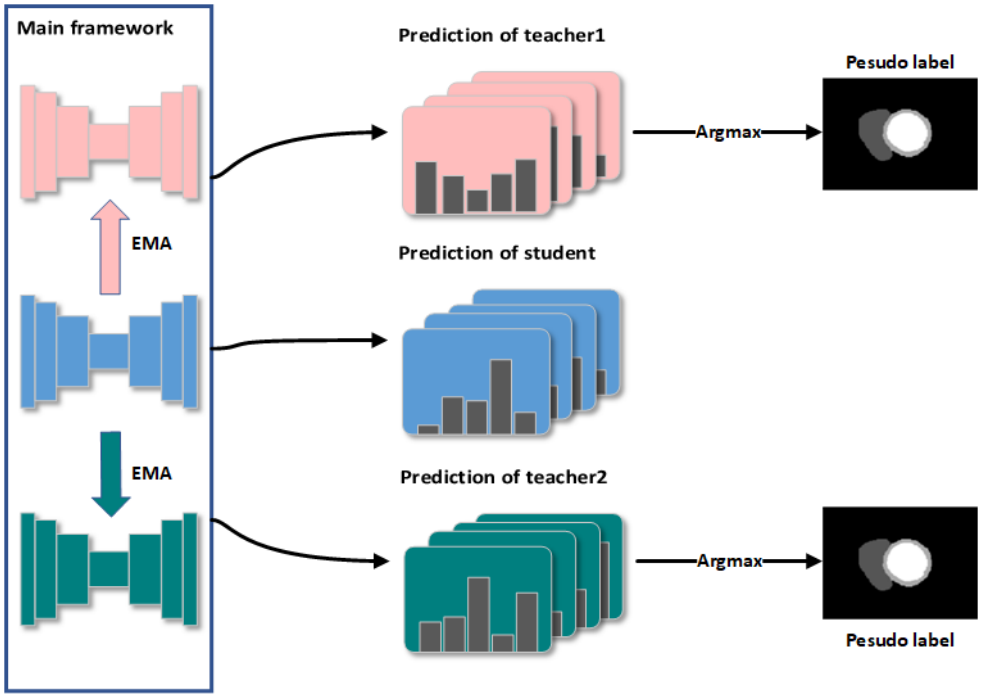

Figure 2. Main segmentation framework.

3.1. Sample selection module with sample-level uncertainty

This section provides a detailed explanation of the SUS module. In order to better utilize unlabeled data, we estimate the sample-level uncertainty of unlabeled samples and prioritize and filter them to obtain a stable and smooth training process. Specifically, we obtain corresponding predictions for unlabeled samples by performing K dropout operations on the pretrained model: $f_1(X_u \mid \theta_u)$, $f_2(X_u \mid \theta_u).\cdots f_k(X_u \mid \theta_u)$. Using the last prediction as the reference, calculate the distance between each prediction and the reference, and take the average of these distances as the sample-level uncertainty, as shown in Eq (3.1).

$$\textit{Sample-level uncertainty} = \sum_{i=1}^{k-1} \text{mean}\left[\left(f_i(X_u \mid \theta_u) - f_k(X_u \mid \theta_u)\right)^2\right] \tag{3.1}$$

The uncertainty at the sample level represents the stability of the segmentation of the sample in different models to some extent. Generally speaking, unlabeled samples with low uncertainty (high segmentation stability) introduce relatively less noise to the model. Based on the above principle, we consider the top 50% of unlabeled samples with low uncertainty as high-priority samples, and the remaining 50% as low-priority samples. In the subsequent training process, high-priority samples are utilized for the first round of pseudo-label learning. After preserving the model parameters, both high-priority and low-priority samples are employed for the second round of pseudo-label learning. In theory, this method can improve the stability and smoothness of the training process to a certain extent.

However, some unlabeled samples may exhibit similar predictions across different model checkpoints but with lower accuracy. This can lead to an improper categorization of unlabeled samples. To address this issue and ensure a smoother self-training process, we apply a softmax operation to

$f_k(X_u \mid \theta_u)$, obtain $Softmax(f_k(X_u \mid \theta_u))$, and calculate the proportion of pixels with probabilities greater than ρ ($\rho = 0.9$) in $Softmax(f_k(X_u \mid \theta_u))$, denoted as S_n. The average value of S_n for all samples is denoted as $\bar{S}_n$. Finally, we select samples from the high-priority set where $S_n < \bar{S}_n$ and replace them with samples from the low-priority set where $S_n > \bar{S}_n$.

3.2. Multi-model pseudo-label generation module with pixel-level uncertainty

In the field of semi-supervised semantic segmentation, traditional self-training methods typically involve the following three steps for training a model: (i) training an initial model using labeled data, (ii) inferring pseudo-labels from the initial model, (iii) jointly training the model on labeled data and pseudo-labeled data, and repeating steps (ii) and (iii) until convergence. While this approach allows for the utilization of unlabeled data to some extent, the inherent noise in pseudo-labels can directly impact the model's training, leading to the accumulation of noise during the iterative process.

