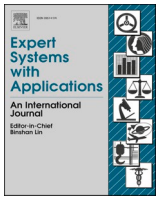
Expert Systems With Applications 244 (2024) 122903

Contents lists available at ScienceDirect

Expert Systems With Applications

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



MLC: Multi-level consistency learning for semi-supervised left atrium segmentation

Zhebin Shi [a], Mingfeng Jiang [a,*], Yang Li [a], Bo Wei [a], Zefeng Wang [b], Yongquan Wu [b], Tao Tan [c], Guang Yang [d,e]

[a] School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
[b] Department of Cardiology, Beijing Anzhen Hospital, Capital Medical University, No. 2 Anzhen Road, Chaoyang District, Beijing 100029, China
[c] Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao, 999078
[d] Cardiovascular Research Centre, Royal Brompton Hospital, London, SW3 6NP, UK
[e] National Heart and Lung Institute, Imperial College London, London, SW7 2AZ, UK

ARTICLE INFO

Keywords:
Left atrium segmentation
Semi-supervised learning
Consistency regularization
Multi-level consistency

ABSTRACT

Atrial fibrillation is the most common type of arrhythmia associated with a high mortality rate. Left atrium segmentation is crucial for the diagnosis and treatment of atrial fibrillation. Accurate left atrium segmentation with limited labeled data is a tricky problem. In this paper, a novel multi-level consistency semi-supervised learning method is proposed for left atrium segmentation from 3D magnetic resonance images. The proposed framework can efficiently utilize limited labeled data and large amounts of unlabeled data by performing consistency predictions under task level, data level, and feature level perturbations. For task consistency, the segmentation results and signed distance maps were used for both segmentation and distance estimation tasks. For data level perturbation, random flips (horizontal or vertical) were introduced for unlabeled data. Moreover, based on virtual adversarial training, we design a multi-layer feature perturbation in the structure of skipping connection. Our method is evaluated on the publicly available Left Atrium Segmentation Challenge dataset version 2018. For the model trained with a label rate of 20%, the evaluation metrics Dice, Jaccard, ASD, and 95HD are 91.69%, 84.71%, 1.43 voxel, and 5.44 voxel, respectively. The experimental results show that the proposed method outperforms other semi-supervised learning methods and even achieves better performance than the fully supervised V-Net.

1. Introduction

Atrial fibrillation (AF) is the most common type of arrhythmia among the population, and is often associated with high morbidity and mortality (Lippi, Sanchis-Gomar, & Cervellin, 2021) Existing studies have shown that there is a correlation between increased left atrial (LA) wall fibrosis and poor ablation outcome in atrial fibrillation(McGann et al., 2014). Therefore, the segmentation and analysis of the left atrium is a significant way to aid diagnosis and treatment of atrial fibrillation.

Medical image segmentation aims to classify images at the pixel level and delineate anatomical structures of interest from raw images (Gao, Zhou, & Metaxas, 2021; Han, Jian, & Wang, 2022).Traditional image segmentation methods (Shotton, Johnson, & Cipolla, 2008; Tong, Zhang, Weng, & Zhu, 2018) need to analyze the difference between the foreground and background of the image to be divided, so as to segment the image from the design features of the image's grayscale, contrast, and texture information, and the segmentation process will lose the image's semantic information. And the deep learning based segmentation methods by using a convolutional neural network (CNN) (Albawi, Mohammed, & Al-Zawi, 2017; da Silva, Silva, de Paiva, & Gattass, 2022; Niyas, Pawan, Anand Kumar, & Rajan, 2022; Su and Wang, 2019) can overcome this limitation of traditional image segmentation methods. With the emergence of U-Net (Ronneberger, Fischer, & Brox, 2015), the encoder-decoder model structure has become the mainstream semantic segmentation method of 2D medical images. After that, 3D U-Net (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016) and V-Net (Milletari, Navab, & Ahmadi, 2016) were successively used to 3D medical image segmentation. These methods have achieved remarkable results in medical image segmentation and have become a commonly used solution for this field.

