BMC Cardiovascular Disorders

RESEARCH

Open Access

Semi-supervised segmentation of cardiac chambers from LGE-CMR using feature consistency awareness

Hairui Wang[1], Helin Huang[1], Jing Wu[1], Nan Li[1], Kaihao Gu[1*] and Xiaomei Wu[1,2,3,4,5*]

Abstract

Background Late gadolinium enhancement cardiac magnetic resonance imaging (LGE-CMR) is a valuable cardiovascular imaging technique. Segmentation of cardiac chambers from LGE-CMR is a fundamental step in electrophysiological modeling and cardiovascular disease diagnosis. Deep learning methods have demonstrated extremely promising performance. However, excellent performance often depended on a large amount of finely annotated data. The purpose of this manuscript was to develop a semi-supervised segmentation method to use unlabeled data to improve model performance.

Methods This manuscript proposed a semi-supervised network that integrates triple-consistency constraints (data-level, task-level, and feature-level) for cardiac chambers segmentation from LGE-CMR. Specifically, we designed a network that integrated segmentation and edge prediction tasks based on the mean teacher architecture. This addressed the problem of ignoring some challenging regions because of excluding low-confidence regions of previous research. We also applied a voxel-level contrastive learning strategy to achieve feature-level consistency, helping the model pay attention to the consistency between features overlooked in previous research.

Results In terms of the Dice, Jaccard, Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD) metrics, for the atrial segmentation dataset, the proposed method achieved scores of 88.34%, 79.30%, 7.92, and 2.02 when trained with 10% labeled data, and 90.70%, 83.09%, 6.41, and 1.72 when trained with 20% labeled data. For the ventricular segmentation task, the results were 87.22%, 77.95%, 2.27, and 0.61 with 10% labeled data, and 88.99%, 80.45%, 1.87, and 0.51 with 20% labeled data, respectively.

Conclusion Experiments demonstrated that our method outperforms previous semi-supervised methods, showing the potential of the proposed network for semi-supervised segmentation problems.

Keywords LGE-CMR, Segmentation, Semi-supervised learning, Consistency, Contrastive learning

*Correspondence:
Kaihao Gu
kaihaogu@aliyun.com
Xiaomei Wu
xiaomeiwu@fudan.edu.cn
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Late gadolinium enhancement cardiac magnetic resonance imaging (LGE-CMR) is a distinctive imaging technique in cardiovascular examination. LGE-CMR is a widely available tool for displaying and quantifying areas of fibrosis and infarction pathology, as gadolinium contrast agents can reveal irreversible myocardial injury and fibrotic changes [1]. Automatic segmentation of the cardiac chambers and lesion regions from LGE-CMR is critical in many clinical applications. Accurate segmentation results are the foundation for diagnosing cardiac diseases and reconstructing cardiac models, which provide essential anatomical information for exploring the pathophysiological mechanisms of cardiac diseases [2–5]. However, segmenting the atria and ventricles from LGE-CMR faces two challenges. First, the morphological characteristics of the atria, ventricles, and cavity sizes pose interpatient variations [6–8]. Second, while gadolinium contrast agents enhance the contrast between lesions and normal tissues, they can also result in the blurring of organ boundaries (myocardial tissue) or simultaneous high-intensity regions in adjacent septa of the cavity, leading to relatively poor image quality [9]. Both of these situations increase the difficulty of LGE-CMR segmentation.

In recent years, the development of medical image segmentation has been greatly facilitated by the emergence of deep learning, particularly the introduction of U-Net [10] and V-Net [11] architectures. For example, a two-stage method was employed to fully automate the left atrium (LA) segmentation [12]. Firstly, Otsu's method was used for left atrium localization, followed by fine segmentation using the U-Net. Despite the high variability of LA anatomy, accurate predictions were still achieved. Similarly, a 3D fully convolutional network was used for automatic left atrium segmentation from LGE-CMR [13]. The model consisted of two convolutional neural networks, the first aimed at coarse segmentation for target localization and the second performing fine segmentation within the localized region. This method ranked first in the 2018 Left Atrial Segmentation Challenge [14], surpassing other traditional approaches such as atlas registration [15] and shape modeling [16]. Automated methods, particularly deep learning-based approaches, have seen increasing adoption in the analysis of ventricular magnetic resonance images. For example, the research [17] proposed an automatic segmentation method combining a Rician-Gaussian mixture model and morphological watershed techniques, effectively segmenting scar regions in the myocardium and distinguishing between boundary zone and core scar. Another study [18] designed a hybrid segmentation framework that integrated traditional computer vision techniques with deep learning pipelines such as multi-atlas approaches, U-Net, and CycleGAN. By leveraging human intervention to combine the strengths of both approaches, this method achieved automatic segmentation of left ventricles with scars from LGE-MR images. Additionally, deep learning-based methods have demonstrated exceptional performance in various ventricular segmentation challenges. In the multi-sequence cardiac MR segmentation challenge [19], deep leaning models that incorporated multi-modal information, such as LGE-CMR and T2-weighted CMR, significantly improved the accuracy of cardiac structure segmentation. Similarly, in the M&Ms challenge [20], U-Net and its variants emerged as the top-performing models overall. These advancements underscore the progress made by deep learning-based segmentation methods in improving segmentation performance. However, these powerful segmentation networks rely heavily on the availability of a substantial amount of finely annotated training data. Nonetheless, high-quality annotated medical images are precious as they require related expertise, and the quality of the images largely depends on clinical experience. To reduce the cost of annotations, recent research has explored the use of unsupervised and weakly supervised approaches for cardiac segmentation. For instance, one study [21] proposed an unsupervised ventricular segmentation algorithm for LGE-CMR. This method transforms readily available Balanced Steady State Free Precession (bSSFP) images into synthetic LGE images employing CycleGAN. Those synthetic images were used to train a segmentation network before being applied to real LGE images. Similarly, another study [22] introduced an unsupervised segmentation framework incorporating intensity and shape constraints to generate realistic synthetic cardiac images through label-based image generation methods. The network was trained with these synthetic images alongside unlabeled real images, achieving effective cardiac segmentation without manual labels. However, those methods depended on the cross-mode image translation and generation. Alternatively, weak-supervision methods have been proposed to reduce labeling efforts using weak labels like bounding boxes [23] and image-level labels [24]. For example, researchers have explored using scribble labels to segment cardiac structures [25]. They proposed a ShapePU network and leveraged constraints on the unlabeled pixels to enhance ventricular segmentation performance. But, collecting these weak labels still requires additional human effort. Therefore, this manuscript proposed to segment LGE-CMR through semi-supervised learning with a small amount of labeled data and a large amount of unlabeled data. This approach aimed to assist algorithms in achieving superior performance.

Semi-supervised learning mainly refers to two issues, which are the generalization of knowledge learned from

limited labeled data to unlabeled data(e.g., pseudo-labeling [26]) and the direct learning based on a large amount of unlabeled data(e.g., entropy minimization [27] and consistency constraints [28–30]). Pseudo-labeling is a simple and intuitive method that utilizes the model's predictions to generate pseudo-labels for unlabeled data, which are then combined with the labeled data for training. The advantage of pseudo-labeling is its simplicity and ease of implementation, but it can introduce noise in the segmentation outputs of under-trained models. Pseudo-labels with noise contained can lead to incorrect knowledge for model learning [31]. In particular, pseudo-labels that are "confident but wrong" can make the training process unstable and result in negative optimization of the model. Therefore, accurately generating pseudo-labels and handling the noise in pseudo-labels are the key focuses of pseudo-labeling methods. Some researchers attempted to generate pseudo-labels based on feature similarity from a reference library [31] or applied a simple thresholding method to select regions with high confidence as pseudo-labels [32]. Entropy minimization constraints aim to reduce the overlap of class probability distributions by minimizing the entropy of unlabeled data, thereby encouraging the model to make low entropy (high confidence) predictions for unlabeled data [27]. Due to its simplicity, entropy minimization, often used as a regularization term, is combined with a supervised loss function or other semi-supervised losses to form a hybrid loss function [33]. Consistency constraints typically introduce perturbations through data augmentation on unlabeled data. The perturbed and unperturbed data are then separately input to the model, aiming for consistent prediction results on a differently perturbed version [34]. This method is more flexible and integrates various design principles for semi-supervised models within the same framework. Therefore, this study primarily utilizes consistency regularization methods to leverage unlabeled data for atrial and ventricular segmentation tasks.

There are multiple implementation approaches for consistency constraints, and one common approach is based on a multi-task network [35, 36]. For example, the level set function prediction task was introduced to impose shape constraints [36]. They argued that different branches focus on different scales of information due to specific tasks, and the different focuses of each task branch can lead to perturbations, thus achieving consistency regularization among tasks. Another common approach is the mean teacher architecture [37–39]. The teacher and student networks share some or all of the parameters. Distance metrics such as mean squared error or divergence metrics such as KL divergence can be used to measure the difference between the outputs of the teacher network and student network, which is then used to enforce consistency between the predictions, thereby achieving semi-supervised learning. In previous work, researchers supposed potential unreliability or noise region existed in the teacher network's predictions when calculating consistency through the mean teacher model. To address this issue, they employed methods such as Monte Carlo dropout [38], setting rules for uncertain regions [37], and error localization subnetworks [40] to exclude low-confidence regions and only compute consistency loss in high-confidence regions. However, these low-confidence regions often correspond to challenging areas such as the edges of cardiac chambers and the junctions between chambers and vessels [41]. We believe that discarding these low-confidence regions may have a negative impact, especially for unique imaging modalities, such as LGE-CMR, where the anatomical structures of the atria can be blurred due to the presence of gadolinium contrast agents. Neglecting these low-confidence regions would not be beneficial for segmentation. In summary, this manuscript identifies the following issues in current research: (1) The methods related to the mean teacher model often use confidence maps to filter out low-confidence regions, which may result in negative impact. It is worth considering how to impose constraints on low-confidence regions. (2) Current research often calculates the consistency loss only on the segmentation results. It is necessary to consider imposing constraints on the semantic scale. (3) Current methods typically calculate consistency constraints under different perturbed versions of the same sample without considering the distribution patterns of the entire dataset. How to impose constraints on the overall dataset is also a worthwhile issue to explore.

To address these issues, we proposed a novel semi-supervised segmentation mean teacher model. *Firstly*, a multi-task network was constructed, consisting of a shared encoder for feature extraction and two independent task-specific output heads to generate segmentation probability maps and edge prediction results. The auxiliary edge prediction task helped the segmentation branch capture more information about border regions, mitigating the loss incurred by filtering out low-confidence regions during consistency computation. *Secondly*, on the encoder's output features, a voxel-level contrastive learning with a Memory Bank was applied. This design not only enforces feature consistency between the original image and its perturbed versions but also encourages the model to consider the distribution patterns of feature representations for voxels belonging to the same or different classes across the entire dataset. *Furthermore*, a cross-task consistency loss between the segmentation and edge prediction tasks was also developed to further exploit the potential of multi-task networks in

semi-supervised segmentation. The proposed method was evaluated on two publicly available LGE-CMR datasets (the 2018 Atria Segmentation Challenge [14] and the EMIDEC challenge [42]) for left atrium and left ventricle segmentation experiments. The experimental results demonstrated that our model improved the performance of cardiac chambers segmentation, which validated the effectiveness of our semi-supervised approach.

In summary, this study has three main contributions:

- We proposed a semi-supervised segmentation network incorporating data-level, task-level, and feature-level consistency constraints. This framework allowed the model to effectively leverage unlabeled data, which enabled it to achieve better segmentation performance.
- This study introduced an edge prediction task within the mean teacher model framework. The multi-task network architecture enhanced segmentation performance by helping the network capture more detailed information from the edge region. Additionally, the edge prediction task established task-level consistency constraints with the primary segmentation task, expanding the ways in which consistency constraints are enforced.
- This manuscript implemented a voxel-level contrastive learning strategy for feature-level consistency. This approach enforced the same category voxel features closer together in the feature space while pushing apart features from different categories, thus preserving contrastive properties while ensuring feature consistency.

Methods

Methods background and design motivation

The mean teacher model has been widely used in semi-supervised learning tasks [39, 43–45]. Inspired by these works, we adopted the mean teacher model as the base architecture. The mean teacher model consists of two branches: a teacher and a student. The different perturbed versions of the same image are input to both branches during training. By minimizing the differences between the teacher and student outputs, the model utilizes unlabeled data for semi-supervised learning. This pattern referred to *data-level consistency (consistency among data).* However, Current semi-supervised segmentation methods using the mean teacher model typically only compute data-level consistency loss in high-confidence regions [38]. To identify the distribution of low-confidence regions, this study applied Monte Carlo dropout to estimate prediction uncertainty which was used in previous research [38]. Specifically, we performed multiple forward passes through the teacher model with random dropout and added gaussian noise for each input volume, calculated softmax probabilities for each voxel, and used predictive entropy as a metric to estimate uncertainty and assess the confidence of each prediction. As shown in Fig. 1, our analysis of the uncertainty map revealed that low-confidence regions were predominantly located around object edges, which are also areas prone to segmentation errors. To enhance the focus on the edge region, this manuscript explicitly introduced an edge prediction task to strengthen the constraint on the segmentation boundary and designed a multi-task network architecture. Different tasks can complement each other, allowing the network to focus on capturing global semantic information and attending to fine-grained details of edge positions.

To further leverage the potential of the edge prediction task, inspired by the DTC network [36], this study introduced consistency between the segmentation task and the edge prediction task. Due to the differences in optimization objectives for specific tasks, segmentation and edge prediction branches may focus on different scales of information, and different focuses of tasks can also introduce perturbations. By mapping/transferring the segmentation results to edge prediction, we can enforce the consistency regularization between the two tasks, thereby establishing *task-level consistency (consistency among tasks).*

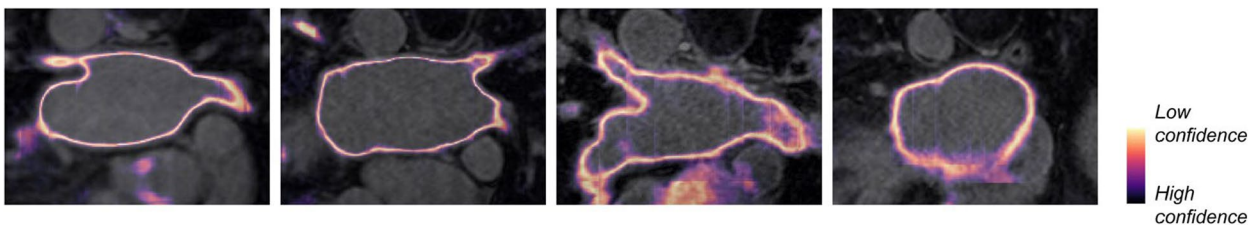

Fig. 1 Example of an uncertainty map from the left atrium dataset. The uncertainty map was obtained using Monte Carlo dropout, with highlighted regions indicating high uncertainty (low confidence). In this dataset, high uncertainty areas were primarily located along the edges of the atrium

Based on this task-level consistency and data-level consistency design, our model can attempt to maintain consistent segmentation masks for the same image and its perturbed version between the teacher and student networks. Moreover, since the original image and its perturbed versions represented the same object, their semantic features should be similar even after different perturbations. This means that the feature embedding obtained by a feature extractor of the teacher and student networks should be similar in the feature space, corresponding to the *feature-level consistency (consistency among features)*. Simple feature-level consistency constraints can be imposed by applying absolute error loss (L1) /absolute error loss (L2) on the encoder output features between the teacher and student networks [46]. However, in addition to ensuring that the encoder outputs of the teacher and the student network are similar, it is essential to ensure the contrastive property of feature embedding in the feature space. In other words, the feature embeddings of voxels belonging to the same category should cluster closely together in the feature space, while those from different categories should be pushed apart. Contrastive learning is perfectly suited to meet this requirement. But current popular contrastive learning approaches like MoCo [47] and SimCLR [48], which treat entire images as instances for contrastive learning, may not be optimal for medical image segmentation tasks. *On the one hand*, instance-level contrastive methods emphasize the minimization of the distance between augmented versions of the same image while maximizing the distance from other images. This approach may potentially overlook the detailed structural information within each image critical for segmentation. *On the other hand*, using a large number of samples for contrast has been proven to be a critical factor in pretraining performance during the construction of positive and negative pairs [48]. However, due to resource limitations, it is challenging to adopt methods like SimCLR [48] that increase the batch size to get a large number of contrastive sample pairs, particularly for 3D medical images with multiple slices. Inspired by previous research [33, 49–51], this study introduced a voxel-level contrastive learning with a memory bank to enforce feature consistency. Specifically, the specified size memory bank stores and dynamically updates voxel features generated during training. So, the model can retrieve a large number of previously stored voxel features from different categories in the memory bank for voxel-level contrastive learning during the training process. This design eliminates the need to recalculate features for each contrastive sample and reduces the dependency on large batch sizes to gather a sufficient number of contrastive pairs. The memory bank effectively increases the number and diversity of contrastive samples without significantly increasing computational overhead.

In summary, shown in Fig. 1, this study designed a semi-supervised medical image segmentation network. The model takes 3D LGE-CMR as input and output object segmentation and edge prediction results. The overall framework consists of three main parts: a multi-task mean teacher structure (shown in khaki on the left), an inter-task transformation module (shown in green on the upper right), and a contrastive learning module for feature consistency (shown in purple on the left). These three parts achieve consistency constraints at the data-level, task-level, and feature-level.

In this manuscript, we defined the semi-supervised problem as follows: Given a semi-supervised training dataset $D_{train} = \{D_l, D_u\}$, where D_{train} consists of N labeled data $D_l = \{x_l, y_l\}$ and M unlabeled data $D_u = \{x_u\}$, with $N << M$. x_l and $y_l$represent the input images and corresponding segmentation annotations from the labeled subset, and x_u represents the input images from the unlabeled subset. Assuming the model's predicted segmentation output is p_{seg}, the semi-supervised approach computes the supervised loss $\mathcal{L}_{sup}$ based on the comparison between p_{seg} and y_l. Additionally, it calculates the unsupervised loss $\mathcal{L}_{semi}$ through consistency measures. The model optimizes its parameters by jointly considering the supervised loss $\mathcal{L}_{sup}$ and the unsupervised loss $\mathcal{L}_{semi}$ as constraints. The overall optimization objective of the model was to minimize the loss function in Eq. (1):

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{semi} = \left(\mathcal{L}_{seg} + \beta \mathcal{L}_{edge} + \gamma L_{consis}^{feat\_l}\right) + \lambda \left(\mathcal{L}_{consis}^{data} + \mathcal{L}_{consis}^{task} + \gamma \mathcal{L}_{consis}^{feat\_u}\right) \tag{1}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{semi}$ represent the supervised loss and unsupervised loss, respectively. $\mathcal{L}_{seg}$ and $\mathcal{L}_{edge}$ are the segmentation loss and edge prediction loss in the multi-task framework (indicated by red dashed arrows), which would be specifically explained in The mean teacher architecture of multiple tasks section $\mathcal{L}_{consis}^{data}$, $\mathcal{L}_{consis}^{task}$, $\mathcal{L}_{consis}^{feat}$ represent the consistency constraints at the data-level, task-level, and feature-level (indicated by yellow, green, and purple arrows, respectively), which will be introduced sequentially in The mean teacher architecture of multiple tasks section to Voxel-level contrastive learning and feature-level consistency section. λ, β and γ are the weighting

coefficients for the supervised loss and unsupervised loss, the segmentation loss and edge prediction loss, and the feature consistency loss and other consistency losses, respectively.

The mean teacher architecture of multiple tasks

As shown in Fig. 2, the khaki-colored parts on the left represent the student and teacher networks, which had the same network structure but were different in parameter update strategies. The student network was updated using gradient descent to minimize the supervised loss on labeled data and the consistency loss on unlabeled data. In contrast, the teacher network was updated using exponential moving average (EMA) of the student network's weights. If we define the weight of the student network at time step t as θ_t, then the weight of the teacher network ξ_t at time step t is:

$$\xi_t = \alpha \xi_{t-1} + (1 - \alpha)\theta_t \tag{2}$$

where α is the update rate for EMA (typically set to 0.99) to balance the proportion of the teacher network weight ξ_t at time step t coming from the student network's weight θ_t and the teacher network's weight ξ_{t-1}. This updated strategy allows the teacher network to

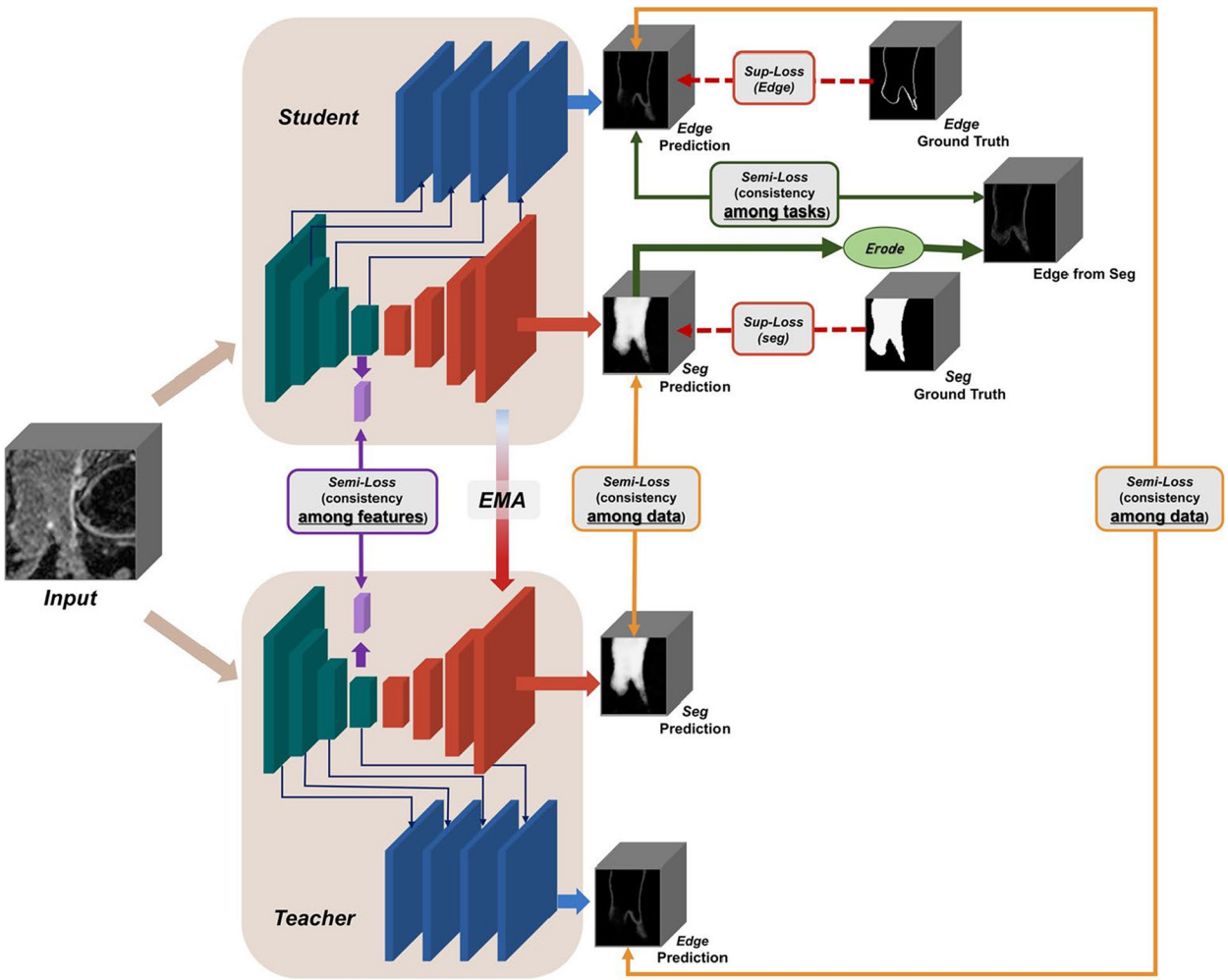

Fig. 2 The overall architecture of the proposed model. The architecture follows the mean teacher model, where student and teacher networks have the same structure. The model consists of an encoder (shown in deep cyan) for feature extraction and two task-specific output heads for segmentation (shown in red) and edge prediction (shown in deep blue). The network processes 3D medical imaging data as input and the dual-task branches simultaneously generate segmentation probability maps and edge prediction results. The model parameter is optimized by minimizing supervised loss (*Sup-Loss*, represented by red arrows) and three types of semi-supervised losses (*Semi-Loss*, indicated by yellow, green, and purple arrows) targeting consistency across data, tasks, and features. The teacher network is updated via the exponential moving average (EMA) of the student network's weights. The "Erode" operation refers to the transformation from segmentation results to edge prediction

provide more stable and reliable predictions, incorporating the knowledge learned by the student network over time.

In the mean teacher model, the architectures of student and teacher networks were identical. For the segmentation branch, this study adopted the V-Net, a classic encoder-decoder architecture widely used in medical imaging, which had demonstrated excellent performance in various medical image segmentation tasks. As shown in Fig. 2, the segmentation branch consisted of four levels of encoders and corresponding decoders. The encoders and decoders were connected through skip connections. Given an input image x, the overall process of the segmentation task can be described as (3):

$$p_{seg} = \mathcal{F}_{seg}(x) \tag{3}$$

where $\mathcal{F}_{seg}$ represents the target segmentation branch, and p_{seg} represents the obtained segmentation result.

For the edge prediction task, the shallow layers of the network tend to generate edges that are unrelated to the classes, while the deeper layers are responsible for detecting class-aware semantic edges. Taking the left atrium segmentation task as an example, we aimed to obtain edge prediction results that complemented the segmentation task and focused only on the edges of the left atrium instead of other cardiac cavities. Therefore, fusing features from shallow and deep layers was particularly important in the model design. Inspired by the DDS network [52], this study adopted a deep supervision-based edge prediction approach to attempt multi-scale fusion and then output. During actual implementation, feature maps from each stage in the encoder were upsampled by trilinear interpolation. Then, the upsampled feature map was concatenated and processed by a 1×1 convolutional layer with a single output channel to generate the edge prediction map p_{edge}. The trainable parameters of the edge prediction module were confined to a convolutional layer dedicated to channel transformation. This design ensured that there was no additional burden imposed on the overall model structure. The overall process can be described as (4):

$$p_{edge} = \mathcal{F}_{edge}\left(\{E^1, up\left(E^2\right), up\left(E^3\right), up\left(E^4\right)\}\right) \tag{4}$$

where E^i represents the output feature map from the i-th stage of the encoder, $i \in [1,4]$;$\{E^1, up(E^2), up(E^3), up(E^4)\}$ represents the result of upsampling and concatenating of feature maps from the four different scales. $\mathcal{F}_{edge}$ represents the edge prediction branch, which was implemented using a 1×1 convolutional layer.

In summary, for the labeled data $D_l = \{x_l, y_l\}$, supervised learning can be performed using the segmentation branch and edge prediction branch. Before calculating the loss, this study first extracted the target edges b_l that matched with input x_l from the segmentation label y_l using an edge extraction algorithm. Since edges appear as single pixels in the image and have weaker constraints, this study empirically extracted edges with equal thickness edges of 2 pixels as supervision signals for optimization. The loss function for the segmentation task is a combination of cross-entropy loss $\mathcal{L}_{ce}$ and Dice loss $\mathcal{L}_{dice}$, given by (5):

$$\mathcal{L}_{seg} = 0.5 \times \left(\mathcal{L}_{ce}\left(p_{seg}, y_l\right) + \mathcal{L}_{dice}\left(p_{seg}, y_l\right)\right) \tag{5}$$

The loss function for the edge prediction task is cross-entropy loss $\mathcal{L}_{ce}$ as (6):

$$\mathcal{L}_{edge} = \mathcal{L}_{ce}\left(p_{edge}, b_l\right) \tag{6}$$

For the unlabeled data $D_u = \{x_u\}$, as the labels y_u were missing, the loss cannot be directly calculated. Ideally, for a same input x experiencing different perturbations, the outputs of the teacher network p^t and the student network p^s should be consistent. Therefore, in the framework of mean teacher model, this study introduced consistency constraints to impose unsupervised loss, encouraging consistent outputs under different perturbations of the same input. During the forward pass of the network, this study applied noise to the input x and used Dropout operations in the network. The differences between the predicted results of the teacher and student networks can serve as unsupervised constraint signals to aid in parameter updates. Since the perturbations were mainly applied to the input images, this approach can be viewed as consistency constraints at the data level. The loss term $\mathcal{L}_{consis}^{data}$ can be described as (7):

$$\mathcal{L}_{consis}^{data} = \mathcal{L}_{consis}\left(p_{seg}^t, p_{seg}^s\right) + \mathcal{L}_{consis}\left(p_{edge}^t, p_{edge}^s\right) \tag{7}$$

where $p_{seg}^t, p_{edge}^t, p_{seg}^s, p_{edge}^s$ represent the segmentation results and edge prediction results from the teacher and student networks, respectively. $\mathcal{L}_{consis}(\cdot)$ is an unsupervised loss used to measure the consistency between the predictions of the teacher and student networks for the same input x with different perturbations. In this study, the Mean Squared Error (MSE) loss was chosen for computing the consistency loss.

Inter-task transformation module and task-level consistency

In order to apply task-level consistency, this study first implemented the transformation from segmentation results to edge prediction conventional ("Erode" in Fig. 2) to minimize the difference in consistency between the two tasks. For a pixel point px belonging to the segmented

object, the segmentation result can be transformed into edge prediction using the formula (8):

$$T(px) = \begin{Bmatrix} 1, \ px \in seg \ \& \ min\{d(px, px' \notin seg)\} < D \\ 0, \ otherwise \end{Bmatrix} \tag{8}$$

where $min\{d(px, px' \notin seg)\}$ describes the minimum distance from the current pixel point px to background pixels (not segmented objects). D is the distance threshold, which can be regarded as the thickness of the edge. In this study, an equal-thickness region of $D = 2$ *pixels* was selected empirically as the target edge. Erosion operation was used to extract the edge from the segmented object. This process was implemented using max-pooling operation. The transformation process did not interrupt gradient backpropagation, making it suitable for parameter optimization using gradient descent.

In this study, the consistency constraint between tasks was only applied to the unlabeled data $D_l = \{x_l, y_l\}$. This loss $\mathcal{L}_{consis}^{task}$ can be described as (9):

$$\mathcal{L}_{consis}^{task} = \mathcal{L}_{consis}(p_{edge}, Erode(p_{seg})) \tag{9}$$

Voxel-level contrastive learning and feature-level consistency

As previously mentioned, this study attempts to enforce feature consistency constraints through voxel-level contrastive learning. Differing from the method of loss calculation for segmentation results based on the aforementioned two consistency constraints, feature consistency constraints attempt loss calculation at the encoder output in the encode-decode structure, thus encouraging the encoder to extract the consistent feature representation for the original image and its perturbed versions. As shown in Fig. 3(a), for an unlabeled sample, it was simultaneously input to both the teacher and student networks. The encoder output of the student network after projection mapping was used as the feature representation of the image. The segmentation result obtained from the teacher network served as a pseudo-label, assigning class information to each voxel in the feature representation. Contrastive learning loss was then computed voxel-wisely. In the specific calculation process, this study treated voxel feature representations of the same class in the Memory Bank as positive samples and different classes as negative samples. If the voxel feature vectors from the student network were considered as queries and the vectors from the Memory Bank were considered as keys, the optimization objective of contrastive learning was to maximize the similarity between queries and keys of the same class and minimize the similarity between queries and keys of different classes. The loss function used in this study is shown in Eq. (10):

$$\mathcal{L}_{consis}^{feat\_u} = \sum_{c=0}^{C} \frac{\frac{1}{N^+}\sum_{i=0}^{N^+} \mathrm{sim}(q_c, k_i^+)}{\frac{1}{N^+}\sum_{i=0}^{N^+} \mathrm{sim}(q_c, k_i^+) + \sum_{j=0}^{N^-} \mathrm{sim}(q_c, k_j^-)} \tag{10}$$

where C is the number of classes, q_c represents the feature representation of a single voxel from the student network, k^+, k^- represent the voxel feature representations from the Memory Bank that belong to the same class and different classes as q_c, respectively, N^+/N^- are the numbers of k^+, k^- in the Memory Bank for q_c, $sim(\cdot)$ is the similarity calculation function using $\mathrm{sim}(q,k) = \exp(q^T k/\tau)$, and τ is the temperature coefficient. During the training phase, the network brought pixels in similar class closer together and these in different class farther apart while optimizing $\mathcal{L}_{consis}^{feat\_u}$ to enforce feature consistency constrains. This ensured the contrastive properties between pixels of the same and different classes and compelling the encoder to learn a good feature representation.

When calculating the loss, a crucial issue is the maintenance of a high-quality Memory Bank. As there is limited storage space for voxel feature representations for all samples, and considering this study focused on semi-supervised segmentation, the Memory Bank only selected high-quality voxel feature vectors from labeled data for storage. The designed Memory Bank was a fixed-length queue with a length of η, and its update rule is shown in Fig. 3(b), following the first-in, first-out principle. The update process of Memory Bank only involved the teacher network. Given labeled data as input, the network outputs the feature representation and segmentation result of the image. The feature quality evaluation rule was applied using the segmentation prediction and segmentation label to select high-quality feature vectors from the feature representation. These selected feature vectors were pushed into the Memory Bank, and the earliest stored feature vectors were popped out. The teacher network was updated using EMA. Its output feature was a smoothed representation of the current and previous time steps. Accordingly, the Memory Bank obtained relatively stable and reliable feature storage. The feature quality evaluation rule was defined as follows: given the output feature representation f^t, segmentation result p_{seg}^t, and segmentation label y_l, the voxel points in f^t that have high-quality feature vectors should satisfy the condition $f^t == \left(Sigmoid\left(p_{seg}^t\right) > \mu\right)$, where μ is the confidence threshold. The selected voxel points were then sorted based on their confidence and the top K voxel points were used as the feature representation for updating the Memory Bank.

To make better use of the labeled data $D_l = \{x_l, y_l\}$ in semi-supervised tasks, we also applied the feature

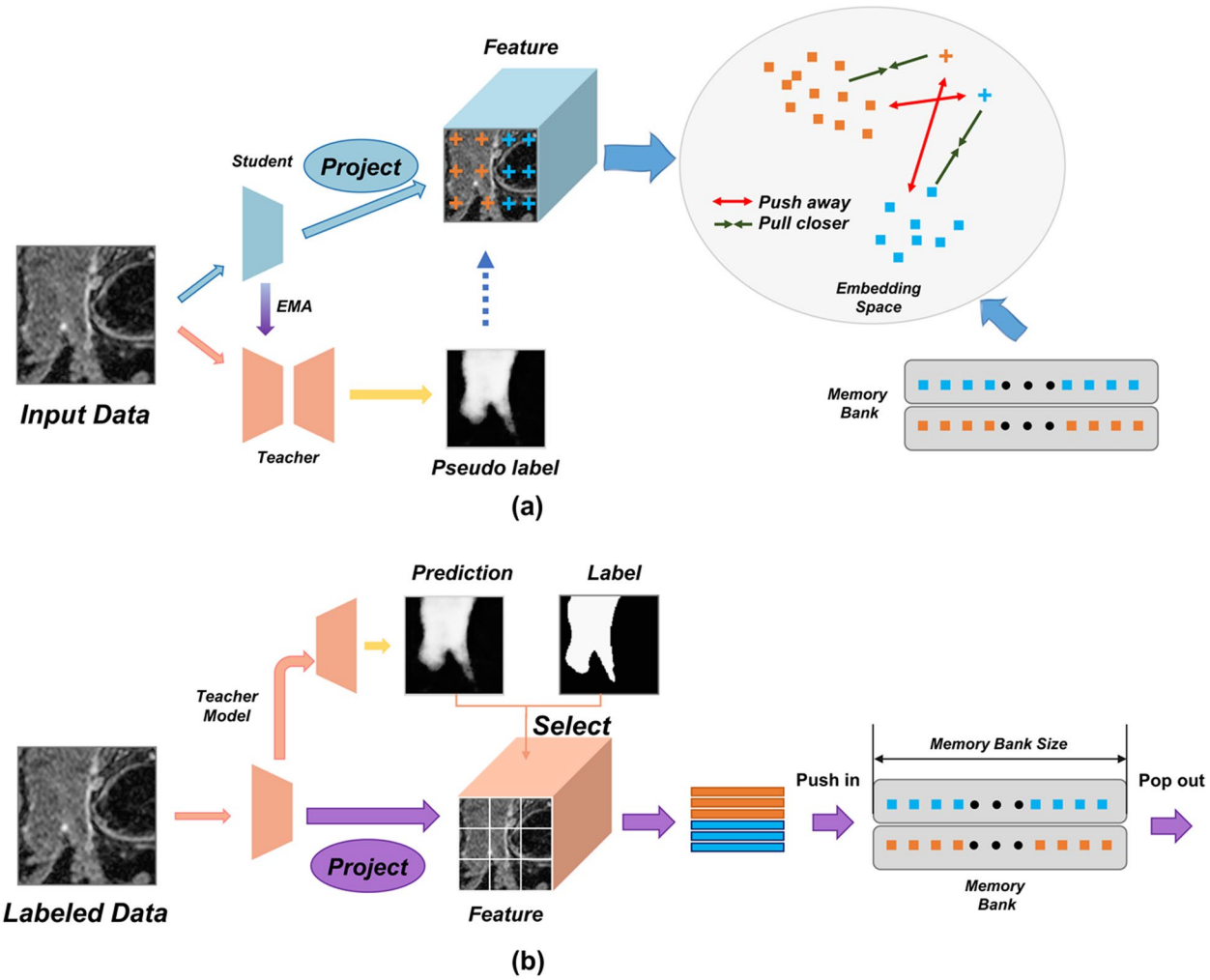

Fig. 3 The schematic diagram of contrastive learning. (**a**) represents a schematic diagram of applying contrastive learning loss. Unlabeled samples are simultaneously fed into both the teacher and student networks. The features generated by the student model's encoder are projected through a mapping layer, and the segmentation results from the teacher model act as pseudo-labels, assigning class information to each voxel of the student model's encoder output. The contrastive learning loss is then calculated. (**b**) demonstrates the updated rules for storing features in the memory bank. The memory bank functions as a fixed-length queue that operates on a first-in, first-out (FIFO) basis. When labeled data is input, the teacher model generates feature representations and segmentation results. High-quality feature vectors are selected based on these segmentation results and corresponding labels and are pushed into the memory bank, while the oldest feature vectors are popped out

consistency loss to D_l. When calculating the loss $L_{consis}^{feat\_l}$, we replaced the pseudo label generated by the teacher network with the actual labels y_l.

Experiments and results

Experimental settings

Dataset

Left atrium segmentation 2018 Atria Segmentation Challenge [14] provides a total of 154 cardiac LGE-CMR scans from 60 patients with atrial fibrillation. 100 LGE-CMR scans have been publicly available, which have been annotated by doctors for left atrium segmentation. Following the approach described in [38], this study divided the data into 80 training images and 20 testing images.

Left ventricle segmentation EMIDEC challenge [42] provides 150 data samples from different patients, including 50 normal cases and 100 cases of myocardial infarction. Currently, 100 training data and corresponding labels have been publicly available, consisting of 67 pathological cases and 33 normal cases. In this study, only the left ventricle labels were used for experimentation. The left ventricle image data from the EMIDEC

challenge was randomly split into five subsets for 5-fold cross-validation, which was used to evaluate the model's performance.

Model implementation details

The overall objective of this study was to minimize the loss function, as shown in Eq.(1). The hyperparameters for the training process were set as follows: the weight coefficients β for segmentation and edge prediction loss was set to 0.6, and the weight coefficients γ for feature-level consistency loss with other consistency losses was set to 0.1. The weight coefficient λ for supervised and unsupervised losses were defined as a ramp-up function $\lambda(t) = 0.1 * e^{-5(1-t/t_{max})}$, where t, t_{max} represent the current iteration step and the maximum iteration number of training, respectively.

During the training process, the initial learning rate was set to 0.01 and decreased by a factor of 0.5 every 2500 iterations. The network parameters were optimized using the SGD optimizer with momentum (weight decay=0.0001, momentum=0.9) for 10,000 iterations. The batch size for training was set to 4, which included two labeled images and two unlabeled data samples. In the preprocessing step, all data were cropped around the heart region and normalized to zero mean and unit variance. Since the original sizes of the data in the two datasets were different, we randomly cropped 112×112×80 sub-volumes as the input to the network for the left atrium segmentation dataset and cropped 48×48×4 sub-volumes as the input for the left ventricle segmentation. Due to the different input sizes in the two datasets, there were slight differences in the 4-level encoding-decoding structure of the segmentation network. For the ventricle dataset, the down-sampling ratio was adjusted to ensure proper functioning.

We believed that the extracted image features were coarse and less reliable at the beginning of the network training. Therefore, in the early stage of training, we did not apply feature consistency constraints to avoid the model learning incorrect feature distributions. We set the iterations of training from 500 to 1000 as the preparation period for the Memory Bank. During this stage, high-quality voxel feature representations generated were stored in the Memory Bank and feature consistency loss is not calculated until after 1000 rounds. The confidence threshold μ in the feature quality evaluation rule was set to 0.95. Each category in the Memory Bank can store up to η=2028 voxel features, and the maximum number of feature vectors stored per image in the Memory Bank is $K = \frac{\eta}{len(x_l)}$. The temperature coefficient τ in the loss calculation was set to 0.1. Regarding the specific model details, the Project used in this manuscript followed the design flow of Conv→ReLu→Conv, where a 1×1 convolution was used to project the 256-channel feature output from the encoder to 128 channels. This Project was designed explicitly for voxel-level contrastive learning and differed from the commonly used Multi-Layer Perceptron (MLP) in instance-level contrastive learning.

Evaluation metrics

To quantitatively evaluate the segmentation results, we used four common metrics for segmentation tasks. They are the Dice coefficient (Dice), Jaccard Index (Jaccard), Average surface distance (ASD), and 95% Hausdorff Distance (95HD). The first two metrics, Dice and Jaccard, primarily measure the overlap between the segmentation result and the ground truth from a regional perspective, with higher values indicating better overlap. The latter two metrics, ASD and 95HD, assess the similarity of all point pairs between the segmentation result and the ground truth from a surface perspective. In this case, smaller values indicate better similarity.

Results

Table 1 presents the quantitative performance evaluation of the left atrium segmentation task. Referring to common practices in semi-supervised segmentation [53, 54], our proposed method achieved the highest performance when trained with both 10% and 20% labeled data. Specifically, with 8 labeled images (10%), the Dice, Jaccard, ASD, and 95HD were 88.34%, 79.30%, 7.92, and 2.02, respectively, While using 16 labeled images (20%), the Dice, Jaccard, ASD, and 95HD improved to 90.70%, 83.09%, 6.41, and 1.72, respectively. Figure 4 provides visual segmentation results trained by 20% labeled data, where the third column shows the 3D reconstruction of four samples from the left atrium segmentation dataset, along with the segmentation results on slices.

The quantitative and visual evaluations of the left ventricle segmentation dataset were performed similarly to the left atrium dataset. Table 2 presents the performance metrics, and Fig. 5 provides visual results trained by 20% Labeled data. To assess the model's performance, the proposed method was compared to the UA-MT [38] and DTC [36], which were reproduced by applying the left ventricular segmentation dataset. As a result, our method trained by 10% Labeled data achieved Dice, Jaccard, ASD, and 95HD scores of 87.22%,77.95%, 2.27, and 0.61, respectively. Our method trained by 20% Labeled data achieved Dice, Jaccard, ASD, and 95HD scores of 88.99%, 80.45%, 1.87, and 0.51, respectively.

Table 1 Quantitative evaluation of the left atrium segmentation

Method	Scans used		Dice(%)	Jaccard(%)	95HD(voxel)	ASD(voxel)
	Labeled	Unlabeled				
V-Net[a]	8	0	78.57	66.96	21.20	6.07
V-Net[a]	16	0	86.03	76.06	14.26	3.51
V-Net[a]	80	0	91.14	83.82	5.75	1.52
UA-MT [38]	8(10%)	72	84.25	73.48	13.84	3.36
SASSNet [55]	8(10%)	72	87.32[b]	77.72[b]	9.62[b]	2.55[b]
LG-ER-MT [56]	8(10%)	72	85.54[b]	75.12[b]	13.29[b]	3.77[b]
DTC [36]	8(10%)	72	87.51	78.17	8.23	2.36
Ours	8(10%)	72	**88.34**	**79.30**	**7.92**	**2.02**
UA-MT [38]	16(20%)	64	88.88[b]	80.21[b]	7.32	2.26[b]
SASSNet [55]	16(20%)	64	89.54[b]	81.24[b]	8.24	2.20[b]
LG-ER-MT [56]	16(20%)	64	89.62[b]	81.31[b]	7.16	2.06[b]
DTC [36]	16(20%)	64	89.42[b]	80.98[b]	7.32[b]	2.10[b]
Ours	16(20%)	64	**90.70**	**83.09**	**6.41**	**1.72**

[a] indicates the segmentation performance trained with only the labeled data sourced from UA-MT [38]

[b] denotes that our method (best value) is significantly better than the reference method (p-value < 0.05) based on a paired t-test). Since the UA-MT method only provided its performance under the 10% labeled data setting and did not release the model files, a paired t-test comparison with this method was not conducted

Discussion

Compare with other methods

As shown in Figs. 4 and 5, we visually compared the typical UA-MT [38] and DTC [36] methods with our proposed method. UA-MT [38] and DTC [36] are commonly used as baselines, which correspond to the mean teacher model and multi-task model, respectively. These network structures and model parameters for the left atrium segmentation task were obtained from publicly available code. For the left ventricle segmentation task, we reproduced these network structures and trained on left ventricle dataset. The last column in Figs. 4 and 5 represents the ground truth provided by the doctors, which serves as the reference standard for performance evaluation. For the left atrium segmentation task, the top two rows and bottom two rows in Fig. 4 display the

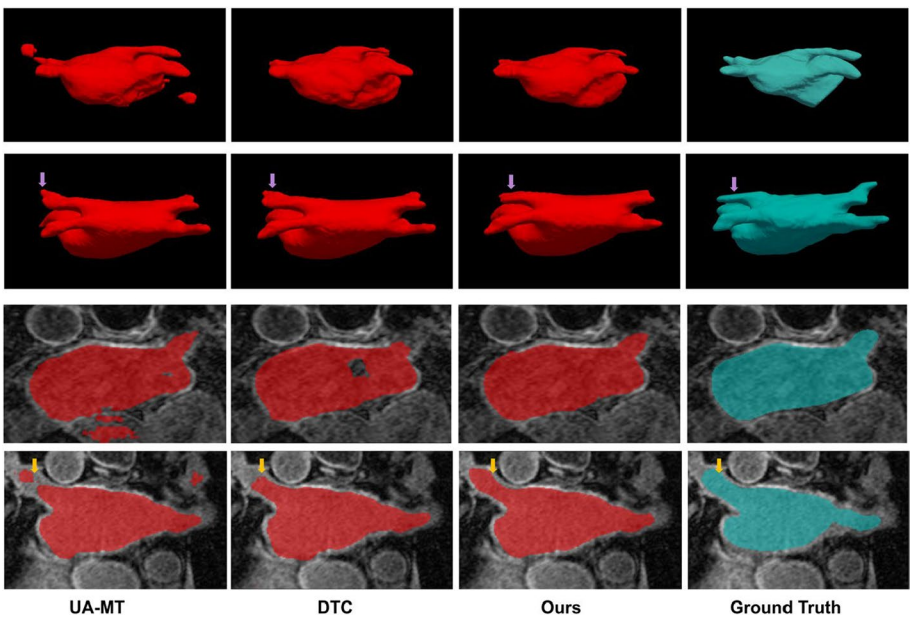

Fig. 4 The left atrium segmentation results. The first two rows show the results of 3D reconstruction; the last two rows display the segmentation results on individual slices. From left to right, the images represent the prediction results from UA-MT [38], DTC [36], our method, and the Ground Truth

Table 2 Quantitative evaluation of the left ventricular segmentation(mean±variance)

Method	Scans used		Dice(%)	Jaccard(%)	95HD(voxel)	ASD(voxel)
	Labeled	Unlabeled				
V-Net[a]	8	0	83.67±9.09	72.86±12.12	7.08±8.65	1.91±2.06
V-Net[a]	16	0	87.77±5.93	78.67±8.77	3.07±4.77	0.90±1.27
V-Net[a]	80	0	91.38±4.00	84.36±6.33	1.49±0.66	0.37±0.38
UA-MT	8(10%)	72	84.68±7.41[b]	74.3±10.47[b]	5.99±3.4[b]	1.71±1.04[b]
DTC	8(10%)	72	85.62±7.41[b]	75.52±10.47[b]	3.08±3.4[b]	0.84±1.04[b]
Ours	8(10%)	72	**87.22±7.00**	**77.95±9.97**	**2.27±1.75**	**0.61±0.5**
UA-MT	16(20%)	64	87.94±5.12[b]	78.93±7.73[b]	3.30±0.91[b]	0.92±0.33[b]
DTC	16(20%)	64	88.58±5.12	79.85±7.73	1.88±0.91	0.52±0.33
Ours	16(20%)	64	**88.99±4.60**	**80.45±7.10**	**1.87±0.84**	**0.51±0.31**

[a] indicates the segmentation performance trained with only the labeled data

[b] denotes that our method (best value) is significantly better than the reference method (p-value < 0.05) based on a paired t-test

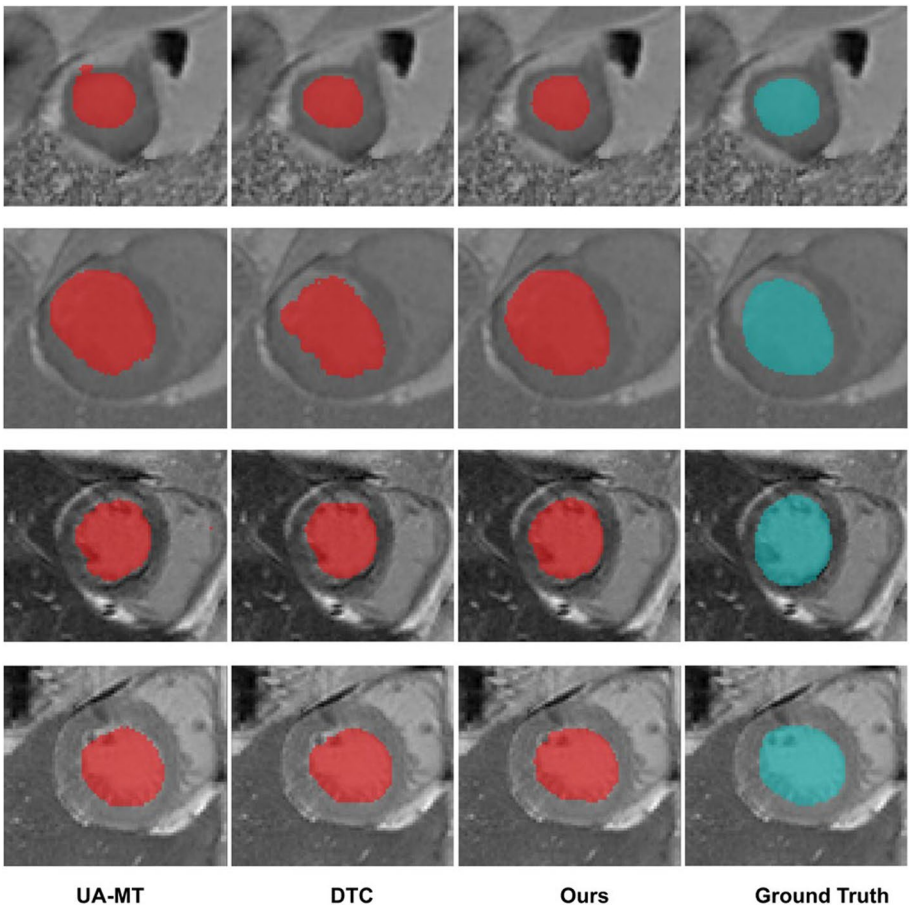

Fig. 5 The left ventricular segmentation results. From left to right, the images represent the prediction results from UA-MT [38], DTC [36], our method, and the Ground Truth

results of 3D reconstruction and single-slice segmentation, respectively. Comparing the first and the third rows, it can be observed that our method preserves the overall structure of the atrium better than the other methods. For example, in the third rows, the prediction of UA-MT incorrectly identifies the regions not

belonging to the atrium, resulting in over-segmentation and the result of DTC exhibits a hole in the atrium segmentation, indicating under-segmentation problem. In contrast, the segmentation of proposed method is more consistent with the Ground Truth. The second and fourth rows demonstrate that our proposed method performs better in capturing details. These areas indicated by purple and yellow arrows in the images correspond to the pulmonary vein orifice connected to the left atrium. For example, in the comparison of fourth rows, the predictions of UA-MT and DTC exhibit discontinuity or missing in the pulmonary vein area, whereas our segmentation are more similar to the Ground Truth in morphology. Our method achieved more consistent results with the ground truth, possibly due to the multitask structure of our proposed model that effectively focuses on edge information (as discussed in Effectiveness of edge prediction tasks section). For the left ventricle segmentation task, as the dataset contains a smaller number of slices (ranging from 4 to 10), the 3D reconstruction visualization is less prominent. Therefore, only single-slice segmentation results are presented in Fig. 5. Similar with the segmentation results of the left atrium, our proposed method achieved better consistency with the ground truth regarding the overall ventricle structure and edge details. Tables 1 and 2 provide the quantitative analysis results, which comprehensively evaluate the performance based on the Dice, Jaccard, ASD, and 95HD metrics. Our proposed model outperforms the other methods across both left atrium and left ventricle segmentation tasks.

Figure 6 utilizes a box plot with a scatter plot overlay to depict the segmentation performance of the model on the LA and LV datasets. It is apparent from the figure that, across both datasets, the proposed method not only shows higher mean and median values in the Dice coefficient but also exhibits a more concentrated distribution of results. This indicates that the method, in addition to achieving better overall performance, also possesses a certain level of stability and robustness.

Ablation experiment

The effectiveness of different components in the proposed method is demonstrated through ablation experiments. It is worth noting that all ablation experiments are conducted on the LA dataset, which consists of 16 labeled data and 64 unlabeled data. The results of the four ablative experiments are presented in Table 3 and Fig. 7, where different components of the network are deactivated by using different losses. Experiments I and II represent supervised learning results using 16 labeled samples (20%) for training. Experiment I exclusively employs segmentation loss for training, which can be considered as the baseline for segmentation performance. Experiment II shows that incorporating edge prediction effectively improves the model's performance. Experiment III adds data-level consistency constraints to the model, resulting in a 0.76% performance gain. Experiments IV and V, based on data consistency, add task-level and feature-level consistency constraints, respectively. The performance is further enhanced, improving Dice scores by 1.06% and 1.20%, respectively. Finally, Experiment VI represents the overall model proposed in this manuscript, achieving the best performance. The results of the ablation experiments demonstrate that the edge prediction task, data-level consistency constraint, task-level consistency constraint, and feature-level consistency constraint in the model can complement each other and further enhance the model's performance. A paired t-test was conducted to analyze the significance of the differences between Experiments II-V and the baseline (Experiment I), as shown in Table 3. Although significant differences were not shown in the comparisons between the baseline and Experiments II-V (p-values>0.05), the comparison between the baseline and the proposed model (Experiment VI) resulted in a p-value<0.05, showing statistically significant improvement. This suggests that the significant improvement is due to the combined effect of the data-level, task-level, and feature-level consistency. Additionally, to further illustrate the ablation experiment, Fig. 7 presents bar graphs for the other evaluation metrics: Jaccard, ASD, and 95HD. The figure shows that the proposed method outperforms the ablation experiments (Experiments I-V) across all evaluation metrics, achieving the best results.

The effectiveness of voxel-level contrast learning

Consistent representation of features

In this study, it is believed that by applying voxel-level contrastive learning, the model can achieve consistent representations of similar data in the feature space, where pixels of the same class cluster together while pixels of different classes are as far apart as possible. To visualize these features, a classical feature distribution visualization method called t-SNE is employed. Figure 8 shows orange points representing foreground voxels (left atrium) and green points representing background voxels (all non-left atrium regions). It can be observed that after applying, the feature distribution exhibits clear clustering of points in same color and clear trend of separation between the orange and green points. This visualization confirms that the voxel-level contrastive learning employed in this study effectively brings similar pixels closer in the feature space while pushing different pixels apart, thereby achieving consistent representations of features.

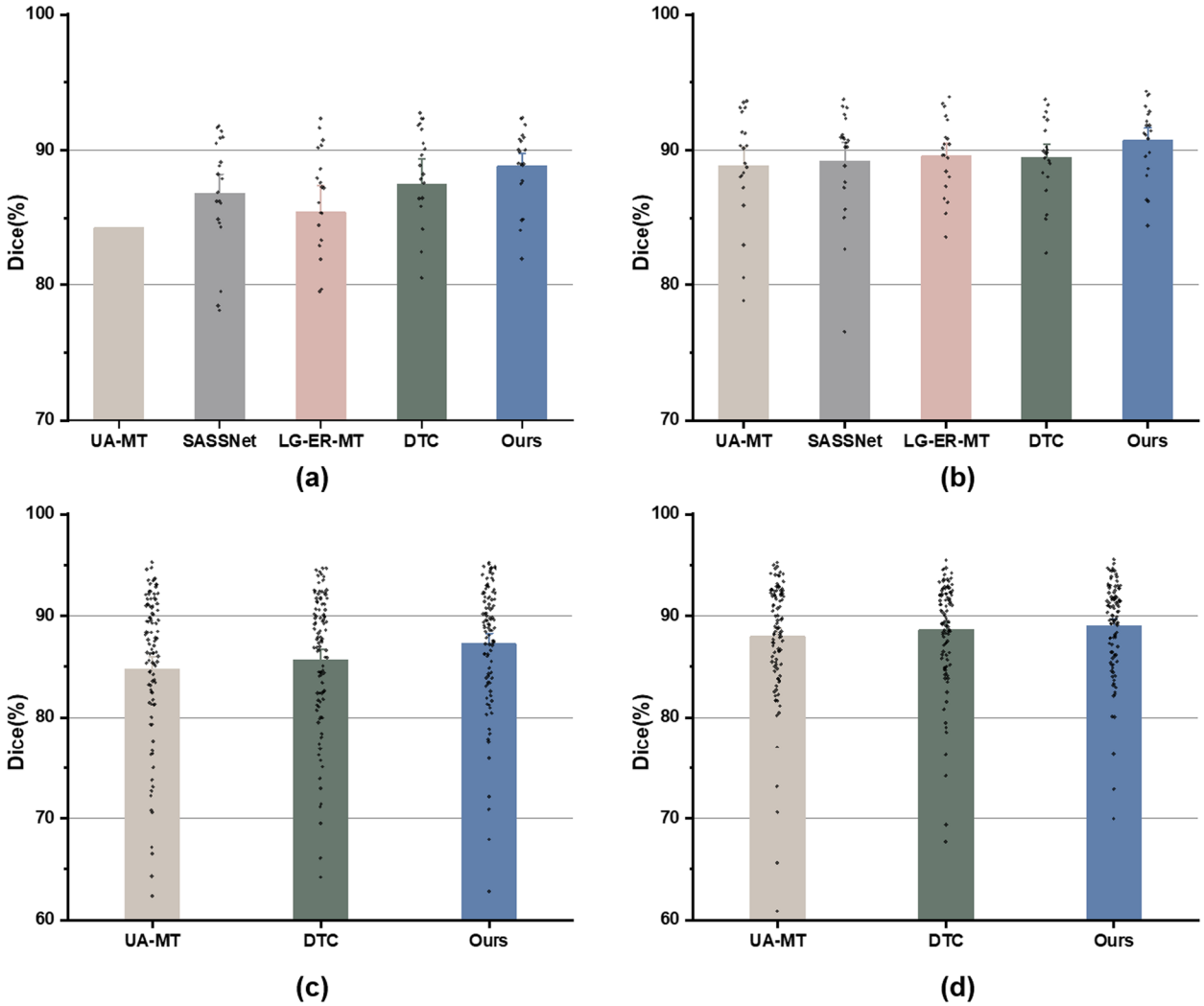

Fig. 6 Bar graph with a scatter plot overlay for different methods on test samples. (**a**) and (**b**) respectively represent the results trained by 10% labeled data and 20% labeled data on the LA dataset. From left to right, the methods compared are UAMT [38], SASSNet [55], LG-ER-MT [56], DTC [36], and the proposed method. In (**a**), only the mean value of the UA-MT method with 10% labeled data is shown in the bar graph since the model files were not publicly available. (**c**) and (**d**) display the results trained with 10% and 20% labeled data on the left ventricle LV dataset. From left to right, the methods are UAMT, DTC, and the proposed method

Table 3 Comparison of ablation experiments

	Methods	$\mathcal{L}_{seg}$	$\mathcal{L}_{edge}$	$\mathcal{L}_{consis}^{data}$	$\mathcal{L}_{consis}^{task}$	$\mathcal{L}_{consis}^{feat}$	Dice(%)
I	Seg	√					89.14
II	Seg+Edge	√	√				89.64 (+0.50)
III	Seg+Edge+Data	√	√	√			89.90 (+0.76)
IV	Seg+Edge+Data+Task	√	√	√	√		90.20 (+1.06)
V	Seg+Edge+Data+Feat	√	√	√		√	90.34 (+1.20)
VI	**Seg+Edge+Data+Task+Feat (Ours Method)**	√	√	√	√	√	**90.70[a] (+1.56)**

[a] denotes that our method (VI, best value) is significantly better than the baseline method (I) based on a paired t-test (p-value < 0.05)

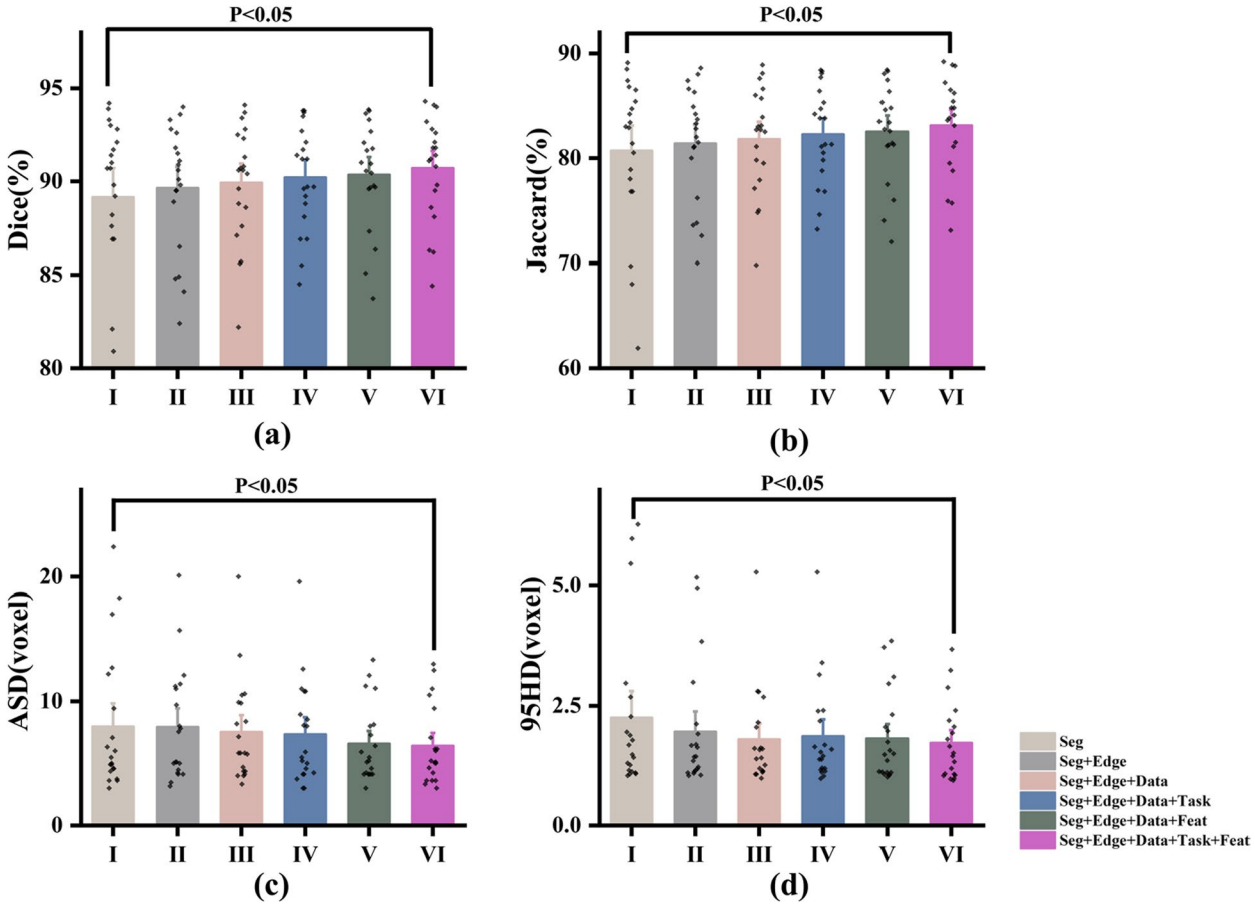

Fig. 7 A bar graph with a scatter plot overlay presenting results for different methods on the test set from the LA dataset. (**a**), (**b**), (**c**), and (**d**) correspond to the four metrics: Dice, Jaccard, ASD, and 95HD. The proposed method (VI) outperforms all ablation experiments (I-IV) across all metrics, demonstrating superior performance in Dice, Jaccard, ASD, and 95HD

Comparison of critical hyperparameters in voxel-level contrast learning

In this study, maintaining a high-quality Memory Bank is considered crucial in implementing voxel-level contrastive learning. The Memory Bank size is a critical hyperparameter that limits the number of stored samples. Apart from its impact on computational efficiency, the number of features in the Memory Bank can affect the effectiveness of contrastive learning. To investigate this issue, experiments were conducted with Memory Bank sizes ranging from 128 to 4096. The results of this experiment are presented in Table 4. Within a certain range, using a larger number of samples during training tends to yield better performance. The best results can be achieved when the Memory Bank size is set to 2048.

Comparison of voxel-level contrastive learning and instance-level contrastive learning

Voxel-level contrastive learning has two main advantages. First, it overcomes the challenge of using instance-level contrastive learning in segmentation tasks, where an image may contain pixels belonging to different classes. Second, it addresses the difficulty of constructing positive-negative sample pairs in contrastive learning due to the limited size of medical image datasets in medical image segmentation. To demonstrate the advantages of voxel-level contrastive learning, this study compared it with instance-level contrastive learning based on the simCLR [48]. As shown in Table 5, the Dice coefficients obtained using instance-level and voxel-level contrastive learning strategies on the left atrium dataset were 89.34% and 90.70%, respectively, indicating a performance improvement with voxel-level contrastive learning.

Effectiveness of edge prediction tasks

In the ablation experiments presented in Table 3, this study conducts a quantitative analysis of the impact of the edge prediction task by comparing Experiments I with II. The comparison reveals that incorporating the edge prediction task leads to an improvement in

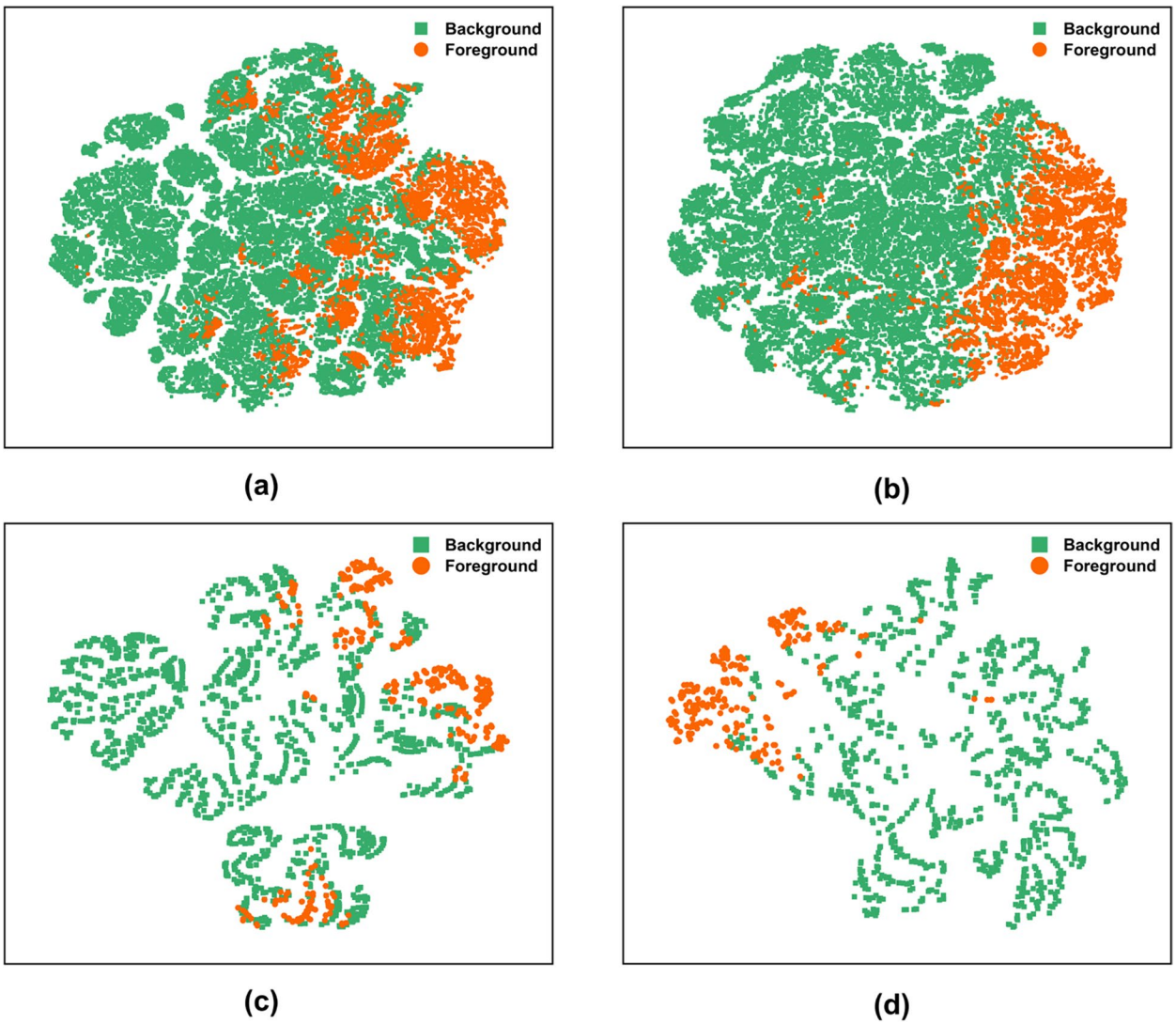

Fig. 8 t-SNE result of voxel features on the test set of the atrial dataset. (**a**) and (**b**) show the overall distribution before and after applying feature consistency loss. Due to the large number of voxels, a random selection of 50,000 voxel points was visualized. (**c**) and (**d**) demonstrate the feature distribution of a single data sample from the test set before and after applying feature consistency loss

Table 4 Influences of different Memory Bank size

MemoryBank Size	128	512	1024	2048	4096
Dice	90.29	90.40	90.20	**90.70**	90.23

Table 5 Comparison of voxel-level contrast learning and instance-level contrastive learning

Contrastive Loss Type	Instance Level	Pixel Level
Dice	89.34	**90.70**

the Dice coefficient from 89.14 to 89.64%. Furthermore, for the proposed semi-supervised segmentation model, the edge prediction task serves as a crucial component of task-level consistency constraints. As depicted in Table 3, Experiment IV, by introducing inter-task consistency constraints on Experiment III, raises the Dice coefficient from 89.90 to 90.20%.

Considering the primary role of the edge prediction task in the model is to guide the model to pay attention to the neglected edge regions. To illustrate the effect of the edge segmentation task, this study presents the results of three edge prediction examples in Fig. 9. From the figure, it can be observed that the edge

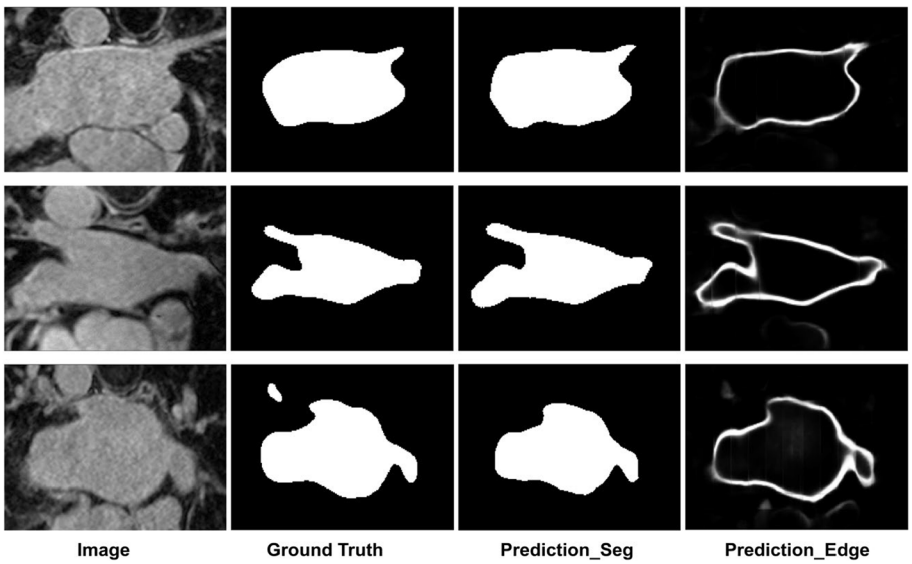

Fig. 9 Results of edge prediction, from left to right are the original image, Ground Truth, predicted segmentation result, and predicted edge result

prediction task successfully achieves the intended goal by capturing the structural information of the edges.

Limitations

Although the overall performance of our method is encouraging. There are still some limitations in our work. This study primarily focused on enforcing consistency at the data-level, task-level, and feature-level. The consideration of model structure-level consistency deserves further exploration. For instance, some researches not only employed CNN to construct models but also attempted to build hybrid models incorporating Transformer and CNN structures [57, 58]. Applying the differences between various model structures for implementing consistency regularization has demonstrated excellent semi-supervised performance. Additionally, our method concentrated on the semi-supervised learning task of cardiac chamber segmentation, specifically on two publicly available datasets containing only 100 samples of LGE-CMR for the left atrium and left ventricle. In the future, we aim to construct datasets with a larger and more diverse patient population for testing and extend our approach to other medical segmentation tasks.

Conclusion

In this work, we focused on the semi-supervised segmentation of the atrium and ventricle on LGE-CMR images. To address the potential neglect of the cardiac cavity border regions, an edge prediction task was introduced within the framework of the mean teacher model to enhance the model's focus on these edge areas. Furthermore, to fully leverage unlabeled data, this study integrated triple-consistency constraints at the data-level, task-level, and feature-level through the mean teacher model, inter-task transformation module, and voxel-level contrastive learning, respectively. In particular, the introduction of voxel-level contrastive learning allowed the model to observe the feature distribution patterns of the entire dataset and encouraged the formation of consistent feature representations. This study demonstrated excellent performance in both atrium and ventricle segmentation tasks compared to other semi-supervised segmentation methods. For the atrium segmentation task, when trained with 20% labeled data, the Dice, Jaccard, ASD, and 95HD were 90.70%, 83.09%, 6.41, and 1.72, respectively. When trained with 10% labeled data, the corresponding values were 88.34%, 79.30%, 7.92, and 2.02. For the ventricle segmentation task, the metrics with 20% labeled data were 88.96%, 80.41%, 1.74, and 0.49, respectively. With 10% labeled data, the Dice, Jaccard, ASD, and 95HD were 87.22%, 77.95%, 2.27, and 0.61. These results validate the effectiveness of the proposed model in both atrium and ventricle segmentation tasks.

Acknowledgements
Not applicable.

Authors' contributions
Methodology and Writing-original draft were performed by Hairui Wang, Data curation and Investigation were performed by Helin Huang, Jing Wu and Nan Li. Supervision and Writing-review & editing were performed by Kaihao Gu and Xiaomei Wu. Funding acquisition was performed by Xiaomei Wu. All authors read and approved the final manuscript.

Funding
This work was supported by National Key Research and Development Program, grant no.2021YFC2400203, Shanghai Municipal Commission of Economy and Information Technology, grant no.GYQJ-2018-2-05, and Medical Engineering Fund of Fudan University, grant no.yg2021-38.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Clinical trial number: not applicable. The data used in this manuscript were obtained from opensource databases (*2018 Atria Segmentation Challenge* and *EMIDEC challenge*).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Biomedical Engineering, School of information Science and Technology, Fudan University, Shanghai 200433, China. [2]Academy for Engineering and Technology, Fudan University, Fudan University, Shanghai 200433, China. [3]Yiwu Research Institute of Fudan University, Yiwu, Zhejiang 322000, China. [4]Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention (MICCAI) of Shanghai, Shanghai 200433, China. [5]Shanghai Engineering Research Center of Assistive Devices, Shanghai 200433, China.

Received: 17 July 2024 Accepted: 10 October 2024
Published online: 17 October 2024

References

1. Hassan S, Barrett CJ, Crossman DJ. Imaging tools for assessment of myocardial fibrosis in humans: the need for greater detail. Biophys Rev. 2020;12:969–87. https://doi.org/10.1007/s12551-020-00738-w.
2. Ma Y, Ding P, Li L, et al. Three-dimensional printing for heart diseases: clinical application review. Biodes Manuf. 2021;4:675–87. https://doi.org/10.1007/s42242-021-00125-8.
3. Martin-Isla C, Campello VM, Izquierdo C, et al. Image-based cardiac diagnosis with machine learning: a review. Front Cardiovasc Med. 2020;7:1. https://doi.org/10.3389/fcvm.2020.00001.
4. Sander J, de Vos BD, Išgum I. Automatic segmentation with detection of local segmentation failures in cardiac MRI. Sci Rep. 2020;10:21769. https://doi.org/10.1038/s41598-020-77733-4.
5. Morais P, Vilaça JL, Queirós S, et al. Automated segmentation of the atrial region and fossa ovalis towards computer-aided planning of inter-atrial wall interventions. Comput Methods Programs Biomed. 2018;161:73–84. https://doi.org/10.1016/j.cmpb.2018.04.014.
6. Heist EK, Refaat M, Danik SB, et al. Analysis of the left atrial appendage by magnetic resonance angiography in patients with atrial fibrillation. Heart Rhythm. 2006;3:1313–8. https://doi.org/10.1016/j.hrthm.2006.07.022.
7. Varela M, Bisbal F, Zacur E, et al. Novel Computational Analysis of Left Atrial Anatomy Improves Prediction of Atrial Fibrillation Recurrence after ablation. Front Physiol. 2017;8. https://doi.org/10.3389/fphys.2017.00068.
8. Colan SD, Shirali G, Margossian R, et al. The ventricular volume variability study of the Pediatric Heart Network: Study Design and impact of beat averaging and variable type on the reproducibility of echocardiographic measurements in children with chronic dilated cardiomyopathy. J Am Soc Echocardiogr. 2012;25:842–e8546. https://doi.org/10.1016/j.echo.2012.05.004.
9. Li L, Zimmer VA, Schnabel JA, Zhuang X. Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: a review. Med Image Anal. 2022;77:102360. https://doi.org/10.1016/j.media.2022.102360.
10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Munich, Germany: Springer International Publishing; 2015. p. 234–41.
11. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). Stanford, CA: IEEE; 2016. p. 565–571.
12. Borra D, Andalò A, Paci M, et al. A fully automated left atrium segmentation approach from late gadolinium enhanced magnetic resonance imaging based on a convolutional neural network. Quant Imaging Med Surg. 2020;10:1894–907. https://doi.org/10.21037/qims-20-168.
13. Xia Q, Yao Y, Hu Z, Hao A. Automatic 3D Atrial Segmentation from GE-MRIs Using Volumetric Fully Convolutional Networks. In: Pop M, Sermesant M, Zhao J, editors. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Cham: Springer International Publishing; 2019. p. 211–220.
14. Xiong Z, Xia Q, Hu Z, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med Image Anal. 2021;67:101832. https://doi.org/10.1016/j.media.2020.101832.
15. Qiao M, Wang Y, van der Geest RJ, Tao Q. Fully Automated Left Atrium Cavity Segmentation from 3D GE-MRI by Multi-atlas Selection and Registration. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Cham: Springer International Publishing; 2019. p. 230–236.
16. Nuñez-Garcia M, Zhuang X, Sanroma G, et al. Left Atrial Segmentation Combining Multi-atlas Whole Heart Labeling and Shape-Based Atlas Selection. In: Pop M, Sermesant M, Zhao J, editors. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Cham: Springer International Publishing; 2019. p. 302–310.
17. Mamalakis M, Pankaj G, Tom N, et al. Automatic development of 3D anatomical models of border zone and core scar regions in the left ventricle. Comput Med Imaging Graph. 2023;103:102152. https://doi.org/10.1016/j.compmedimag.2022.102152.
18. Mamalakis M, Panka G, Tom N, et al. Artificial Intelligence framework with traditional computer vision and deep learning approaches for optimal automatic segmentation of left ventricle with scar. Artif Intell Med. 2023;143:102610. https://doi.org/10.1016/j.artmed.2023.102610.
19. Zhuang X, Xu J, Luo X, et al. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. Med Image Anal. 2022;81:102528. https://doi.org/10.1016/j.media.2022.102528.
20. Martin-Isla C, Campello VM, Izquierdo C, et al. Deep learning segmentation of the right ventricle in Cardiac MRI: the M&Ms challenge. IEEE J Biomed Health Inf. 2023;27:3302–13. https://doi.org/10.1109/JBHI.2023.3267857.
21. Yu X, Chen J, Fang B, et al. Cardiac LGE MRI Segmentation with Cross-modality Image Augmentation and Improved U-Net. IEEE J Biomedical Health Inf. 2023;27:588–97. https://doi.org/10.1109/JBHI.2021.3139591.
22. Wang S, Wu F, Li L, et al. Unsupervised Cardiac Segmentation utilizing synthesized images from anatomical labels. In: Camara O, Puyol-Antón E, Qin C, et al. editors. Statistical atlases and computational models of the heart. Regular and CMRxMotion Challenge Papers. Cham: Springer Nature Switzerland; 2022. pp. 349–58.
23. Rajchl M, Lee MCH, Oktay O, et al. DeepCut: object segmentation from bounding Box annotations using Convolutional neural networks. IEEE Trans Med Imaging. 2017;36:674–83. https://doi.org/10.1109/TMI.2016.2621185.
24. Xiong H, Liu S, Sharan RV, et al. Weak label based bayesian U-Net for optic disc segmentation in fundus images. Artif Intell Med. 2022;126:102261. https://doi.org/10.1016/j.artmed.2022.102261.
25. Zhang K, Zhuang X. ShapePU: a New PU Learning Framework regularized by global consistency for Scribble supervised Cardiac Segmentation. In: Wang L, Dou Q, Fletcher PT, et al. editors. Medical Image Computing and Computer assisted intervention – MICCAI 2022. Cham: Springer Nature Switzerland; 2022. pp. 162–72.

26. Lee D-H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In: Workshop on challenges in representation learning. Atlanta: ICML; 2013. p. 896.
27. Grandvalet Y, Bengio Y. Semi-supervised learning by Entropy Minimization. Advances in neural information Processing systems. MIT Press; 2004.
28. Ke Z, Wang D, Yan Q et al. (2019) Dual Student: breaking the limits of the teacher in Semi-supervised Learning. pp 6728–36.
29. Jeong J, Lee S, Kim J, Kwak N. Consistency-based semi-supervised learning for object detection. Advances in neural information Processing systems. Curran Associates, Inc; 2019.
30. Kumar A, Rawat YS. End-to-End Semi-Supervised Learning for Video Action Detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE; 2022. p. 14680–14690.
31. Seibold CM, Reiß S, Kleesiek J, Stiefelhagen R. Reference-guided Pseudo-label Generation for Medical Semantic Segmentation. AAAI. 2022;36:2171–9. https://doi.org/10.1609/aaai.v36i2.20114.
32. Wang X, Yuan Y, Guo D, et al. SSA-Net: spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. Med Image Anal. 2022;79:102459. https://doi.org/10.1016/j.media.2022.102459.
33. Alonso I, Sabater A, Ferstl D, et al. Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-wise Memory Bank. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. p. 8199–208.
34. Yang X, Song Z, King I, Xu Z. A Survey on Deep Semi-supervised Learning. IEEE Trans Knowl Data Eng. 2022;1–20. https://doi.org/10.1109/TKDE.2022.3220219.
35. Wang K, Zhan B, Zu C, et al. Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation. In: de Bruijne M, Cattin PC, Cotin S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing; 2021. p. 450–460.
36. Luo X, Chen J, Song T, Wang G. Semi-supervised medical image segmentation through dual-task consistency. Proc AAAI Conf Artif Intell. 2021;35:8801–9.
37. Zhang Y, Jiao R, Liao Q, et al. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artif Intell Med. 2023;138:102476. https://doi.org/10.1016/j.artmed.2022.102476.
38. Yu L, Wang S, Li X, et al. Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation. In: Shen D, Liu T, Peters TM, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Cham: Springer International Publishing; 2019. p. 605–613.
39. Xiang J, Qiu P, Yang Y. FUSSNet: Fusing two sources of uncertainty for Semi-supervised Medical Image Segmentation. Medical Image Computing and Computer assisted intervention – MICCAI 2022. Cham: Springer Nature Switzerland; 2022. pp. 481–91.
40. Kwon D, Kwak S. Semi-supervised Semantic Segmentation with Error Localization Network. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE; 2022. p. 9947–57.
41. Wu Y, Xu M, Ge Z, et al. Semi-supervised Left Atrium Segmentation with Mutual Consistency Training. In: de Bruijne M, Cattin PC, Cotin S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing; 2021. p. 297–306.
42. Lalande A, Chen Z, Decourselle T, et al. Emidec: a database usable for the Automatic evaluation of myocardial infarction from delayed-enhancement Cardiac MRI. Data. 2020;5:89. https://doi.org/10.3390/data5040089.
43. Xu A, Wang S, Ye S, et al. Ca-Mt: A Self-Ensembling Model for Semi-Supervised Cardiac Segmentation with Elliptical Descriptor Based Contour-Aware. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). Kolkata: IEEE; 2022. p. 1–5.
44. Liu J, Desrosiers C, Zhou Y. Semi-supervised Medical Image Segmentation Using Cross-Model Pseudo-Supervision with Shape Awareness and Local Context Constraints. In: Wang L, Dou Q, Fletcher PT, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Cham: Springer Nature Switzerland; 2022. p. 140–150.
45. Wang Y, Wang H, Shen Y, et al. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE; 2022. p. 4238–4247.
46. Yang X, Tian J, Wan Y, et al. Semi-supervised medical image segmentation via cross-guidance and feature-level consistency dual regularization schemes. Med Phys. 2023;50(7):4269–81. https://doi.org/10.1002/mp.16217.
47. He K, Fan H, Wu Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2020. p. 9726–35.
48. Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. PMLR; 2020. p. 1597–1607.
49. Lai X, Tian Z, Jiang L, et al. Semi-Supervised Semantic Segmentation With Directional Context-Aware Consistency. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE; 2021. p. 1205–1214.
50. Wu H, Wang Z, Song Y, et al. Cross-patch Dense Contrastive Learning for Semi-supervised Segmentation of Cellular Nuclei in Histopathologic Images. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022. p. 11656–65.
51. Zhong Y, Yuan B, Wu H, et al. Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE; 2021. p. 7253–7262.
52. Liu Y, Cheng M-M, Fan D-P, et al. Semantic edge detection with Diverse Deep Supervision. Int J Comput Vis. 2022;130:179–98. https://doi.org/10.1007/s11263-021-01539-8.
53. Peiris H, Chen Z, Egan G, Harandi M. Duo-SegNet: Adversarial Dual-Views for Semi-supervised Medical Image Segmentation. In: De Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing; 2021. p. 428–38.
54. Lou A, Tawfik K, Yao X, et al. Min-Max Similarity: a contrastive Semi-supervised Deep Learning Network for Surgical Tools Segmentation. IEEE Trans Med Imaging. 2023;42:2832–41. https://doi.org/10.1109/TMI.2023.3266137.
55. Li S, Zhang C, He X. Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham: Springer International Publishing; 2020. p. 552–561.
56. Hang W, Feng W, Liang S, et al. Local and Global Structure-Aware Entropy Regularized Mean Teacher Model for 3D Left Atrium Segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham: Springer International Publishing; 2020. p. 562–571.
57. Luo X, Hu M, Song T, et al. Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. PMLR; 2022. p. 820–833.
58. Li W, Yang H. Collaborative Transformer-CNN Learning for Semi-supervised Medical Image Segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Las Vegas: IEEE; 2022. p. 1058–1065.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.