Differing from traditional self-training, we employ two sets of average teacher models $f(\theta')$, $f(\theta'')$ to generate two sets of pseudo-labels. During the training process, we develop these teacher models in conjunction with the student model $f(\theta)$ using different weightings of EMA (Eqs (3.2) and (3.3)). This ensemble approach enhances the model's tolerance to inaccurate pseudo-labels while enabling the generation of high-quality predictions for unlabeled samples, thereby improving overall model performance.

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{3.2}$$

$$\theta''_t = \beta\theta''_{t-1} + (1-\beta)\theta_t \tag{3.3}$$

As our model has multiple sets of outputs, pixel-level uncertainty can be obtained by measuring the discrepancies between them. Specifically, pixel-level uncertainty is quantified by computing the KL divergence between the outputs of the student model and the average outputs of the two teacher models, as shown in Eq (3.4).

$$Pixel\text{-}level\ uncertainty\ = D_{kl} = f(X_u \mid \theta_u) \log \frac{f(X_u \mid \theta_u)}{\left(f(X_u \mid \theta'_u) + f(X_u \mid \theta''_u)\right)/2} \tag{3.4}$$

$f(X_u \mid \theta'_u)$ represents predictions generated by the teacher model $f(\theta')$, while $f(X_u \mid \theta''_u)$ represents predictions generated by the teacher model $f(\theta'')$ and D_{kl} represents the KL divergence between two sets of discrete variables. After applying softmax to both sets of predictions and taking their average, the D_{kl}(KL divergence) is computed with respect to the predictions of the student model $f_k(X_u \mid \theta_u)$. Finally, we can obtain the discrete value of each pixel, which we call pixel-level uncertainty. The pixel with a low discrete value represents a stable prediction result and is less prone to prediction errors, while the pixel with a high discrete value represents an unstable prediction result and is more prone to prediction errors. Based on the above analysis, we further utilize this uncertainty to modify the pseudo-label-based unsupervised loss, emphasizing the reliable portions in pseudo-labels and disregarding the unreliable ones. The modified unsupervised loss is shown in Eqs (3.5) and (3.6).

$$L_{unsup1} = \frac{1}{m}\sum_{t=1}^{m}\left[\exp\{-D_{kl}\}\, L_{ce}\left(f(X_u \mid \theta_u), \hat{y}'_u\right) + D_{kl}\right] \tag{3.5}$$

$$L_{unsup2} = \frac{1}{m}\sum_{t=1}^{m}\left[\exp\left\{-D_{kl}\right\} L_{ce}\left(f\left(X_u \mid \theta_u\right), \hat{y}_u''\right) + D_{kl}\right] \tag{3.6}$$

where $\hat{y}_u'$ and $\hat{y}_u''$ are pseudo-labels generated by the teacher models $f(\theta')$ and $f(\theta'')$, respectively.

3.3. The overall loss function

The training procedure of the proposed method is outlined in Algorithm 1. In the pretraining phase, supervised training is conducted using the labeled data:

$$L_{sup} = \left[L_{ce}\left(f\left(X_l \mid \theta_l\right), Y_l\right) + L_{dice}\left(Softmax\left(f\left(X_l \mid \theta_l\right)\right), Y_l\right)\right]/2 \tag{3.7}$$

In the subsequent self-training phase, training is conducted using both labeled and unlabeled data:

$$L_{total} = L_{sup} + L_{unsup} \tag{3.8}$$

where $L_{unsup} = \left(L_{unsup1} + L_{unsup2}\right)/2$, as defined by Eqs (3.5) and (3.6). Additionally, the values of α and β in Eqs (3.2) and (3.3) are set to 0.999 and 0.99, respectively.

Algorithm 1: The process of our proposed method

Input: Labeled training set $D_l = \{X_l, Y_l\}_{i=1}^{n}$;
Unlabeled training set $D_u = \{X_u\}_{i=1}^{m}$;
Student model S and two teacher model T', T''.
Output: Trained student model S.

1 Train model S on D_l based on Eq (3.7) and save pretrained model;
2 Prioritize and screen the unlabeled samples according to Section 3.1;
3 $D_u = D_{u1} \cup D_{u2}$;
4 Assign the parameters θ of model S to the teacher models T' and T'';
5 **for** T **in [1,**T_{max}**] do**
6 Use teacher models T' and T'' to generate two sets of pseudo -labels for D_{u1};
7 Train model S on D_l and D_{u1} based on Eq (3.8);
8 Update the student model's parameters θ using optimizer;
9 Update the two teacher model's parameters θ' and θ'' based on Eqs (3.2) and (3.3);
10 **end for**
11 Retrain model S on D_l and D_{u1} and D_{u2}, following the steps 4–10;
12 **Return:** S.

4. Experiment

4.1. Datasets

The experiments for evaluating the model performance were carried out on two public datasets: (a) ACDC dataset: It contains 100 short-axis MR-cine T1 3D volumes of cardiac anatomy acquired using 1.5T and 3T scanners. In this experiment, we cut these 100 3D samples into 2D slices, resulting

in a total of 1903 2D datasets. The expert annotations are provided for three structures: right ventricle, myocardium, and left ventricle. It was hosted as part of the MICCAI ACDC Challenge 2017.
(b) ISIC dataset: It is a collection of images dedicated to the challenge of skin lesion segmentation. It contains 2694 images and the corresponding ground truth mask images. This dataset was released by the ISIC and represents a substantial repository of skin disease images.

4.2. Implementation details

Dataset Split: For the ACDC dataset, we used 10 samples for validation, 20 samples for testing, and the remaining 70 samples for training. Within the training set, we conducted a series of experiments with 10% and 20% of the samples used as labeled data. For the ISIC dataset, we employed 270 samples for validation, 538 samples for testing, and the remaining samples for training. Within the training set, we also conducted experiments using 5% and 10% of the samples as labeled data.

Data preprocessing: For the ACDC dataset, we normalized each 3D volume (x) during the generation of 2D slices. Afterward, we enhanced the data by randomly rotating and flipping the 2D slices, and randomly cropping them to a size of 256×256. For the ISIC dataset, we applied random horizontal and vertical flips to the data, followed by a 90-degree rotation, and finally resized it to a size of 256×256.

Parameter settings: We used the SGD optimizer with a learning rate between 0.001 and 0.01, momentum of 0.9, and weight decay factor of 0.0001 for both datasets. During network training, the batch size was set to 8, and each training phase was iterated for 200 epochs. Our method was implemented using PyTorch on Nvidia GeForce RTX 3090 GPU.

Validation metrics: For the ACDC dataset, we used four metrics, including Dice, Jaccard, Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD), to quantitatively evaluate the performance of the model. For the ISIC dataset, we used Dice and MIOU (Mean Intersection over Union) as validation metrics.

5. Results

In this section, we conduct quantitative analyses of our model on two publicly available medical image segmentation datasets. We compare our model's performance with six well-established methods in the field of semi-supervised image segmentation: MT [15], UA-MT [27], DCT-seg [40], URPC [11], DTML [44], ST++ [41], and the most recent work, DUST [18]. Additionally, we perform ablation experiments on two modules proposed in our model. The results presented in the following sections are reproducible under the same conditions.

5.1. Results on the ACDC dataset

Table 1 shows the specific results of the comparative experiment on the ACDC dataset. The first three rows of the table show the prediction accuracy of the U-net model trained with 10%, 20%, and 100% labeled data as the baseline for this comparative study. It is evident that in the case of semi-supervised training using 10% and 20% labeled data, our method has improved the Dice metric from 80.52% to 84.70%, and from 82.83% to 88.07%, which is an average increase of 5% compared to the baseline. Regarding segmentation boundary metrics, the 95HD (Hausdorff Distance) has decreased from 10.27 and 8.54 to 4.52 and 2.24 mm, while the ASD has decreased from 2.97 and 2.51 to 1.32 and

0.58 mm. Furthermore, under the same settings, our proposed method outperforms several recent semi-supervised methods. For example, compared to the traditional self-training method ST++ [41], our model achieves an improvement of approximately 3% in the Dice metric. Compared to methods based on consistency and uncertainty [11], we exhibit significant advantages with Jaccard indices increased from 71.96% to 74.33% and from 76.65% to 79.40%.

Figure 3 shows the visual segmentation results of the proposed method and the comparative method on the ACDC dataset. It can be seen from the figure that some early semi-supervised learning methods (such as MT [15], UA-MT [27], DCT-seg [40]) often produce inconsistent segmentation results. These methods often misclassify background pixels far away from the foreground region into the wrong category, resulting in suboptimal visual segmentation results mainly because they rely on a single consistency loss, which limits their ability to capture image information. Additionally, they lack pixel-level uncertainty estimation, leading to misclassifications in sensitive regions. In contrast, the segmentation results from recent methods (URPC [11], ST++ [41]) demonstrate that effective uncertainty estimation allows the model to accurately identify foreground regions, yielding more coherent and smooth segmentation images with only minor flaws at foreground-background boundaries. Our method combines pseudo-label learning and consistency and estimates the uncertainty of the model, thus solving some problems that exist in the comparison method. From the segmentation results in the figure, the prediction of our model is the closest to the true label.

Table 1. Comparisons of the proposed method with other semi-supervised methods on the ACDC2017 dataset. Reported values are averages and standard deviations for 5 runs with different random seeds.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
U-net	7 (10%)	0	80.52 ± 0.92	69.24 ± 1.12	10.27 ± 2.06	2.97 ± 0.60
U-net	14 (20%)	0	82.83 ± 0.88	72.34 ± 1.20	8.54 ± 1.19	2.51 ± 0.14
U-net	All (70)	0	90.52 ± 0.65	83.12 ± 1.03	1.78 ± 0.34	0.56 ± 0.11
MT (2017) [15]	7 (10%)	63 (90%)	81.53 ± 0.30	70.60 ± 0.44	10.04 ± 1.54	2.95 ± 0.33
UA-MT (2019) [27]			81.80 ± 0.34	70.53 ± 0.39	10.61 ± 1.38	3.14 ± 0.42
DCT-Seg (2020) [40]			82.04 ± 0.44	71.22 ± 0.46	8.23 ± 0.46	2.61 ± 0.15
URPC (2022) [11]			82.49 ± 0.45	71.96 ± 0.18	6.06 ± 0.81	1.94 ± 0.77
DTML (2021) [44]			83.01 ± 0.90	72.41 ± 1.11	5.72 ± 0.93	1.71 ± 0.27
ST++ (2022) [41]			80.14 ± 0.36	68.22 ± 0.56	14.05 ± 2.69	4.19 ± 0.47
DUST [18]			83.44 ± 0.60	72.60 ± 0.86	7.07 ± 1.21	1.94 ± 0.28
Ours			**84.70 ± 0.30**	**74.33 ± 0.41**	**4.52 ± 0.46**	**1.32 ± 0.15**
MT (2017) [15]	14 (20%)	56 (80%)	84.61 ± 0.74	75.09 ± 0.88	9.38 ± 2.81	2.66 ± 0.54
UA-MT (2019) [27]			84.76 ± 0.91	75.34 ± 1.13	7.42 ± 1.32	2.37 ± 0.47
DCT-Seg (2020) [40]			85.26 ± 0.36	75.88 ± 0.62	7.56 ± 0.98	2.29 ± 0.17
URPC (2022) [11]			85.86 ± 0.66	76.65 ± 0.90	5.72 ± 1.20	1.68 ± 0.34
DTML (2021) [44]			85.71 ± 0.42	76.43 ± 0.42	6.14 ± 0.25	1.86 ± 0.08
ST++ (2022) [41]			84.57 ± 0.53	74.45 ± 0.91	8.88 ± 2.03	2.47 ± 0.52
DUST [18]			86.19 ± 0.18	76.75 ± 0.27	5.02 ± 0.84	1.47 ± 0.22
Ours			**88.07 ± 0.16**	**79.40 ± 0.23**	**2.24 ± 0.67**	**0.58 ± 0.17**

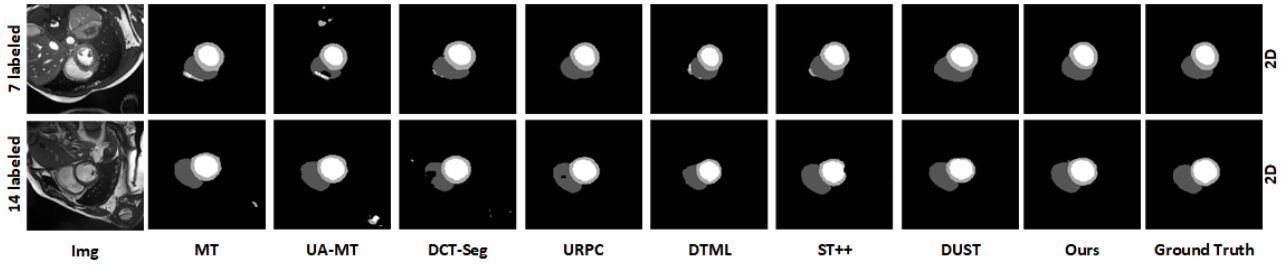

Figure 3. There are the segmentation results of MT [15], UA-MT [27], DCT-Seg [40], URPC [11], DTML [44], ST++ [41], DUST [18], and our proposed method trained on 10% and 20% labeled data on the test images of the ACDC dataset, along with their corresponding ground truth. The black regions represent the background, while different shades of gray represent the three different foregrounds.

5.2. Results on the ISIC dataset

Table 2. Comparisons of the proposed method with other semi-supervised methods on the ISIC2018 dataset. Reported values are averages and standard deviations for 5 runs with different random seeds.

Method	Scans used		Metrics	
	Labeled	Unlabeled	Dice (%)	Jaccard (%)
U-net	5%	0	80.11 ± 0.83	69.19 ± 1.03
U-net	10%	0	82.66 ± 0.69	72.01 ± 0.69
U-net	100%	0	89.96 ± 0.31	80.63 ± 0.38
MT (2017) [15]	5%	95%	82.60 ± 0.85	71.66 ± 0.97
UA-MT (2019) [27]			82.73 ± 0.61	71.84 ± 0.74
DCT-Seg (2020) [40]			82.58 ± 0.81	72.22 ± 0.97
URPC (2022) [11]			82.18 ± 0.46	71.38 ± 0.60
DTML (2021) [44]			83.77 ± 0.58	73.38 ± 0.76
ST++ (2022) [41]			83.45 ± 0.75	72.67 ± 0.63
DUST [13]			84.23 ± 0.57	73.56 ± 0.42
Ours			**85.20 ± 0.38**	**74.25 ± 0.56**
MT (2017) [10]	10%	90%	84.22 ± 0.63	74.00 ± 0.61
UA-MT (2019) [22]			83.74 ± 0.82	73.50 ± 0.92
DCT-Seg (2020) [31]			84.02 ± 0.86	73.78 ± 0.87
URPC (2022) [8]			84.64 ± 1.16	74.61 ± 1.24
DTML (2021) [44]			85.20 ± 0.43	75.17 ± 0.44
ST++ (2022) [32]			85.02 ± 0.76	74.89 ± 0.44
DUST [13]			85.86 ± 0.84	75.33 ± 0.76
Ours			**86.65 ± 0.62**	**75.89 ± 0.70**

Table 2 presents the quantitative experimental results of the proposed method and the comparative methods on the ISIC dataset. Overall, our method improves the segmentation accuracy by approxi-

mately 4% compared to the baseline. When using only 5% labeled data, the Dice score increases from 80.11% to 85.20%, and with 10% labeled data, it improves from 82.66% to 86.52%. In certain random scenarios, our method can approach the performance of fully supervised training. Due to the single foreground involved in the ISIC dataset, only 5% of labeled data is required to achieve 80% segmentation accuracy, and further increasing labeled data has a small impact on model performance. Additionally, through comparative experiments, we can observe that methods based on pseudo-labels (such as DUST [18], ST++ [41]) outperform consistency-based methods (such as MT [15], URPC [11]) on this dataset. Our method outperforms these comparative methods by 1% to 2%, demonstrating the superiority of our framework in utilizing information from unlabeled data.

Figure 4 shows the visual segmentation results of the proposed method and the comparison method on the ISIC dataset. From the figures, it can be observed that when using only 5% labeled data, most semi-supervised methods produce discontinuous and incomplete prediction maps. In contrast, our model generates the most accurate and smooth prediction maps. When using 10% labeled data, our model also obtains prediction maps closest to the ground truth. This further demonstrates the effectiveness of the method we have proposed.

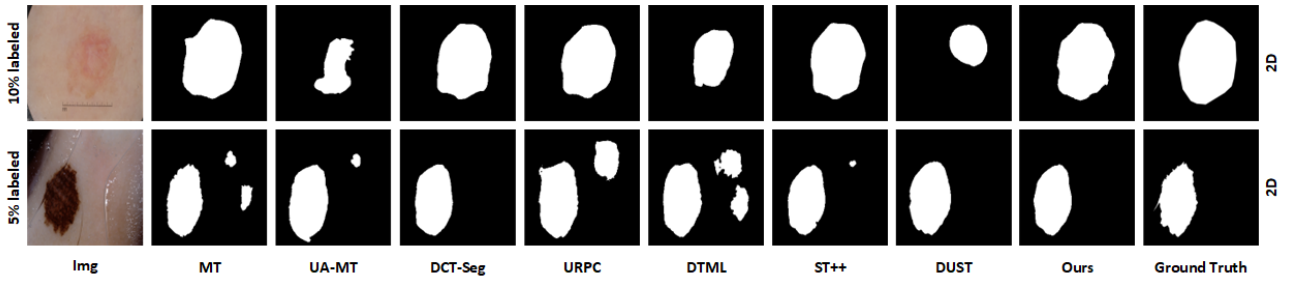

Figure 4. There are the segmentation results of MT [15], UA-MT [27], DCT-Seg [40], URPC [11], DTML [44], ST++ [41], DUST [18], and our proposed method trained on 10% and 5% labeled data on the test images of the ISIC dataset, along with their corresponding ground truth.

5.3. Ablation experiments

In the introduction and method section, we detail the proposed method, which is based on pseudo-label learning. We propose a sample selection module with the SUS module and a multi-model pseudo-label generation module with the PUM module to address the problems in traditional self-training methods. To verify the contribution of the proposed key modules, we conducted ablation experiments on the ACDC dataset.

Table 3 shows the results of ablation experiments. It can be observed that the traditional pseudo-labeling method can utilize unlabeled data to some extent, but due to the uncertainty of the model, iteratively generated pseudo-labels will accumulate noise, resulting in poor and unstable results. After introducing the SUS module, the Dice score increased by 1.5%. In addition, the variance of the results from multiple random experiments indicates that the training process became stable. When the PUM module is introduced separately, the segmentation performance is significantly improved, indicating that the PUM module can effectively reduce the accumulation of noise and generate high-quality pseudo-labels. When the SUS module and the PUM module are simultaneously introduced, all

validation metrics showed significant improvement. Among them, the Dice score increased by 6.5% compared to the baseline, and the boundary distance metrics 95HD and ASD approached the level of full supervision.

Through comparative experiments, we have demonstrated the effectiveness of the proposed modules. Specifically, both the SUS module and the PUM module have contributed to our method, with the PUM module having a more significant impact on improving segmentation performance. Additionally, the results indicate that our approach significantly elevates the level of pseudo-label learning.

Table 3. Ablation analysis on key modules on ACDC2017 dataset. Reported values are averages and standard deviations for 5 runs with different random seeds.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
U-net	14 (20%)	0	82.83 ± 0.88	72.34 ± 1.20	8.54 ± 1.19	2.51 ± 0.14
U-net	All (70)	0	90.52 ± 0.65	83.12 ± 1.03	1.78 ± 0.34	0.56 ± 0.11
PL (Pseudo label) [6]	14 (20%)	56 (80%)	83.50 ± 1.01	73.13 ± 1.30	10.50 ± 3.96	3.10 ± 0.95
PL+SUS			84.90 ± 0.21	74.85 ± 0.27	9.94 ± 0.10	2.95 ± 0.04
PL+PUM			87.30 ± 0.40	78.38 ± 0.53	2.70 ± 0.58	0.78 ± 0.09
PL+SUS+PUM			**88.07 ± 0.16**	**79.40 ± 0.23**	**2.24 ± 0.67**	**0.58 ± 0.17**

5.4. Effects of different sample-level uncertainty estimates

In the method, we mentioned that sample-level uncertainty is obtained by estimating the distance between multiple predictions. Specifically, we perform K dropout operations on the pretrained model to obtain multiple predictions, and we calculate their variance as the uncertainty of unlabeled samples. In Eq (3.1), K is set to 4. To explore the effects of different sample-level uncertainty estimation on experimental results, we conducted a series of comparative experiments on the ACDC2017 dataset.

Table 4. Analysis of sample-level uncertainty estimation methods on ACDC2017 dataset. Reported values are averages and standard deviations for 5 runs with different random seeds.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
U-net	14 (20%)	0	82.83 ± 0.88	72.34 ± 1.20	8.54 ± 1.19	2.51 ± 0.14
U-net	All (70)	0	90.52 ± 0.65	83.12 ± 1.03	1.78 ± 0.34	0.54 ± 0.11
Checkpoints (K = 4)	14 (20%)	56 (80%)	87.70 ± 0.22	78.84 ± 0.27	2.63 ± 0.58	0.72 ± 0.17
Checkpoints (K = 3)			87.76 ± 0.34	78.97 ± 0.45	**2.21 ± 0.44**	0.62 ± 0.14
Dropout (K = 3)			87.93 ± 0.14	79.16 ± 0.23	2.28 ± 0.13	0.66 ± 0.05
Dropout (K = 4)			**88.07 ± 0.16**	**79.40 ± 0.23**	2.24 ± 0.67	**0.58 ± 0.17**
Dropout (K = 5)			88.07 ± 0.20	79.39 ± 0.27	2.31 ± 0.22	0.62 ± 0.06

We conducted experiments under the settings of the checkpoints (K = 3, 4) and the dropout (K = 3, 4, 5), and the experimental results are shown in Table 4. It can be seen that when K = 3, that is, 1/2 epochs, 2/3 epochs, and epochs, the model is a checkpoint and the final training model has better

validation metrics than K = 4. In addition, when we do not use checkpoints but use dropout to obtain multiple outputs and estimate sample-level uncertainty, the model's segmentation metric reaches its highest value.Specifically, when the number of dropout executions (K = 3, 4, 5) has little effects on the performance of the model, the model performs best when K = 4.

Based on the above results, using the model during pretraining as a checkpoint is not the best choice. The model may perform poorly during the early stages of pretraining and cannot produce stable predictions for unlabeled samples. At the same time, using a more stable model as a checkpoint will achieve better results.

5.5. Effects of different teacher models

The proposed PUM module consists of two teacher models and one student model. The teacher model generates pseudo-labels to supervise the student model, and the different momentum of the student model updates the teacher model. In order to explore the effects of different numbers of teacher models and different momentum on the experimental results, we conducted a series of experiments on the ACDC2017 dataset.

From the results in Table 5 (M represents the number of teacher models, and the values in parentheses represent different momentum), it can be seen that the number of integrated teacher models and different momentum have a certain impact on the performance of the model. Specifically, the performance is best when there are two sets of teacher models in the PUM module. Additionally, differences in momentum also lead to slight variations in overall performance.

Table 5. Analysis of different teacher methods on ACDC2017 dataset. Reported values are averages and standard deviations for 5 runs with different random seeds.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
U-net	14 (20%)	0	82.83 ± 0.88	72.34 ± 1.20	8.54 ± 1.19	2.51 ± 0.14
U-net	All (70)	0	90.52 ± 0.65	83.12 ± 1.03	1.78 ± 0.34	0.56 ± 0.11
M = 1 (0.99)	14 (20%)	56 (80%)	87.38 ± 0.24	78.40 ± 0.27	2.78 ± 0.58	1.44 ± 0.49
M = 2 (0.999,0.99)			**88.07 ± 0.16**	**79.40 ± 0.23**	**2.24 ± 0.67**	**0.58 ± 0.17**
M = 2 (0.99,0.9)			88.05 ± 0.16	79.25 ± 0.33	2.32 ± 0.35	0.63 ± 0.10
M = 3 (0.999,0.995,0.99)			87.72 ± 0.14	78.95 ± 0.18	2.58 ± 0.82	0.70 ± 0.23

6. Conclusions

While existing deep learning-based medical image segmentation methods have achieved significant success, they are limited by the need for extensive expert-annotated ground truth data. Semi-supervised methods can be trained on a small amount of labeled data and a large amount of unlabeled data to deal with the scarcity of labeled data and, thus, can be applied to the field of medical image segmentation. Traditional semi-supervised methods typically iteratively generate pseudo-labels for unlabeled data and include both the unlabeled data and their pseudo-labels in training. This allows the model to learn from the unlabeled data to some extent. This article introduced some excellent works in the field, pointed out the shortcomings of traditional pseudo-label learning, and analyzed the shortcomings of some recent

SSL methods. We proposed a dual uncertainty-guided multi-model pseudo-label learning framework for semi-supervised medical image segmentation, which mainly consists of two modules, SUS and PUM. The SUS module prioritizes unlabeled samples based on sample-level uncertainty to obtain a stable and smooth training process; the PUM module integrates multiple teacher and student models and obtains high-quality pseudo-labels based on pixel-level uncertainty. The comparative experiments conducted on two public medical image segmentation datasets have demonstrated the superiority of our proposed method. However, the proposed method has high complexity and a large number of parameters, resulting in heavy time consumption. At the same time, the training results of the proposed method on 3D datasets do not have much advantage compared to the comparison method, which may be due to the noise in the generated 3D data pseudo-labels. Some semi-supervised methods based on consistency regularization often achieve good results on 3D datasets, such as [42–44]. Future work will focus on investigating the feasibility of pseudo-labeling methods on 3D datasets.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the High-level Talents Fund of Hubei University of Technology under grant No.GCRC2020016, Natural Science Foundation of China under grant No. 62201203 and 62306106.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, (2015), 234–241. https://doi.org/10.1007/978331924574428
2. F. Milletari, N. Navab, S. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*, (2016), 565–571. https://doi.org/10.1109/3DV.2016.79
3. L. Qiu, H. Ren, RSegNet: A joint learning framework for deformable registration and segmentation, *IEEE Trans. Autom. Sci. Eng.*, **19** (2021), 2499–2513. https://doi.org/10.1109/TASE.2021.3087868
4. W. Kim, A. Kanezaki, M. Tanaka, Unsupervised learning of image segmentation based on differentiable feature clustering, *IEEE Trans. Image Process.*, **29** (2020), 8055–8068. https://doi.org/10.1109/TIP.2020.3011269

5. W. Lei, Q. Su, T. Jiang, R. Gu, N. Wang, X. Liu, et al., One-shot weakly-supervised segmentation in 3D medical images, *IEEE Trans. Med. Imaging*, **43** (2024), 175–189. https://doi.org/10.1109/TMI.2023.3294975

6. D. H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in *Workshop on Challenges in Representation Learning, ICML*, **3** (2013), 896. https://doi.org/10.1007/978331966185829

7. W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, et al., Semi-supervised learning for network-based cardiac MR image segmentation, in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part II 20*, (2017), 253–260. https://doi.org/10.1007/978-3-030-32248-9_51

8. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, **15** (2014), 1929–1958.

9. S. Chen, G. Bortsova, A. Garcia-Uceda Juarez, G. Van Tulder, M. De Bruijne, Multi-task attention-based semi-supervised learning for medical image segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, (2019), 457–465. https://doi.org/10.1007/978303032248951

10. L. Sun, J. Wu, X. Ding, Y. Huang, G. Wang, Y. Yu, A teacher-student framework for semi-supervised medical image segmentation from mixed supervision, preprint, arXiv:2010.12219. https://doi.org/10.48550/arXiv.2010.12219

11. X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, et al., Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency, *Med. Image Anal.*, **80** (2022), 102517. https://doi.org/10.1016/j.media.2022.102517

12. Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, et al., Mutual consistency learning for semi-supervised medical image segmentation, *Med. Image Anal.*, **81** (2022), 102530. https://doi.org/10.1016/j.media.2022.102530

13. Y. Xie, J. Zhang, Z. Liao, J. Verjans, C. Shen, Y. Xia, Intra-and inter-pair consistency for semi-supervised gland segmentation, *IEEE Trans. Image Process.*, **31** (2021), 894–905. https://doi.org/10.1109/TIP.2021.3136716

14. C. Chen, K. Zhou, Z. Wang, R. Xiao, Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA, *IEEE Trans. Med. Imaging*, **42** (2022), 346–353. https://doi.org/10.1109/TMI.2022.3184675

15. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in *Advances in Neural Information Processing Systems*, **30** (2017).

16. Y. Zhang, R. Jiao, Q. Liao, D. Li, J. Zhang, Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation, *Artif. Intell. Med.*, **138** (2023), 102476. https://doi.org/10.1016/j.artmed.2022.102476

17. K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, et al., Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning, *Med. Image Anal.*, **79** (2022), 102447. https://doi.org/10.1016/j.media.2022.102447

18. Z. Qiu, H. Gan, M. Shi, Z. Huang, Z. Yang, Self-training with dual uncertainty for semi-supervised medical image segmentation, preprint, arXiv:2304.04441. https://doi.org/10.48550/arXiv.2304.04441

19. K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, et al., Fixmatch: Simplifying semi-supervised learning with consistency and confidence, in *Advances in Neural Information Processing Systems*, **33** (2020), 596–608.