* Corresponding author at: School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China.
E-mail address: m.jiang@zstu.edu.cn (M. Jiang).

https://doi.org/10.1016/j.eswa.2023.122903
Received 22 February 2023; Received in revised form 25 August 2023; Accepted 7 December 2023
Available online 21 December 2023

However, deep learning based segmentation methods often require a large amount of annotation information for training. The pixel-by-pixel annotation of the left atrium on cardiac magnetic resonance image (MRI), which relies heavily on expert experience, often requires a huge workload, meaning that such labeling work is a tricky problem. Therefore, semi-supervised learning methods (Lee et al., 2022; Yang, Song, King, & Xu, 2023; Zeng, Wang, Zhang, & Wu, 2016; Zhang et al., 2022) using samples with limited labels are of great significance.

At present, many studies have proved that the combination of different levels of consistency regularization can improve the performance of semi-supervised models. (Wang et al., 2022) combined model consistency and spatial consistency, and then proposed a dual consistency model for left atrium segmentation. (Li, Luo, Lin, Chen, & Heng, 2021). proposed a semi-supervised segmentation model combining data consistency and feature consistency for COVID-19 lesion segmentation from CT Images. Therefore, a model with more consistencies combined can make full use of a small amount of labeled information and learn the feature information of unlabeled data, thereby improving the segmentation accuracy of the semi-supervised model.

In this paper, a novel multi-level consistency (MLC) learning framework is proposed by combining task level, data level, and feature level consistency regularization. A signed distance map (SDM) with a neural network is designed to learn more geometric contour information. The task consistency is presented by using the SDM regression task and the segmentation task. This part is inspired by dual task consistency (DTC) (Luo, Chen, Song, & Wang, 2021). However, the difference is that we decouple the dual-task output of DTC into two decoders. In terms of data consistency, the use of random flips allows the network to learn multiple views of the data at the same time and also improves the robustness of the model. In order to make full use of the unlabeled data, we flip the unlabeled data randomly (horizontally or vertically) to form unlabeled sample pairs, and use an auxiliary decoder to decode the features of the images perturbed by the flip, forming a data level consistency regularization with the main segmentation decoder. The feature representation of the model can also be enriched by learning feature perturbation invariance. Therefore, we introduced virtual adversarial training (VAT) (Miyato, Maeda, Koyama, & Ishii, 2019) to perturb features before being fed to the auxiliary decoder to achieve feature level consistency. In order to improve feature richness, we designed a multi-layer feature perturbation in V-Net's skip connection structure. The proposed model consists of one shared encoder and three decoders. In order to validate the performance of the proposed framework on MRI left atrium segmentation, some comparisons are investigated with state-of-the-art (SOTA) semi-supervised learning frameworks. The main dataset used for analysis in this work is from STACOM 2018 left atrial segmentation challenge (STACOM 2018 dataset).

In summary, our main contributions are as follows

1. In order to exploit the unlabeled data and improve the robustness of the semi-supervised learning model, a multi-level consistency learning method, combining task level, data level, and feature level consistency, is proposed for 3D MRI left atrial segmentation
2. In terms of feature consistency regularization, a multi-layer feature perturbation based on VAT is designed in the skip connection structure to improve feature richness.
3. Our method is experimented with and validated on the dataset of the 2018 Left Atrial Segmentation Challenge, and the results showed that our method has better segmentation performance and robustness than the SOTA semi-supervised learning methods for LA segmentation.

The rest of this paper is organized as follows: related work is introduced in Section 2. The details of the proposed model are presented in Section 3. Section 4 presents the experiments and results. Discussion and conclusion are given in Section 5 and Section 6 respectively.

2. Related work

2.1. Semi-supervised medical image segmentation

Semi-supervised learning usually trains a model using labeled and unlabeled samples, to reduce the cost of manual labeling data. There are many methods of semi-supervised learning. Using knowledge distillation (Chen, Lopes, Cheng, Collins, Cubuk, Zoph, Adam, & Shlens, 2020; Gou, Yu, Maybank, Tao, & o. C. V., 2021; Mi, