Computational Biology and Chemistry 119 (2025) 108579

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

DSMT-Net: Dual-student mean teacher network with pixel-level pseudo-label optimization for semi-supervised medical image segmentation

Jun Su [a], Wenlong Sun [a], Bogdan Adamyk [b,*]

[a] School of Computer Science, Hubei University of Technology, Wuhan, 430068, China
[b] Aston Business School, Aston University, Birmingham, B4 7ET, UK

ARTICLE INFO

Keywords:
Semi-supervised learning
Medical image segmentation
Dual student network
Pseudo labels optimization
Mamba architecture
Mean teacher

ABSTRACT

Medical imaging technologies, as essential tools for the precise visualization of internal anatomical structures, play a crucial role in early disease detection and ensuring accurate diagnosis. Recently, semi-supervised learning has become a key strategy in medical image segmentation to reduce reliance on scarce annotated data. However, existing frameworks, such as the Mean Teacher (MT), often suffer from low-quality pseudo-labels and limited robustness due to structural homogeneity and noise amplification in complex medical scenarios. To address these issues, this study presents a novel Dual-Student Mean Teacher Network (DSMT-Net), which enhances performance through a collaborative complementary architecture and pixel-level pseudo-label optimization. First, DSMT-Net combines U-Net and Mamba-UNet as dual students, utilizing the former's local boundary accuracy and the latter's global dependency modeling via a state-space model. Second, a pixel-level pseudo-label enhancement mechanism is introduced, combining pixel-level similarity analysis, adaptive confidence threshold setting, and iterative propagation to improve pseudo-label quality while maintaining structural consistency. Third, a self-supervised contrastive loss is adopted to enforce feature consistency between the dual students, alleviating noise propagation and improving the efficiency of unsupervised learning. Comprehensive evaluations on the ACDC and LA datasets confirm the effectiveness of DSMT-Net, highlighting its substantial capability to lower annotation requirements in medical image segmentation tasks. This provides a robust and scalable framework for semi-supervised learning in medical image segmentation, advancing clinical diagnostic efficiency and accuracy. Our code is available at https://github.com/sunwenlong1/DSMT.git.

1. Introduction

As medical imaging technologies have advanced and the volume of medical images has surged, automated medical image analysis has progressed rapidly. Automated medical image analysis and processing can effectively help physicians improve their efficiency of work and promote rapid advancements in medical care (Bian et al., 2020; Wenderott et al., 2024; Masood et al., 2015). At present, convolutional neural networks (CNNs) have shown outstanding performance in medical image segmentation and various other visual computing tasks (Ronneberger et al., 2015; Milletari et al., 2016). However, their success often requires a large amount of accurately labeled data. For medical image analysis tasks, labeled data in clinical medical imaging are often scarce, and data annotation methods almost always require high-quality pixel or voxel-level annotations from experts (Tajbakhsh et al., 2020). This makes acquiring enough pixel-level labeled medical data both expensive and time-intensive, with labeled training data continuing to be a key limitation for deep models (Jiao et al., 2024). Extracting meaningful information from large quantities of unlabeled medical data and improving the segmentation performance of deep learning models have become a central focus. Consequently, many semi-supervised learning methods for medical image segmentation have been increasingly explored.

In recent years, semi-supervised learning has increasingly emerged as one of the leading approaches for medical image segmentation, owing to its ability to effectively integrate labeled and unlabeled data. Among these, the Mean Teacher (MT) framework (Tarvainen and Valpola, 2017) based on teacher–student networks has shown high robustness and excellent performance. The Mean Teacher framework generates pseudo-labels through the teacher network and guides the learning of the student network, making it possible to utilize unlabeled data. However, this framework still faces issues such as low-quality pseudo-labels and blurry boundaries when handling complex medical

* Corresponding author.
E-mail addresses: sujuncs@hbut.edu.cn (J. Su), sunwenlong@hbut.edu.cn (W. Sun), b.adamyk@aston.ac.uk (B. Adamyk).

https://doi.org/10.1016/j.compbiolchem.2025.108579
Received 16 April 2025; Received in revised form 17 June 2025; Accepted 27 June 2025
Available online 9 July 2025

images (Basak and Yin, 2023). Additionally, due to the complexity of organs or lesions in medical images, relying on a single network to generate pseudo-labels may introduce noise, thus limiting the model's performance improvement (Ke et al., 2019). Therefore, improving the quality of pseudo-label generation and optimizing the teacher–student network structure are key research topics.

To address the problem of low-quality pseudo-labels, previous studies have proposed various label enhancement methods. For instance, one approach improves pseudo-label quality by filtering based on confidence scores (Rizve et al., 2021), specifically by discarding low-confidence areas to reduce noise interference. These methods typically rely on the confidence distribution of model predictions and set a threshold to select the high-confidence regions as reliable pseudo-supervision signals for training unlabeled data. By focusing on high-confidence regions, such methods can effectively reduce error propagation in pseudo-labels and enhance model stability. However, such threshold-based methods are sensitive to the choice of threshold and can suffer from information loss in regions with fuzzy boundaries or complex structures (Li et al., 2022). Another method involves improving pseudo-label generation quality through data augmentation techniques such as contrastive learning or mixed-label strategies (Wu et al., 2023; Chaitanya et al., 2023; Chen et al., 2024). However, these methods still have limitations: on the one hand, label enhancement strategies perform inconsistently in high-noise environments and struggle to adapt to complex situations with different data distributions; on the other hand, existing label enhancement methods often fail to fully consider the role of pixel similarity or global semantic information, leading to poor pseudo-label performance in boundary regions.

In addition to pseudo-label optimization methods, the choice of network architecture is also critical for the performance of semi-supervised learning. In recent years, the classic U-Net model (Ronneberger et al., 2015) has gained significant popularity in medical image segmentation tasks, owing to its strong capacity to capture local features and restore fine-grained details. However, U-Net has certain limitations in capturing global semantic information and handling complex scenarios. To address this, modules based on global sequence modeling (e.g. Transformer) (Vaswani et al., 2017; Xie et al., 2021; Chen et al., 2021a) and state-space models (SSM) (Wang et al., 2023) have been gradually incorporated into medical image segmentation tasks. Among these, the Mamba module (Gu and Dao, 2023) has demonstrated strong adaptability and efficiency by modeling long-range dependencies with a linear complexity. Drawing inspiration from Mamba's success, Vision Mamba (Liu et al., 2024) and Segmamba (Xing et al., 2024) have made notable advancements in performance within the broader vision domain. Moreover, U-Mamba (Ma et al., 2024), Mamba-UNet (Wang et al., 2024), and VM-UNet (Ruan et al., 2024) have showcased the potential of integrating CNNs with the Mamba architecture in visual computing tasks.

In this context, we propose an innovative Dual-Student Mean Teacher Network (DSMT-Net) that combines U-Net (Ronneberger et al., 2015) and Mamba-UNet (Wang et al., 2024) as dual complementary architectures to enhance pseudo-label quality and optimize the efficiency of utilizing unlabeled data. Specifically, the contributions of this paper are as follows:

1. A dual-student network architecture is proposed within the MT framework, employing both U-Net and Mamba-UNet models. U-Net focuses on the precise detection of local boundaries, while Mamba-UNet captures long-range dependencies using state-space models (SSM), enhancing the extraction of global semantic information. The complementary nature of these two models effectively improves the utilization efficiency of unlabeled data and significantly alleviates the performance limitations of using a single network structure.

2. We introduce an innovative dynamic pseudo-label optimization method based on pixel similarity and high-confidence label propagation. By incorporating dynamic thresholds and calculating pixel-level local similarities, high-confidence regions are efficiently selected, and pseudo-labels are propagated iteratively, improving both the quality and stability of pseudo-labels. This mechanism fully considers local structures and global information in images, enabling the model to adapt more effectively to complex scenarios.

3. To further enhance the model's robustness and generalization ability, we incorporate a self-supervised contrastive learning strategy between the dual-student networks. By applying contrastive loss to enforce feature consistency, this strategy not only effectively reduces noise propagation in pseudo-labels but also takes full advantage of the complementarity between different networks, optimizing the learning of unlabeled data.

4. Through experiments on two public datasets, we compare the effectiveness of this method with existing semi-supervised medical segmentation techniques. The results demonstrate its powerful performance and provide convincing evidence of its superiority over current state-of-the-art methods.

2. Related work

2.1. Medical image segmentation

Since the introduction of U-Net (Ronneberger et al., 2015), it has demonstrated excellent performance in medical image segmentation due to its strong ability to capture local features and restore fine-grained details, leading to the development of various U-Net-based methods (Chen et al., 2018; Gu et al., 2020; Chen et al., 2023). Based on U-Net, DenseU-Net (Dong et al., 2019) (a densely connected U-Net), U-Net++(Zhou et al., 2018) (a nested U-Net), and lightweight MobileNet-based UNet networks (Jing et al., 2022) have been developed, along with 3D U-Net (Çiçek et al., 2016) and V-Net (Milletari et al., 2016). Since the introduction of Transformers, some researchers have applied Transformers and self-attention mechanisms to medical image segmentation, achieving advanced results. For example, TransFuse (Zhang et al., 2021) combines CNNs and ViTs to process both local and global information and uses a parallel decoder structure to improve model performance. Swin-UNet (Cao et al., 2022) combines the Swin Transformer with U-Net and proposes the first purely Transformer-based medical image segmentation model. Recently, the introduction of Mamba (Gu and Dao, 2023) has provided strong competition to CNN and Transformer. The Mamba module, through the SSM mechanism, achieves long-range feature modeling with linear complexity. Inspired by Mamba's success, Vision Mamba (Liu et al., 2024) has achieved significant performance improvements in the general vision field. Additionally, U-Mamba (Ma et al., 2024) and Mamba-UNet (Wang et al., 2024) by combining CNNs and Mamba, are beginning to showcase the potential of SSM-based models in medical image segmentation.

2.2. Semi-supervised medical image segmentation

In recent years, as medical image data has grown in scale and annotation costs have risen, semi-supervised learning has garnered significant attention. By integrating a small amount of labeled data with a large volume of unlabeled data, semi-supervised learning reduces the dependence on labeled data while greatly enhancing model performance and generalization capabilities. Consistency regularization is commonly used as a standard approach to achieve this goal. Li et al. (2020) and Basak et al. (2022), where the model's predictions are made consistent under different input perturbations or model states to enhance robustness.

For instance, Y et al. proposed the CCT (Ouali et al., 2020) based on cross-model consistency, which builds a robust semi-supervised segmentation framework through cross-supervision from different sub-models. Li et al. developed the DSST (Li et al., 2024) with a dual-network structure, utilizing consistency constraints between the primary and auxiliary segmentation networks to improve the model's

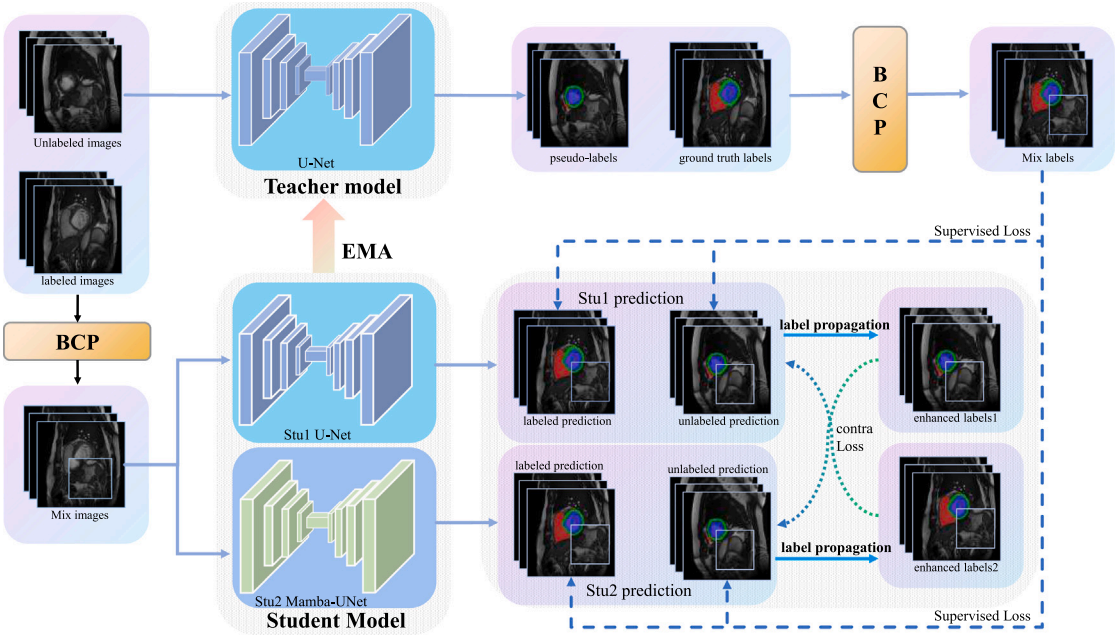

Fig. 1. Overall network architecture. The illustration uses 2D input for better visualization.

utilization of unlabeled data. Additionally, Wu et al. proposed MC-Net (Wu et al., 2021), which employs a mutual learning scheme that transforms the prediction discrepancies between two decoders into unsupervised losses to encourage relative consistency.

The Mean Teacher (Tarvainen and Valpola, 2017) is a classic semi-supervised learning framework. This method introduces a relationship between the teacher model and the student model, imposing a consistency loss to guide the student model's learning of unlabeled data. Based on this, Chen et al. further proposed the MTANS (Chen et al., 2021b), combining a multi-scale consistency strategy to significantly improve model performance in medical image segmentation tasks.

Furthermore, Rizve et al. (2021) enhanced the use of unlabeled data by selecting high-confidence pseudo-labels and integrating spatial enhancement strategies. Lu et al. (2024) significantly improved segmentation accuracy, especially in boundary details, by using hierarchical pseudo-label generation and dynamic pseudo-label correction modules, outperforming existing methods. Several innovative approaches have been developed to further optimize pseudo-label generation and utilization. For instance, BCP (Bai et al., 2023) employs a bidirectional copy-paste strategy to address distribution mismatches between labeled and unlabeled data. DiHC-Net (Koo, 2024) uses diagonal hierarchical consistency and leverages the synergy among multiple sub-models to generate high-quality pseudo-labels.

Building on the Mean Teacher framework, we propose the DSMT-Net. This model extends MT by introducing two structurally different student networks to enhance model performance and robustness. Additionally, to improve pseudo-label quality, we propose an innovative dynamic pseudo-label optimization method based on pixel similarity and high-confidence label propagation. Furthermore, contrastive learning is introduced as a feature consistency constraint between the two student networks to further refine the pseudo-labels, reducing the noise interference during model learning. Through this innovative design, the DSMT-Net effectively improves performance under the scenario of limited labeled samples and enhances the model's learning capacity and robustness for unlabeled data.

3. Method

3.1. Model architecture

In semi-supervised medical image segmentation, a major challenge is effectively leveraging the semantic information in unlabeled data while maintaining a balance between labeled and unlabeled data. Furthermore, the effectiveness of semi-supervised learning algorithms that rely on pseudo-labels is largely influenced by the quality of these pseudo-labels. High-quality pseudo-labels offer more precise supervision signals for unlabeled data, significantly improving the segmentation results. Therefore, our research goals are twofold: first, to fully mine useful information from unlabeled data while maintaining an equilibrium between labeled and unlabeled data; second, to enhance pseudo-label generation through an enhanced process to boost segmentation accuracy.

Inspired by previous studies (Ronneberger et al., 2015; Tarvainen and Valpola, 2017; Wang et al., 2024), we propose a semi-supervised learning framework that combines the teacher–student network architecture, U-Net, and Mamba-UNet models. Fig. 1 illustrates the overall structure of the framework, which is composed of three core parts, each designed and optimized for key challenges in semi-supervised learning.

Specifically, the first part consists of a data augmentation module. To tackle the issue of limited labeled data, we introduce the Bidirectional copy-paste (BCP) algorithm (Bai et al., 2023), which enhances the dataset. BCP performs bidirectional copy-paste operations on labeled and unlabeled images, expanding the training dataset and significantly improving pseudo-label quality. Specifically, BCP conducts region mixing on input images by combining the ground truth labels from the labeled data with the pseudo-labels predicted by the teacher network from the unlabeled data. This mixing operation greatly improves the quality of the mixed labels, and the exchange process can be represented as follows:

$$x_{\text{mix}} = M \otimes x_l + (1 - M) \otimes x_u \tag{1}$$

$$y_{\text{mix}} = M \otimes y_l + (1 - M) \otimes y_p \tag{2}$$

In this context, x_l and x_u represent labeled samples and unlabeled samples, respectively. y_l is the true label of the labeled samples, y_p is the pseudo-label generated by the teacher network, and M is the exchange mask.

Through BCP data augmentation, the accuracy is significantly improved as some pseudo-labels are replaced with true labels, thus providing higher-quality supervision signals to the student networks.

The second part consists of a multi-network collaborative learning architecture composed of a teacher network and two student networks.

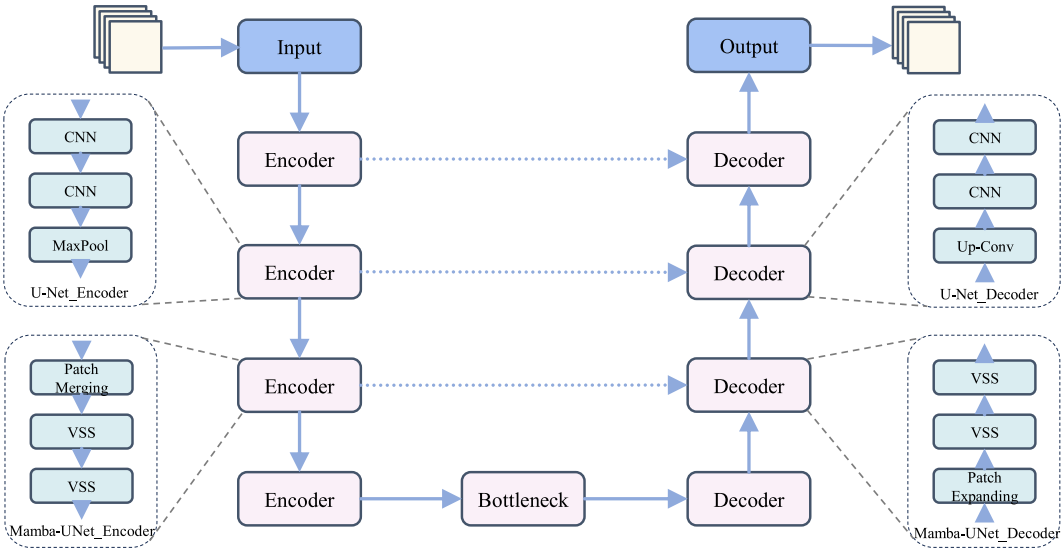

Fig. 2. The encoder and decoder structures of the $f_{S1}(\theta_2)$ and $f_{S2}(\theta_3)$.

We define these networks as $f_T(\theta_1)$, $f_{S1}(\theta_2)$, and $f_{S2}(\theta_3)$, where $f_T(\theta_1)$ represents the teacher network, and $f_{S1}(\theta_2)$, and $f_{S2}(\theta_3)$ represent the student networks. The teacher network uses a classical U-Net model based on CNNs, and its primary task is to generate pseudo-labels for the unlabeled images. The two student networks employ different model structures: $f_{S1}(\theta_2)$ uses the U-Net (Ronneberger et al., 2015), while $f_{S2}(\theta_3)$ uses the Mamba-UNet (Wang et al., 2024), which incorporates the Visual Mamba module. Mamba-UNet enhances the model's global semantic information extraction through state-space modeling (SSM), complementing U-Net's strong boundary detection capabilities. The pseudo-labels generated and augmented by BCP are used to guide the training of the student networks, acting as supervision signals for both labeled and unlabeled data. To further optimize the learning effect of the student networks, we apply not only supervised loss but also contrastive loss. Contrastive loss forces the student networks to maintain consistency in their predictions for the same input, which helps detect and correct prediction biases in the pseudo-labels.

In addition, in semi-supervised learning, the student networks' predictions are often less accurate, which may lead to instability in the contrastive loss process. To address this issue, we propose a mechanism for enhancing the predictions of the student networks based on pixel correlation and high-confidence label propagation. Specifically, we utilize the predictions from the unlabeled images, calculate the pixel-level correlation, and use high-confidence label distributions. For each pixel's prediction, we perform a weighted average based on its similarity with other pixels, generating an enhanced label. This enhanced label is used to optimize the contrastive loss of the student networks, thereby improving the utilization of the unlabeled data.

Next, we will detail the proposed Dual-Student Mean Teacher Network, focusing on the pixel correlation and high-confidence label propagation method, as well as the model loss design.

3.2. Dual-student mean teacher network

As a semi-supervised learning framework, the Mean Teacher method aims to improve the model's generalization performance by establishing a strong interaction between labeled and unlabeled samples during training. The main concept is to integrate the predictions of the unlabeled samples into the model's training process, thereby reinforcing the learning. The key components of the framework include a teacher network and a student network, which share the same network structure. The teacher network updates its parameters using exponential moving averages (EMA), thus providing smoother prediction signals. However, this design may face limitations when handling complex scenarios, especially when the pseudo-labels generated by the teacher network are of low quality. In such cases, the student network is more likely to encounter a "noise amplification" issue, where the low-quality pseudo-labels continuously impact model performance, creating a vicious cycle.

To solve the above issues, we propose a DSMT-Net, which introduces two student networks with different characteristics into the Mean Teacher framework. This significantly enhances the model's diversified learning ability and robustness to unlabeled data. This design makes the model more flexible in utilizing unlabeled data and effectively alleviates the bias and uncertainty caused by a single network structure.

Specifically, the dual-student network adopts different designs: one student network uses the classical U-Net model, with its encoder and decoder both consisting of double convolutional layers (CNNs), which are particularly advantageous in capturing fine-grained boundary information and recovering local features; the second student network uses the Mamba-UNet model based on the Visual Mamba module, which integrates two Visual Mamba blocks that model long-range dependencies using state-space models (SSM). The core mechanism of the Mamba module is represented as:

$$h_t = A \cdot h_{t-1} + B \cdot x_t \tag{3}$$

$$y_t = C \cdot h_t + D \cdot x_t \tag{4}$$

where h_t is the hidden state at time step t, x_t is the input feature, y_t is the output feature, and A, B, C, and D are learnable parameter matrices. The Mamba module achieves long-range feature modeling with linear complexity, which enhances Mamba-UNet's ability to capture global features and perform well in complex structures and blurry boundaries.

Fig. 2 illustrates the different module structures in the two student networks. This modular distinction is key to the method as it directly impacts the network's ability to handle and interpret complex details in medical images. By combining the U-Net and Mamba-UNet architectures, the dual-student network fully leverages their complementary strengths. U-Net focuses on precise boundary detection, while Mamba-UNet strengthens the extraction of global structural information, allowing the model to account for both local details and global consistency in medical image segmentation tasks.

During training, the two student networks generate predictions for the same sample, and the pseudo-labels are enhanced through pixel similarity and high-confidence label propagation mechanisms,

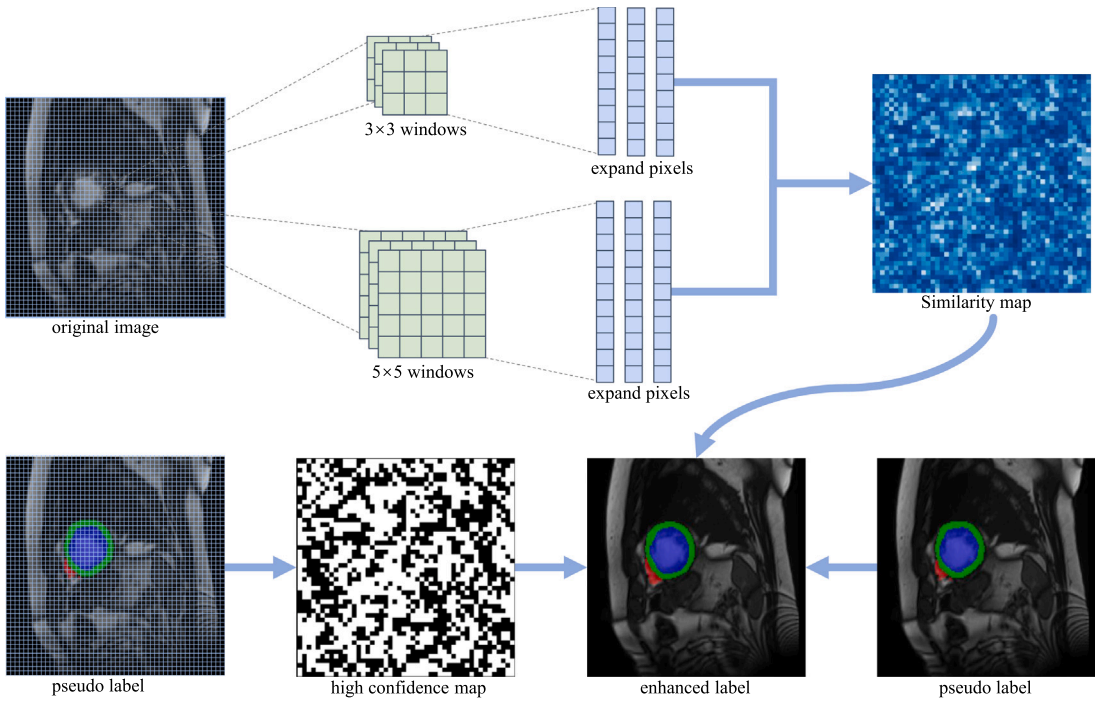

Fig. 3. Pixel-level correlation and high-confidence label propagation mechanism.

improving prediction reliability. The enhanced predictions are then used in the contrastive supervision mechanism between the two student networks. Through contrastive loss, the student networks are forced to maintain consistency in the feature space, which helps to mutually correct prediction biases in the pseudo-labels. This contrastive supervision mechanism not only constrains the differences between the two student networks but also reduces the noise propagation in pseudo-labels, further enhancing the model's robustness and segmentation performance.

3.3. Pixel-level correlation and high-confidence label propagation mechanism

The core challenge of semi-supervised learning using pseudo-labels lies in how to fully utilize the limited labeled data alongside the vast amounts of unlabeled data, with the quality of pseudo-labels being critical to the model's generalization ability. However, traditional pseudo-label generation methods face three main issues: first, noise interference and boundary information loss can lead to inaccuracies in some pseudo-label regions; second, there is a failure to fully leverage high-confidence regions; and third, there is a lack of effective capture of multi-scale features. To address these issues, we propose a new label optimization method from three aspects: pixel-level similarity calculation, high-confidence region filtering, and iterative pseudo-label propagation.

Medical images often have significant local features, such as texture information at organ boundaries or the continuity of anatomical structures. By calculating pixel-level local similarity and analyzing the feature differences between each pixel and its neighboring pixels, we can capture the spatial correlations between pixels, providing stronger support for pixel classification. Furthermore, since the generation of pseudo-labels is affected by model prediction uncertainty, we enhance the pseudo-labels by utilizing a weighted propagation mechanism. This mechanism strengthens the reliable pseudo-labels from high-confidence regions and propagates them to similar pixel areas, thus reducing the risk of noise propagation. As shown in Fig. 3, the method consists of three steps: pixel-level similarity calculation, high-confidence region filtering, and similarity-driven pseudo-label propagation.

3.3.1. Pixel-level similarity calculation

Pixel-level similarity is a core metric for measuring the correlation between a pixel and its neighborhood, serving as the weight for label propagation. We propose a method for calculating local similarity with dynamic adjustment and gradient-aware capabilities using a sliding window. The window size is crucial for similarity calculation, directly affecting the spatial range of the computation: smaller windows are more suited to capturing fine details, while larger windows are better for macro-structural analysis. However, traditional fixed-window sizes are not adaptive to complex regions. Therefore, we use gradient magnitude-based dynamic adjustment to modify the window size according to the feature complexity of different areas. The window size calculation formula is as follows:

$$P = \text{clamp}\left(3 + \left\lfloor 4 \cdot \frac{G_x^2 + G_y^2}{\sqrt{G_x^2 + G_y^2}} \right\rfloor,\ 3,\ 7\right) \tag{5}$$

In this context, G_x and G_y represent the Sobel gradients in the horizontal and vertical directions, respectively.

Next, local neighborhoods are extracted using the expanded operation, and the squared differences of pixel values are calculated to measure the similarity between each pixel and its neighborhood, generating a local similarity map that is then normalized:

$$S_{i,j} = -\frac{1}{P^2}\sum_{k=1}^{P^2}(P_{i,j,k} - I_{i,j})^2 \tag{6}$$

$$S_{i,j}^{norm} = \exp(S_{i,j} - \max(S)) \tag{7}$$

In this context, $P_{i,j,k}$ represents the kth pixel in the neighborhood window, and $I_{i,j}$ is the value of the center pixel.

The normalized similarity map $S_{i,j}^{norm}$ is mapped to the range (0, 1], with higher values indicating smaller differences between the center pixel and its neighbors. This result will serve as the weight for label propagation in the next steps.

3.3.2. High-confidence region filtering

High-confidence region filtering is a crucial step in the pseudo-label propagation process. Its core task is to select the regions with

higher prediction confidence from the output results in order to reduce the interference of pseudo-label noise in unsupervised learning and enhance the model's generalization ability. To solve this problem, we propose a dynamic threshold combined with a smoothing strategy for high-confidence filtering. This approach is elaborated in three aspects: dynamic threshold calculation, smoothing processing, and generation of high-confidence masks.

Traditional fixed-threshold methods struggle to adapt to changes in confidence distribution across different samples, which may either lead to the neglect of high-confidence regions or introduce noisy regions. To address this, we employ a dynamic threshold strategy based on confidence distribution, adjusting the threshold dynamically by calculating the mean and standard deviation of the current prediction confidence. The dynamic threshold calculation formula is as follows:

$$T = \mu \cdot C + \gamma \cdot \sigma_c \tag{8}$$

In this context, T represents the result of the dynamic threshold calculation, $\mu \cdot C$ is the mean of the maximum classification probability for each pixel, σ_c is the standard deviation of the prediction confidence, and γ is the dynamic adjustment factor.

This mechanism dynamically adjusts the filtering threshold to ensure the adaptability of high-confidence regions while suppressing low-confidence areas from affecting label propagation. Compared with fixed-threshold methods, the dynamic threshold approach significantly improves the accuracy and stability of pseudo-label generation.

In addition, applying the dynamic threshold for high-confidence region filtering may cause abrupt transitions at the boundaries where high-confidence and low-confidence regions meet. This can negatively impact pseudo-label propagation. To solve this, we apply a smoothing operation to the high-confidence regions by processing the confidence mask continuously, which results in a smoother transition between high- and low-confidence regions. The smoothing mask calculation is as follows:

$$M_{\text{smooth}} = \sigma\big(\alpha \cdot (\mu \cdot C - T)\big) \tag{9}$$

In this context, M_{smooth} is the smoothed high-confidence mask, $\sigma(x)$ is the Sigmoid function, and α is the smoothing intensity coefficient, which controls the degree of smoothing.

After applying smoothing, the boundary transition of high-confidence regions becomes more natural while the primary confidence information is retained. This significantly reduces boundary errors in pseudo-label propagation.

On the basis of dynamic thresholding and smoothing processing, a high-confidence mask is ultimately generated to guide the weighted propagation of pseudo labels. Firstly, we perform an initial screening of high-confidence regions:

$$M_{\text{binary}} = \begin{cases} 1 & \text{if } \mu \cdot C \geq T \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The final high-confidence mask is generated by combining the initial mask and the smoothed mask:

$$M_{final} = M_{binary} \cdot M_{smooth} \tag{11}$$

The final generated mask M_{final} is used for weight distribution in pseudo-label propagation, indicating the weight each pixel holds during propagation. The closer M_{final} is to 1, the higher the weight it will receive in the label propagation process.

3.3.3. Similarity-driven pseudo-label propagation

Similarity-driven pseudo-label propagation is a key step in further optimizing the quality of pseudo-labels. The core idea is to utilize the local similarity graph $S_{i,j}^{norm}$ and the high-confidence region mask M_{final}. Through multiple iterations of propagation, the pseudo-labels are progressively refined, ensuring that predictions in the unlabeled regions become more accurate and stable.

In the initial propagation stage, pseudo-labels are updated using the predicted results $\hat{y}$, the local similarity graph $S_{i,j}^{norm}$, and the high-confidence region mask M_{final}. A single round of propagation can improve the quality of pseudo-labels, but its impact is often limited to the local neighborhood of the high-confidence regions. To further expand the influence of high-confidence regions, this paper introduces a multi-round iterative propagation mechanism. The goal of iterative propagation is to gradually expand the coverage of label propagation, ensuring that the quality of pseudo-labels is enhanced across the entire image. For each pixel, the label propagation formula is as follows:

$$\hat{y}_{enhanced}^{(k)} = \hat{y}_{enhanced}^{(k-1)} \cdot \left(S_{i,j}^{norm} \cdot M_{final}\right) + \hat{y}_{enhanced}^{(k-1)} \cdot \left(1 - M_{final}\right) \tag{12}$$

In this context, $S_{i,j}^{norm} \cdot M_{final}$ represents the weighted contribution of the high-confidence region to the pseudo-labels. Through the combined effect of the local similarity graph $S_{i,j}^{norm}$ and the high-confidence mask M_{final}, the influence range of high-confidence regions is gradually expanded. $1 - M_{final}$ represents retaining the original pseudo-label for low-confidence regions to prevent the complete loss of information in those areas. $\hat{y}_{enhanced}^{(k-1)}$ denotes the pseudo-labels after the previous round of propagation.

To ensure that the propagation results adhere to the probability distribution constraints, the pseudo-labels after propagation need to undergo Softmax normalization, ensuring that the sum of the class probabilities for each pixel equals 1:

$$\hat{y}_{enhanced}^{(k)} = softmax(\hat{y}_{enhanced}^{(k)}) \tag{13}$$

After multiple rounds of iterative propagation, the optimized pseudo-labels can not only cover the unlabeled regions but also effectively smooth the boundary areas, reducing the noise interference caused by model prediction errors. The enhanced visualization is shown in Fig. 4. The final generated pseudo-labels $\hat{y}_{enhanced}^{(k)}$ can fully utilize the information from the high-confidence regions while providing reliable supervisory signals for the unlabeled areas, offering high-quality pseudo-label support for semi-supervised medical image segmentation tasks.

3.4. Loss function

In this section, we provide a comprehensive explanation of the model's loss function structure. The basic loss function of our model combines Dice loss and Cross-entropy loss. Since our model includes two student networks, $f_{S1}(\theta_2)$, and $f_{S2}(\theta_3)$, the loss function for each student network is divided into two components: one part is the pseudo-supervision loss from the teacher network to the student network, and the other part is the self-supervised contrastive loss between the two student networks. The primary objective during training is to continuously update the network parameters to minimize the total loss. The total loss can be represented as follows:

$$L_{\text{total}} = \alpha(L_{P\text{-sup}}^{1} + \delta L_{\text{contra}}^{1}) + \beta(L_{P\text{-sup}}^{2} + \delta L_{\text{contra}}^{2}) \tag{14}$$

In this context, $L_{P\text{-sup}}^{1}$, L_{contra}^{1}, $L_{P\text{-sup}}^{2}$, and L_{contra}^{2} represent the pseudo-supervision loss for $f_{S1}(\theta_2)$, the contrastive loss for $f_{S1}(\theta_2)$, the pseudo-supervision loss for $f_{S2}(\theta_3)$, and the contrastive loss for $f_{S2}(\theta_3)$, respectively. α and β are weight adjustment factors, where α increases from 0.5 to 1 as training progresses, becoming the dominant network, and β decreases from 0.5 to 0, becoming the auxiliary network. σ is the contrastive loss weight, which controls the influence of the contrastive loss in the total loss.

In each iteration, we update the parameters of $f_{S1}(\theta_2)$ and $f_{S2}(\theta_3)$ using stochastic gradient descent, while the teacher network's parameters are updated by the dominant student network $f_{S1}(\theta_2)$ through EMA.

For the pseudo-supervision loss, since labeled and unlabeled data coexist in the network after the bidirectional copy-paste operation, the pseudo-supervision loss is divided into two parts. As both $f_{S1}(\theta_2)$ and

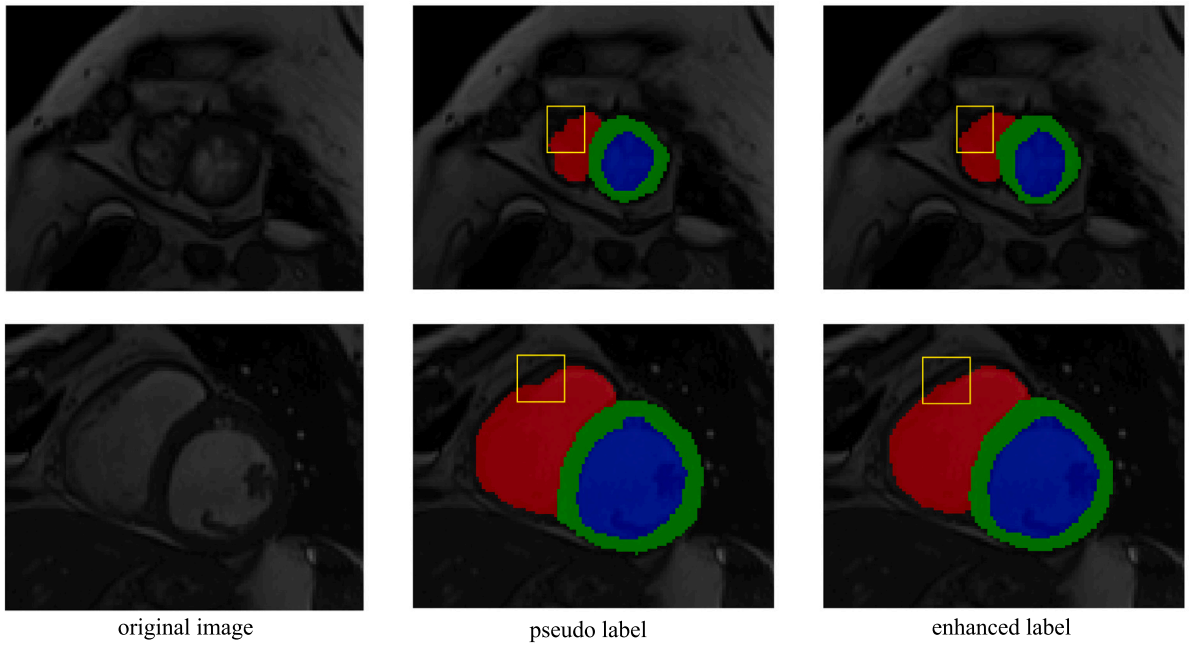

Fig. 4. Label propagation effect, where pseudo label is the pseudo label generated by the student network, and enhanced label is the enhanced pseudo label.

$f_{S2}(\theta_3)$ use the same pseudo-supervision loss, we only list the loss for $f_{S1}(\theta_2)$.

$$L^1_{P\text{-sup}} = L^1_{l\text{-sup}} + L^1_{u\text{-sup}} \tag{15}$$

$$L^1_{l\text{-sup}} = L_{\text{seg}}(x_l, \hat{y}_l) \odot M + \gamma L_{\text{seg}}(x_l, \hat{y}_l) \odot (1 - M) \tag{16}$$

$$L^1_{u\text{-sup}} = L_{\text{seg}}(x_u, \hat{y}_u) \odot M + \gamma L_{\text{seg}}(x_u, \hat{y}_u) \odot (1 - M) \tag{17}$$

In this context, x_l, $\hat{y}_l$, x_u, and $\hat{y}_u$ represent the labeled images, pseudo-labels of labeled images, unlabeled images, and pseudo-labels of unlabeled images, respectively, after the BCP operation. M is the exchange mask in BCP, and γ is used to control the weight of the non-exchanged parts. Following the BCP, γ is set to 0.5.

For self-supervised contrastive loss, its goal is to maintain consistency between the feature representations of the two student networks. Let the outputs of the two student networks on labeled data be p^1_l and p^2_l, and the outputs on unlabeled data be p^1_u and p^2_u. After enhancement through the label propagation mechanism, the outputs are defined as $\tilde{p}^1_l$, $\tilde{p}^2_l$, $\tilde{p}^1_u$, and $\tilde{p}^2_u$. The self-supervised contrastive loss is then defined as:

$$L^1_{\text{contra}} = \frac{1}{2}\left(L_{\text{dice}}(p^1_l, \tilde{p}^2_l) + L_{\text{dice}}(p^1_u, \tilde{p}^2_u)\right) \tag{18}$$

$$L^2_{\text{contra}} = \frac{1}{2}\left(L_{\text{dice}}(p^2_l, \tilde{p}^1_l) + L_{\text{dice}}(p^2_u, \tilde{p}^1_u)\right) \tag{19}$$

By combining pseudo-supervision loss and self-supervised contrastive loss, a fully collaborative optimization is achieved in the utilization of both labeled and unlabeled data. This also maximizes the strengths of the two different models, achieving complementary advantages. The pseudo-supervision loss effectively integrates the features of labeled and unlabeled data through the bidirectional copy-paste operation while dynamically adjusting the weight factors (α, β, σ), balancing the contributions of the dominant network and the auxiliary network at different stages of training. Additionally, the self-supervised contrastive loss, through feature alignment between the student networks, significantly enhances the model's generalization ability and robustness.

4. Experiments

4.1. Dataset

To evaluate the proposed framework, we used two representative datasets in this field and conducted experiments alongside several advanced semi-supervised medical image segmentation methods. The details of the datasets are summarized in Table 1.

The ACDC dataset (Bernard et al., 2018) is MRI scans dataset. It focuses on heart disease diagnosis and segmentation. The dataset includes MRI images of the short-axis heart for 100 patients, covering five types of cardiac conditions: Normal (NOR), Left Ventricular Hypertrophy (HYP), Dilated Cardiomyopathy (DCM), Myocardial Infarction (MINF), and Right Ventricular Dysfunction (RV). The ACDC dataset is a four-class segmentation dataset (background, left ventricle, right ventricle, and myocardium). We split it into 70 cases (1930 slices) for training, 10 for validation, and 20 for testing.

The LA dataset (Xiong et al., 2021) is a medical imaging dataset focused on the segmentation task of the left atrium. It contains 3D MRIs from multi-center scans, all annotated by experts to accurately mark the anatomical structure of the left atrium, providing a reliable benchmark for segmentation and shape analysis tasks. The dataset includes 100 labeled samples, split into 80 (7040 slices) for training and 20 for testing.

4.2. Implementation details

All experiments were run on Ubuntu 22.04 with PyTorch 2.1.0+ CUDA121, using two Nvidia GeForce RTX 3080 GPUs and an Intel Core i5-12400F CPU. The initial learning rate was set to 0.01, decaying by 10% every 2.5 K iterations. The pretraining and self-training iterations were set to 10 K and 30 K, respectively. An evaluation was conducted every 200 iterations, and the network weights were saved if the validation performance surpassed the previous best.

For the ACDC dataset, we extracted 256 × 256 2D slices as input. Prior to feeding the images into the neural network, we applied random flipping and rotation as data augmentation strategies. In the experiment, a small subset of the training data was considered as the labeled training set, with the rest as the unlabeled training set. Specifically,

Table 1
Dataset composition and partition details.

Attribute	ACDC Dataset	LA Dataset
Dataset name	Automated cardiac diagnosis challenge	Left atrium segmentation Challenge
Source	MICCAI 2017 challenge	2018 atrial segmentation challenge
Total scans	100 patients	100 3D MRI scans
Data partition	Training: 70 patients Validation: 10 patients Test: 20 patients	Training: 80 patients Test: 20 patients
Annotated structures	• Left Ventricle (LV) • Myocardium (MYO) • Right Ventricle (RV)	Left Atrium (LA)
Slice range per scan	6–12 slices (per phase)	88–132 slices
Training set slices	1930 slices	7040 slices
Spatial resolution	1.37–1.68 mm^2 (in-plane)	$0.625 \times 0.625 \times 0.625$ mm^3

we considered three different setups: using 5%, 10%, and 20% labeled data for training. In this method, we used U-Net and Mamba-UNet as the segmentation backbones for the two student networks. The batch size was set to 24. During inference, a patch-based pipeline with a sliding window strategy was used, combining predictions from patches to generate the final segmentation output.

On the LA dataset, similar to ACDC, random flipping and rotation were used as data augmentation strategies. The patch size was set to $112 \times 112 \times 80$. As in the ACDC experiments, a small subset of the training data was selected as the labeled training set, with the rest used as the unlabeled training set. Three different setups with 5%, 10%, and 20% labeled data were considered for evaluation. In this method, V-Net and 3D Mamba-UNet were used as the segmentation backbones for the two student networks. The batch size was set to 8.

To verify the performance of the proposed framework, we compared it with previously proposed methods such as MT (Tarvainen and Valpola, 2017), URPC (Luo et al., 2022), SS-Net (Wu et al., 2022b), MC-Net+(Wu et al., 2022a), BCP (Bai et al., 2023), CONTRMIX (Zhang et al., 2024), and DSST (Li et al., 2024). During the evaluation, we adopted four standard metrics for medical image segmentation: Dice, Jaccard, 95HD, and ASD. The Dice coefficient is used to assess the similarity between the predicted and ground truth regions. A higher Dice score signifies improved segmentation accuracy. The Jaccard coefficient quantifies the similarity between two samples; a higher Jaccard score indicates greater similarity between the samples. In the study, both Dice and Jaccard were used to evaluate regional overlap, while 95HD and ASD evaluated boundary overlap. 95HD measures the maximum distance between the predicted and true label surfaces, with a smaller value indicating better prediction accuracy. ASD calculates the average distance between corresponding points on the predicted and ground truth surfaces, with a smaller value indicating better performance.

4.3. Comparison with state-of-the-art methods

4.3.1. Results on the ACDC dataset

Table 2 presents the average results on the ACDC dataset for different models at labeling data usage rates of 5%, 10%, and 20%. It also presents the average results of fully supervised training using U-Net as the backbone network at four different labeling data usage rates. The experimental results show that DSMT-Net delivers significant improvements in performance. With 10% labeled data, the average Dice score increases from 77.34% (fully supervised U-Net) to 88.70%. Additionally, compared to the best-performing models, the DSMT-Net's segmentation results show improvements of 0.51 (10% labels) and 0.63 (20% labels). Notably, with 20% labeled data, DSMT-Net achieves a Dice score of 90.79%, which is close to the fully supervised U-Net (100% labels), whose Dice score is 91.65%. As shown in Table 2, except for the suboptimal scores in 95HD and ASD with 5% labeled data, the DSMT-Net consistently outperforms other methods across various metrics at different labeled data usage rates.

Fig. 5 shows the segmentation masks produced by various semi-supervised methods using 10% labeled data from the ACDC dataset. In the figure, the right ventricle is represented in red, the myocardium in green, and the left ventricle in blue. As clearly demonstrated in Fig. 5, our method achieves more accurate segmentation results compared to the other seven methods, with the generated masks closely resembling the ground truth. It is also worth noting that no additional processing was applied to the segmentation masks after their generation by the model.

Additionally, we compared the Dice scores of different models under varying labeling data usage rates, providing a clear comparison of four methods, as shown in Fig. 6. U-Net was trained in a fully supervised manner using only a small portion of labeled data, without the use of any semi-supervised techniques. It is evident that when only 5% of the data is labeled, U-Net's performance is subpar. However, by incorporating various semi-supervised methods, U-Net's segmentation performance is notably enhanced, demonstrating the effectiveness of semi-supervised learning approaches. Additionally, our method outperforms other semi-supervised techniques in segmentation performance across different labeling data utilization rates.

4.3.2. Results on the LA dataset

To confirm the reliability and scalability of our method, we further evaluated its performance on the LA dataset. The experimental process on the LA dataset was similar to that on the ACDC dataset. Table 3 shows the comparison results between the DSMT-Net and other methods on the LA dataset. In all semi-supervised settings, our method significantly outperforms other state-of-the-art approaches. With 10% labeled data, our method's Dice score improves from 90.29% (second-best) to 91.03% (best), and with 20% labeled data, the Dice score improves from 91.07% (second-best) to 92.12% (best). Notably, DSMT-Net, using only 20% labeled data, already surpasses the segmentation results of the fully supervised V-Net method, which uses 100% labeled data.

Along with the quantitative results, the visual outcomes in Fig. 7 clearly highlight the superior performance of our method, showing the segmentation masks obtained by different methods using 10% labeled data. It is evident from the figure that, compared to other seven methods, our method provides more accurate segmentation results, particularly in handling complex regions.

Furthermore, the results in Fig. 8 are similar to those in Fig. 6, where V-Net was trained fully supervised using only partially labeled data, without incorporating any semi-supervised techniques. It is clearly evident from the figure that DSMT-Net, using only 5% labeled data, already outperforms V-Net, which uses 20% labeled data. Additionally, compared to the other three methods, DSMT-Net consistently achieves significantly better segmentation results at all labeling data rates, highlighting the great potential of our approach.

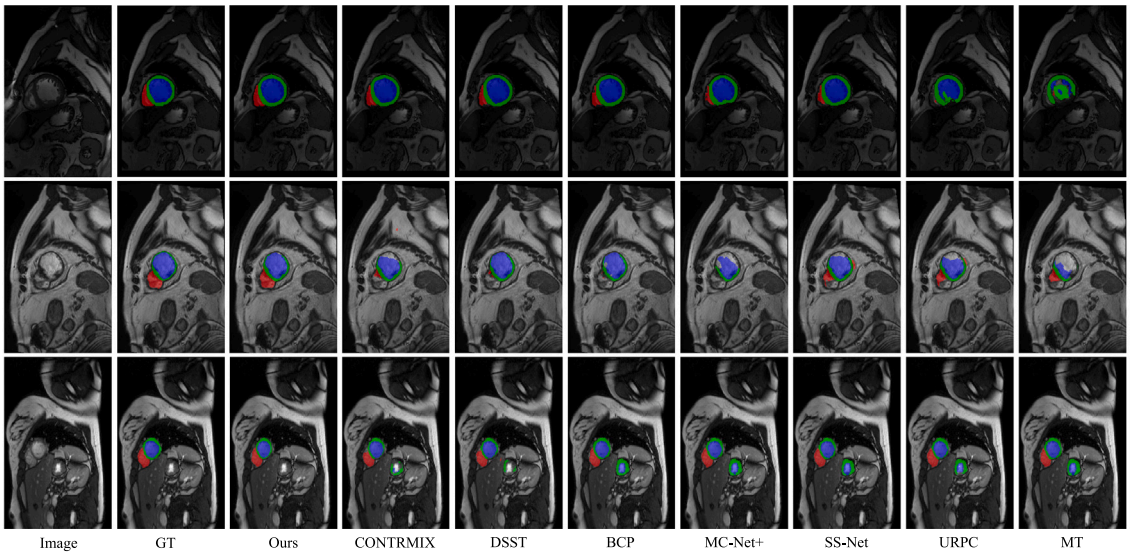

Fig. 5. Visualization of various methods using 10% labeled data on the ACDC dataset.

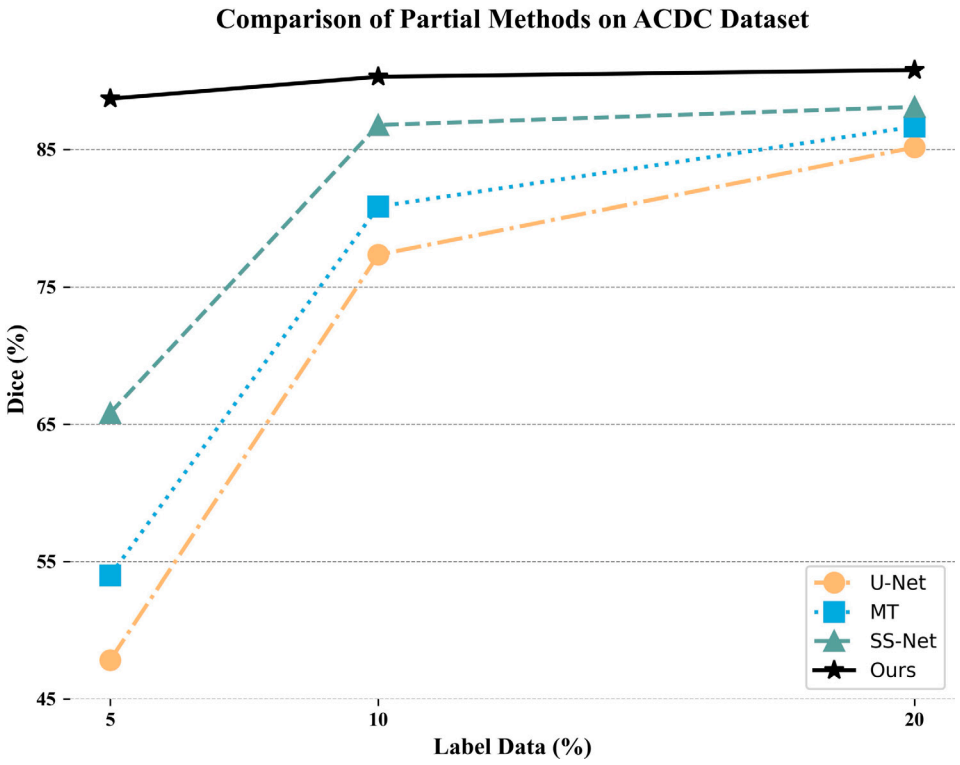

Fig. 6. Compared to supervised and three semi-supervised methods on the ACDC dataset.

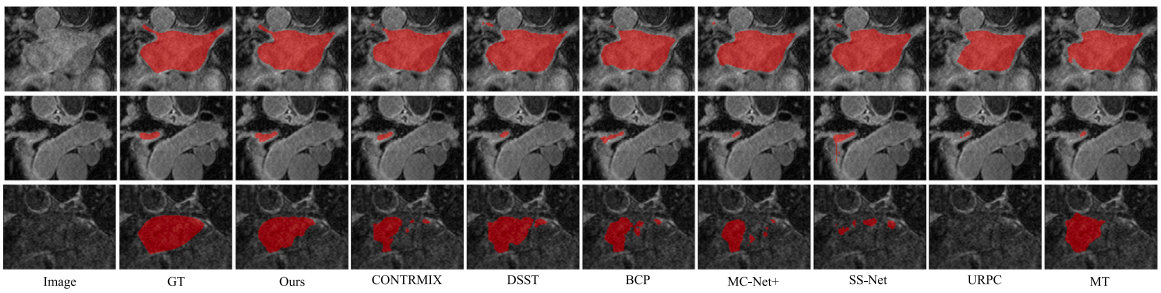

Fig. 7. Visualization of various methods using 10% labeled data on the LA dataset.

Table 2
Results of various methods on the ACDC dataset in labeled 5%, 10%, and 20%.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
U-Net	3(5%)	0	47.83	37.01	31.16	12.62
U-Net	7(10%)	0	77.34	66.20	9.18	2.45
U-Net	14(20%)	0	85.15	75.48	6.20	2.12
U-Net	70(All)	0	91.65	84.93	1.89	0.56
MT	3(5%)	67(95%)	53.98	43.25	14.82	7.63
URPC			55.87	44.64	13.60	3.74
SS-Net			65.82	55.38	6.67	2.28
MC-Net+			64.77	54.06	12.39	3.45
BCP			87.59	78.67	1.90	**0.67**
DSST			87.97	79.14	**2.18**	0.86
CONTRMIX			87.69	78.70	2.47	0.80
Ours			**88.70**	**80.31**	2.38	0.79
MT	7(10%)	63(90%)	80.84	69.71	9.62	2.56
URPC			81.77	70.85	5.04	1.41
SS-Net			86.78	77.67	6.07	1.40
MC-Net+			87.10	78.06	6.68	2.00
BCP			88.84	80.62	3.98	1.17
DSST			88.81	80.55	2.08	1.06
CONTRMIX			89.79	81.98	1.93	0.88
Ours			**90.30**	**82.76**	**1.82**	**0.73**
MT	14(20%)	56(80%)	86.66	73.89	7.51	2.18
URPC			86.07	75.61	6.26	1.77
SS-Net			88.10	79.70	3.34	0.95
MC-Net+			88.51	80.19	5.35	1.54
BCP			89.12	81.03	3.40	0.97
DSST			90.16	82.33	2.58	0.86
CONTRMIX			90.16	82.67	1.56	0.64
Ours			**90.79**	**83.59**	**1.31**	**0.52**

Table 3
Results of various methods on the LA dataset in labeled 5%, 10%, and 20%.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
V-Net	3(5%)	0	52.55	39.60	47.05	9.87
V-Net	7(10%)	0	78.57	66.96	21.20	6.07
V-Net	14(20%)	0	86.96	77.31	11.85	3.22
V-Net	70(All)	0	91.71	84.76	5.57	1.59
MT	3(5%)	67(95%)	83.76	74.08	10.88	3.16
URPC			82.48	71.35	14.65	3.65
SS-Net			86.33	76.15	9.97	2.31
MC-Net+			82.07	70.38	20.49	5.72
BCP			88.02	78.72	7.90	2.15
DSST			89.57	79.76	**7.53**	2.04
CONTRMIX			88.02	78.80	7.94	2.17
Ours			**90.04**	**80.88**	8.02	**1.96**
MT	7(10%)	63(90%)	88.49	79.55	9.13	2.68
URPC			85.01	74.36	15.37	3.96
SS-Net			88.43	78.43	7.95	2.55
MC-Net+			88.96	80.25	7.93	1.86
BCP			89.62	81.31	6.81	1.76
DSST			90.29	82.42	6.32	1.78
CONTRMIX			89.69	81.42	7.04	1.89
Ours			**91.03**	**83.55**	**6.13**	**1.46**
MT	140(20%)	56(80%)	89.76	81.53	7.90	2.32
URPC			88.74	79.93	12.73	3.66
SS-Net			89.86	81.70	7.01	1.87
MC-Net+			91.07	83.67	5.84	1.87
BCP			90.34	82.50	6.75	1.77
DSST			91.04	83.63	5.75	1.59
CONTRMIX			90.42	81.82	6.96	1.58
Ours			**92.12**	**85.67**	**5.31**	**1.47**

4.4. Ablation studies

To assess the influence of different factors on the overall performance of DSMT-Net, we performed a series of ablation experiments in this section. We first examined the impact of different combinations of segmentation networks on performance, followed by an investigation into the effects of pixel-level correlation and high-confidence label propagation mechanism iterations. To validate the effectiveness of our proposed label propagation strategy, we conducted a comparison with existing label enhancement methods. We then analyzed the role of the contrastive loss weight σ, and evaluated the model's robustness to noise in unlabeled data, simulating real-world imperfect annotations.

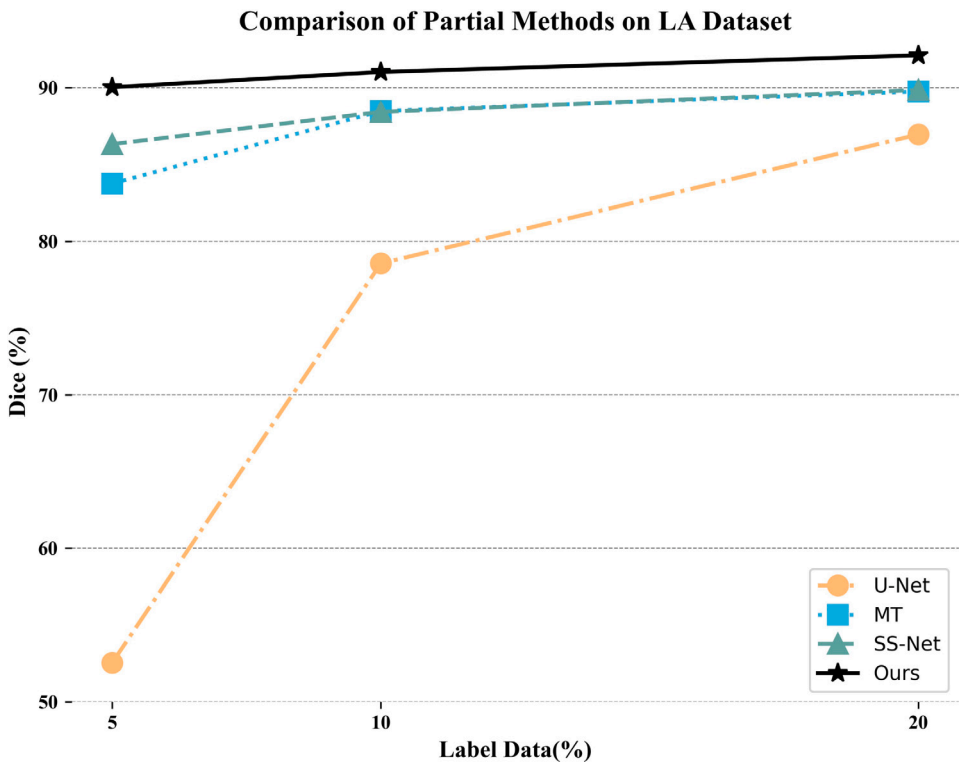

Fig. 8. Compared to supervised and three semi-supervised methods on the LA dataset.

Table 4
The impact of different combinations of segmentation networks on performance using 10% labeled data in the ACDC dataset.

Teacher model	Students model	Label ratio	Metrics			
			Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
U-Net	U-Net + U-Net	10%	88.92	80.79	2.06	0.96
	U-Net + Swin-UNet		89.48	81.87	1.77	0.84
	U-Net + Mamba-UNet		**90.30**	**82.76**	**1.82**	**0.73**
Swin-UNet	Swin-UNet + U-Net	10%	87.87	79.69	2.40	1.26
	Swin-UNet + Swin-UNet		86.82	78.65	4.56	1.51
	Swin-UNet + Mamba-UNet		87.12	79.45	2.98	1.39
Mamba-UNet	Mamba-UNet + U-Net	10%	88.78	80.40	2.16	1.32
	Mamba-UNet + Swin-UNet		87.01	79.23	3.62	1.78
	Mamba-UNet + Mamba-UNet		86.76	78.22	4.70	1.58

Furthermore, we assessed the contribution of each proposed model component. Finally, an inference efficiency analysis was performed to evaluate the practical deployability of our model. Throughout these experiments, individual components were modified while all other variables were strictly controlled, ensuring both the accuracy and fairness of the results. This approach provides clearer insights into the contribution of each part and highlights their synergistic effects within the model.

4.4.1. Impact of different combinations of segmentation networks on performance

To assess the various combinations of segmentation networks in DSMT-Net, we conducted ablation experiments using various network configurations on both the ACDC and LA datasets with 10% labeled data. Specifically, we used three popular models in medical image segmentation: the CNN-based U-Net, the Transformer-based Swin-Unet, and the Mamba-based Mamba-UNet. These models were combined in different ways, resulting in three main configurations. For each configuration, we tested the models as the teacher network and main student network while also using each model sequentially as the auxiliary student network.

The results, clearly shown in Tables 4 and 5, indicate that using different models for the two student networks yields better performance than using the same model for both. The optimal performance is achieved when U-Net is used as both the teacher network and the main student network, with Mamba-UNet acting as the auxiliary student network. This combination takes advantage of the strengths of U-Net for local feature learning and Mamba-UNet's global feature learning, leading to optimal segmentation results.

The differences between the networks enable multi-view learning, where the advantages of local and global feature learning complement each other, ultimately improving the segmentation performance. These results highlight the effectiveness of combining diverse network architectures in the DSMT-Net to enhance segmentation accuracy.

4.4.2. Impact of pixel-level correlation and high-confidence label propagation mechanism iterations on performance

To assess the impact of the pixel-level correlation and high-confidence label propagation mechanism iterations on pseudo-label generation quality and model segmentation performance, we designed a series of ablation experiments. These experiments were conducted on

Table 5
The impact of different combinations of segmentation networks on performance using 10% labeled data in the LA dataset.

Teacher model	Students model	Label ratio	Metrics			
			Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
U-Net	U-Net + U-Net	10%	90.31	82.67	6.82	1.30
	U-Net + Swin-UNet		90.59	83.18	6.77	1.41
	U-Net + Mamba-UNet		**91.03**	**83.55**	**6.13**	**1.46**
Swin-UNet	Swin-UNet + U-Net	10%	90.48	82.99	7.19	1.58
	Swin-UNet + Swin-UNet		88.66	81.10	7.90	2.08
	Swin-UNet + Mamba-UNet		89.43	82.48	6.75	1.48
Mamba-UNet	Mamba-UNet + U-Net	10%	90.87	83.23	6.45	1.44
	Mamba-UNet + Swin-UNet		88.70	81.31	7.76	1.87
	Mamba-UNet + Mamba-UNet		88.62	80.88	7.97	2.48

Table 6
The impact of different label propagation iterations on overall performance using 10% labeled data on two datasets.

Dataset	Iterations	Label ratio	Metrics			
			Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
ACDC	1	10%	88.82	80.59	3.15	1.14
	2		90.11	82.68	2.08	0.89
	3		**90.30**	**82.76**	**1.82**	**0.73**
	4		89.98	82.48	2.29	1.02
	5		89.32	81.69	2.40	1.11
LA	1	10%	89.78	81.66	6.92	1.72
	2		90.96	83.48	6.24	1.48
	3		**91.03**	**83.55**	**6.13**	**1.46**
	4		90.42	81.98	6.94	1.58
	5		89.88	82.50	6.74	1.58

Table 7
Comparison of different label enhancement strategies using 10% labeled data in the ACDC dataset.

Model	Label optimization	Label ratio	Metrics			
			Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
DSMT-Net	w/o Label optimization	10%	87.48	78.66	3.35	1.22
	SASSNet		88.02	79.12	2.91	1.07
	HPLG+DPLC		88.45	80.06	2.68	1.15
	DiHc-Net		89.02	80.44	3.01	0.95
	Ours		**90.30**	**82.76**	**1.82**	**0.73**

both the ACDC and LA datasets with 10% labeled data. We tested the performance of 1, 2, 3, 4, and 5 iterations of label propagation while keeping other experimental parameters unchanged. The experimental results are shown in Table 6. It can be observed that when the number of iterations increases to 3, all metrics achieve their highest values. Further increasing the number of iterations leads to diminishing returns and may even result in slight performance degradation. This is because too many iterations could introduce cumulative noise in the pseudo-labels, negatively impacting the model's generalization ability.

4.4.3. Comparison with existing label enhancement methods

To further demonstrate the effectiveness of our proposed pixel-level label optimization module, we conducted a comprehensive ablation study by replacing it with several representative label enhancement techniques. Specifically, we evaluated the following alternatives: (1) without Label Optimization, where our refinement module is removed entirely; (2) SASSNet (Rizve et al., 2021), which applies shape-aware consistency constraints and uses sharpened predictions as pseudo-labels; (3) HPLG+DPLC (Lu et al., 2024), which employs hierarchical pseudo-label generation guided by uncertainty and a dual-path learning strategy; and (4) DiHC-Net (Koo, 2024), which enhances pseudo-label quality through inter-pixel consistency and hard sample filtering. For fairness, all methods were evaluated under the same DSMT dual-student framework with consistent training settings.

As shown in Table 7, all tested methods benefit from incorporating some form of label enhancement. However, our proposed pixel-level optimization achieves the highest Dice score and boundary accuracy. Compared with DiHC-Net, the best-performing baseline, our method still gains a 1.28% improvement in Dice and significantly reduces boundary errors. This demonstrates that fine-grained spatial refinement and confidence-aware propagation in our method are more effective than existing region- or sample-based label strategies, especially for capturing anatomical details.

4.4.4. Effect of contrastive loss weight

To verify the role of the contrastive loss weight σ in improving model consistency and unsupervised learning performance, we designed ablation experiments to examine the effect of different values of σ on model performance. Experiments were conducted on both the ACDC and LA datasets using 5%, 10%, and 20% labeled data. We adjusted the contrastive loss weight σ and analyzed its impact with values of σ = 0,0.25,0.5,1.0. Notably, when $\sigma = 0$, only pseudo-supervision loss is used, and the model degenerates into the Mean Teacher framework, relying solely on the teacher network's guidance for the student networks. The results, shown in Fig. 9, indicate that as σ increases, the contrastive loss significantly improves the model's generalization ability. When $\sigma = 0.5$, the Dice coefficient reaches its highest value. However, as σ increases to 1.0, the contrastive loss dominates the training, potentially causing over-amplification of pseudo-label noise, which can lead to instability in the training process.

4.4.5. Robustness to unlabeled data noise

To simulate varying noise conditions that may occur in real-world clinical scenarios, we introduced zero-mean Gaussian noise with standard deviations $\sigma \in \{0.01, 0.05, 0.1\}$ into the unlabeled input images in the ACDC dataset. This setup enables us to evaluate the resilience

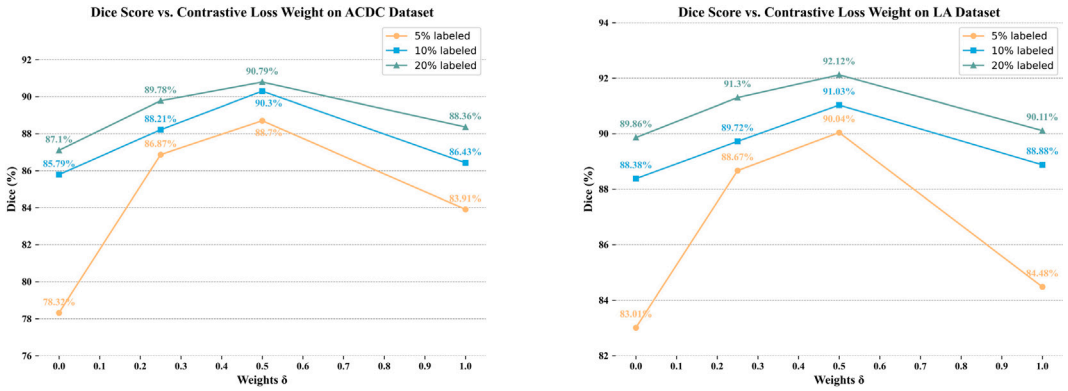

Fig. 9. Impact of contrastive loss weight on overall performance.

Table 8
Robustness analysis of the model under different unlabeled data noise levels using 10% labeled data in the ACDC dataset.

Method	Noise level	Label ratio	Metrics			
			Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
MT	0.00	10%	80.84	69.71	9.62	2.56
	0.01		79.61	68.50	10.87	2.81
	0.05		77.03	65.86	12.43	3.42
	0.10		76.58	64.97	13.12	3.81
DSMT-Net	0.00	10%	**90.30**	**82.76**	**1.82**	**0.73**
	0.01		89.84	82.18	2.01	0.77
	0.05		88.92	82.01	2.34	0.80
	0.10		88.75	81.12	2.68	0.86

Table 9
Ablation results on the ACDC dataset using 10% of labeled data to evaluate the contribution of each component of the proposed method.

Module			Label ratio	Metrics			
DS	PHLP	BCP		Dice (%)↑	Jaccard (%)↑	95HD (voxel)↓	ASD (voxel)↓
			10%	80.84	69.71	9.62	2.56
✓				85.26	78.55	3.87	1.35
✓	✓			89.66	81.23	2.01	0.89
✓	✓	✓		**90.30**	**82.76**	**1.82**	**0.73**

of our model to different levels of noise contamination in unlabeled data, which is common in practical medical imaging pipelines due to variability in scanners, acquisition protocols, and patient motion.

As shown in Table 8, the performance of the baseline Mean Teacher (MT) model degraded significantly with increasing noise, experiencing a Dice score drop of 4.26% at $\sigma = 0.1$. In contrast, our DSMT-Net exhibited superior robustness, with only a minor reduction of 1.55% in Dice score under the same noise level. This stability can be attributed to two key design choices: (1) the use of a dual-student architecture combining U-Net and Mamba-UNet, which enables multi-perspective feature extraction and mitigates noise amplification from a single model; and (2) the proposed pixel-level pseudo-label refinement mechanism that leverages high-confidence regions and similarity-driven label propagation, effectively filtering out unreliable supervision signals.

These findings underscore the robustness of DSMT-Net in handling noisy unlabeled data and its practical suitability for deployment in clinical environments, where data quality may be suboptimal or inconsistent.

4.4.6. Proposed model components

To evaluate the contribution of each module to the overall performance, we conducted experiments on the ACDC dataset using 10% labeled data, as shown in Table 9. Here, DS, PHLP, and BCP represent the Dual-Student Mean Teacher network, the Pixel-Level Correlation and High-Confidence Label Propagation Mechanism, and the Bidirectional Copy-Paste algorithm, respectively. The first row of the table shows the performance of the classic MT method without all the modules. By systematically adding each component and observing its impact on model performance, we can assess the contribution of each module. The results show that each individual component in DSMT significantly improves model performance. Furthermore, when these modules work synergistically, the model achieves optimal performance.

Table 10
Inference time per 2D slice on the ACDC dataset using 10% of labeled data.

Model	Inference time (ms)	Dice (%)↑
U-Net	12.3	77.34
MT	14.5	80.84
DSMT-Net	19.2	**90.30**

4.4.7. Inference efficiency analysis

To assess the practicality of deploying DSMT-Net in real-world clinical settings, we further measured the average inference time per image slice. All evaluations were performed on an NVIDIA RTX 3080 GPU with batch size 1. The results are shown in Table 10, DSMT-Net requires approximately 19.2 ms per slice, compared to 12.3 ms for U-Net and 14.5 ms for the classical Mean Teacher baseline. Despite the

modest increase in inference time, DSMT-Net delivers significant performance gains, indicating its potential for clinical deployment where segmentation accuracy is critical.

5. Conclusion

In this research, we introduce a novel Dual-Student Mean Teacher Network (DSMT-Net) for semi-supervised medical image segmentation. By combining the strengths of U-Net and Mamba-UNet models within a dual-student framework, our method effectively addresses the challenges of pseudo-label generation and the utilization of unlabeled data. The key innovations of DSMT-Net include a dynamic pseudo-label optimization strategy based on pixel similarity and high-confidence label propagation, as well as the incorporation of self-supervised contrastive learning to enforce feature consistency between the dual students. Our extensive experimental results on two publicly available datasets, ACDC and LA, demonstrate the advantages of DSMT-Net compared to current state-of-the-art semi-supervised methods. Specifically, DSMT-Net achieves significant improvements in segmentation performance, particularly with a limited amount of labeled data, reducing the reliance on expensive annotation efforts. The performance enhancements are attributed to the synergistic effects of the complementary architectures (U-Net and Mamba-UNet), the effective utilization of dynamic pseudo-label optimization, and the incorporation of contrastive learning for improved generalization and robustness. The results emphasize DSMT-Net's potential as a robust and scalable framework for medical image segmentation, offering substantial advantages in clinical applications where labeled data is limited. By improving the efficiency of semi-supervised learning, our framework not only paves the way for more accurate segmentation models but also contributes to advancing the clinical diagnostic process. Future work will focus on refining the dynamic pseudo-label optimization mechanism and exploring its integration with other advanced network architectures for broader medical image applications.

CRediT authorship contribution statement

Jun Su: Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wenlong Sun:** Writing – review & editing, Validation, Software, Methodology, Formal analysis, Data curation. **Bogdan Adamyk:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bai, Yunhao, Chen, Duowen, Li, Qingli, Shen, Wei, Wang, Yan, 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11514–11524.

Basak, Hritam, Bhattacharya, Rajarshi, Hussain, Rukhshanda, Chatterjee, Agniv, 2022. An exceedingly simple consistency regularization method for semi-supervised medical image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–4.

Basak, Hritam, Yin, Zhaozheng, 2023. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19786–19797.

Bernard, Olivier, Lalande, Alain, Zotti, Clement, Cervenansky, Frederick, Yang, Xin, Heng, Pheng-Ann, Cetin, Irem, Lekadir, Karim, Camara, Oscar, Ballester, Miguel Angel Gonzalez, et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

Bian, Cheng, Yuan, Chenglang, Wang, Jiexiang, Li, Meng, Yang, Xin, Yu, Shuang, Ma, Kai, Yuan, Jin, Zheng, Yefeng, 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. Med. Image Anal. 64, 101732.

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

Chaitanya, Krishna, Erdil, Ertunc, Karani, Neerav, Konukoglu, Ender, 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med. Image Anal. 87, 102792.

Chen, Liang, Bentley, Paul, Mori, Kensaku, Misawa, Kazunari, Fujiwara, Michitaka, Rueckert, Daniel, 2018. Drinet for medical image segmentation. IEEE Trans. Med. Imaging 37 (11), 2453–2462.

Chen, Jieneng, Lu, Yongyi, Yu, Qihang, Luo, Xiangde, Adeli, Ehsan, Wang, Yan, Lu, Le, Yuille, Alan L, Zhou, Yuyin, 2021a. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, Gaoxiang, Ru, Jintao, Zhou, Yilin, Rekik, Islem, Pan, Zhifang, Liu, Xiaoming, Lin, Yezhi, Lu, Beichen, Shi, Jialin, 2021b. MTANS: multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation. NeuroImage 244, 118568.

Chen, Yuanbin, Wang, Tao, Tang, Hui, Zhao, Longxuan, Zong, Ruige, Chen, Shun, Tan, Tao, Zhang, Xinlin, Tong, Tong, 2023. Dual-decoder consistency via pseudo-labels guided data augmentation for semi-supervised medical image segmentation. arXiv preprint arXiv:2308.16573.

Chen, Yifei, Zhang, Chenyan, Ke, Yifan, Huang, Yiyu, Dai, Xuezhou, Qin, Feiwei, Zhang, Yongquan, Zhang, Xiaodong, Wang, Changmiao, 2024. Semi-supervised medical image segmentation method based on cross-pseudo labeling leveraging strong and weak data augmentation strategies. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Çiçek, Özgün, Abdulkadir, Ahmed, Lienkamp, Soeren S, Brox, Thomas, Ronneberger, Olaf, 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp. 424–432.

Dong, Rongsheng, Pan, Xiaoquan, Li, Fengying, 2019. Denseu-net-based semantic segmentation of small objects in urban remote sensing images. IEEE Access 7, 65347–65356.

Gu, Albert, Dao, Tri, 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gu, Ran, Wang, Guotai, Song, Tao, Huang, Rui, Aertsen, Michael, Deprest, Jan, Ourselin, Sébastien, Vercauteren, Tom, Zhang, Shaoting, 2020. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40 (2), 699–711.

Jiao, Rushi, Zhang, Yichi, Ding, Le, Xue, Bingsen, Zhang, Jicong, Cai, Rong, Jin, Cheng, 2024. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Comput. Biol. Med. 169, 107840.

Jing, Junfeng, Wang, Zhen, Rätsch, Matthias, Zhang, Huanhuan, 2022. Mobile-Unet: An efficient convolutional neural network for fabric defect detection. Text. Res. J. 92 (1–2), 30–42.

Ke, Zhanghan, Wang, Daoye, Yan, Qiong, Ren, Jimmy, Lau, Rynson WH, 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6728–6736.

Koo, Heejoon, 2024. Diagonal hierarchical consistency learning for semi-supervised medical image segmentation. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 1–4.

Li, Boliang, Wang, Yan, Xu, Yaming, Wu, Chen, 2024. Dsst: A dual student model guided student–teacher framework for semi-supervised medical image segmentation. Biomed. Signal Process. Control. 90, 105890.

Li, Hengduo, Wu, Zuxuan, Shrivastava, Abhinav, Davis, Larry S, 2022. Rethinking pseudo labels for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 1314–1322.

Li, Xiaomeng, Yu, Lequan, Chen, Hao, Fu, Chi-Wing, Xing, Lei, Heng, Pheng-Ann, 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 523–534.

Liu, Yue, Tian, Yunjie, Zhao, Yuzhong, Yu, Hongtian, Xie, Lingxi, Wang, Yaowei, Ye, Qixiang, Jiao, Jianbin, Liu, Yunfan, 2024. Vmamba: Visual state space model. Adv. Neural Inf. Process. Syst. 37, 103031–103063.

Lu, Yayuan, Zhang, Zheyu, Zhang, Yueyi, Ou, Yunwei, Sun, Xiaoyan, 2024. Semi-supervised medical image segmentation via dynamic pseudo-label refinement. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Luo, Xiangde, Wang, Guotai, Liao, Wenjun, Chen, Jieneng, Song, Tao, Chen, Yinan, Zhang, Shichuan, Metaxas, Dimitris N, Zhang, Shaoting, 2022. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Med. Image Anal. 80, 102517.

Ma, Jun, Li, Feifei, Wang, Bo, 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

Masood, Saleha, Sharif, Muhammad, Masood, Afifa, Yasmin, Mussarat, Raza, Mudassar, 2015. A survey on medical image segmentation. Curr. Med. Imaging 11 (1), 3–14.

Milletari, Fausto, Navab, Nassir, Ahmadi, Seyed-Ahmad, 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. 3DV, Ieee, pp. 565–571.

Ouali, Yassine, Hudelot, Céline, Tami, Myriam, 2020. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684.

Rizve, Mamshad Nayeem, Duarte, Kevin, Rawat, Yogesh S, Shah, Mubarak, 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329.

Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. 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. Springer, pp. 234–241.

Ruan, Jiacheng, Li, Jincheng, Xiang, Suncheng, 2024. Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491.

Tajbakhsh, Nima, Jeyaseelan, Laura, Li, Qian, Chiang, Jeffrey N, Wu, Zhihao, Ding, Xiaowei, 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693.

Tarvainen, Antti, Valpola, Harri, 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. 30.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Ziyang, Zheng, Jian-Qing, Zhang, Yichi, Cui, Ge, Li, Lei, 2024. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079.

Wang, Jue, Zhu, Wentao, Wang, Pichao, Yu, Xiang, Liu, Linda, Omar, Mohamed, Hamid, Raffay, 2023. Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397.

Wenderott, Katharina, Krups, Jim, Zaruchas, Fiona, Weigl, Matthias, 2024. Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis. NPJ Digit. Med. 7 (1), 265.

Wu, Yicheng, Ge, Zongyuan, Zhang, Donghao, Xu, Minfeng, Zhang, Lei, Xia, Yong, Cai, Jianfei, 2022a. Mutual consistency learning for semi-supervised medical image segmentation. Med. Image Anal. 81, 102530.

Wu, Yicheng, Wu, Zhonghua, Wu, Qianyi, Ge, Zongyuan, Cai, Jianfei, 2022b. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 34–43.

Wu, Yicheng, Xu, Minfeng, Ge, Zongyuan, Cai, Jianfei, Zhang, Lei, 2021. Semi-supervised left atrium segmentation with mutual consistency training. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. Springer, pp. 297–306.

Wu, Huisi, Zhang, Baiming, Chen, Cheng, Qin, Jing, 2023. Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning. IEEE Trans. Med. Imaging 43 (2), 649–661.

Xie, Enze, Wang, Wenhai, Yu, Zhiding, Anandkumar, Anima, Alvarez, Jose M, Luo, Ping, 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xing, Zhaohu, Ye, Tian, Yang, Yijun, Liu, Guang, Zhu, Lei, 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 578–588.

Xiong, Zhaohan, Xia, Qing, Hu, Zhiqiang, Huang, Ning, Bian, Cheng, Zheng, Yefeng, Vesal, Sulaiman, Ravikumar, Nishant, Maier, Andreas, Yang, Xin, et al., 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. 67, 101832.

Zhang, Yundong, Liu, Huiye, Hu, Qiang, 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 14–24.

Zhang, Meisheng, Wang, Chenye, Zou, Wenxuan, Qi, Xingqun, Sun, Muyi, Zhou, Wanting, 2024. Contrmix: Progressive mixed contrastive learning for semi-supervised medical image segmentation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2260–2264.

Zhou, Zongwei, Rahman Siddiquee, Md Mahfuzur, Tajbakhsh, Nima, Liang, Jianming, 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp. 3–11.