20. D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C. A. Raffel, Mixmatch: A holistic approach to semi-supervised learning, in *Advances in Neural Information Processing Systems*, **32** (2019).

21. A. Kurakin, C. Raffel, D. Berthelot, E. D. Cubuk, H. Zhang, K. Sohn, et al., Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring, 2020. Available from: https://research.google/pubs/remixmatch-semi-supervised-learning-with-distribution-matching-and-augmentation-anchoring/.

22. S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, preprint, arXiv:1610.02242. https://doi.org/10.48550/arXiv.1610.02242

23. J. Li, C. Xiong, S. Hoi, Comatch: Semi-supervised learning with contrastive graph regularization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 9475–9484.

24. M. Zheng, S. You, L. Huang, F. Wang, C. Qian, C. Xu, Simmatch: Semi-supervised learning with similarity matching, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 14471–14481.

25. W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, A. G. Wilson, A simple baseline for bayesian uncertainty in deep learning, in *Advances in Neural Information Processing Systems*, **32** (2019).

26. M. N. Rizve, K. Duarte, Y. S. Rawat, M. Shah, In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning, preprint, arXiv:2101.06329. https://doi.org/10.48550/arXiv.2101.06329

27. L. Yu, S. Wang, X. Li, C. W. Fu, P. A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, (2019), 605–613. https://doi.org/10.1007/978303032245867

28. J. Fan, B. Gao, H. Jin, L. Jiang, Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 9947–9956.

29. Z. Shen, P. Cao, H. Yang, X. Liu, J. Yang, O. R. Zaiane, Co-training with high-confidence Pseudo labels for semi-supervised medical image segmentation, preprint, arXiv:2301.04465. https://doi.org/10.48550/arXiv.2301.04465

30. Z. Xu, J. Luo, D. Lu, J. Yan, S. Frisken, J. Jagadeesan, et al., Double-uncertainty guided spatial and temporal consistency regularization weighting for learning-based abdominal registration, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2022), 14–24.

31. J. Zhang, J. Lyu, X. Ma, J. Yan, J. Yang, L. Wan, et al., Uncertainty-driven trajectory truncation for model-based offline reinforcement learning, preprint, arXiv:2304.04660. https://doi.org/10.48550/arXiv.2304.04660

32. X. Wang, Y. Yuan, D. Guo, X. Huang, Y. Cui, M. Xia, et al., SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning, *Med. Image Anal.*, **79** (2022), 102459. https://doi.org/10.1016/j.media.2022.102459

33. Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, et al., Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation, *IEEE Trans. Med. Imaging*, **41** (2021), 608–620. https://doi.org/10.1109/TMI.2021.3117888

34. Y. Zhang, B. Zhou, L. Chen, Y. Wu, H. Zhou, Multi-transformation consistency regularization for semi-supervised medical image segmentation, in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, (2021), 485–489. https://doi.org/10.1109/ICAIBD51990.2021.9459059

35. H. Basak, R. Bhattacharya, R. Hussain, A. Chatterjee, An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation, preprint, arXiv:2202.00677. https://doi.org/10.48550/arXiv.2202.00677

36. H. Basak, Z. Yin, Pseudo-label guided contrastive learning for semi-supervised medical image segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 19786–19797.

37. Y. Bai, D. Chen, Q. Li, W. Shen, Y. Wang, Bidirectional copy-paste for semi-supervised medical image segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 11514–11524.

38. Z. Xu, D. Lu, J. Yan, J. Sun, J. Luo, D. Wei, et al., Category-level regularized unlabeled-to-labeled learning for semi-supervised prostate segmentation with multi-site unlabeled data, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2023), 3–13. https://doi.org/10.1007/97830314390181

39. W. Pan, J. Yan, H. Chen, J. Yang, Z. Xu, X. Li, et al., Human-machine interactive tissue prototype learning for label-efficient histopathology image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2023), 3–13. https://doi.org/10.1007/978303134048252

40. J. Peng, G. Estrada, M. Pedersoli, C. Desrosiers, Deep co-training for semi-supervised image segmentation, *Pattern Recognit.*, **107** (2020), 107269. https://doi.org/10.1016/j.patcog.2020.107269

41. L. Yang, W. Zhuo, L. Qi, Y. Shi, Y. Gao, ST++: Make self-training work better for semi-supervised semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 4268–4277.

42. Y. Shi, Y. Zhang, S. Wang, Competitive ensembling teacher-student framework for semi-supervised left atrium MRI segmentation, preprint, arXiv:2310.13955. https://doi.org/10.48550/arXiv.2310.13955

43. Z. Xu, Y. Wang, D. Lu, X. Luo, J. Yan, Y. Zheng, Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation, *Med. Image Anal.*, **88** (2023), 102880. https://doi.org/10.1016/j.media.2023.102880

44. Y. Zhang, J. Zhang, Dual-task mutual learning for semi-supervised medical image segmentation, in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*, (2021), 548–559. https://doi.org/10.1007/978303088010146

AIMS Press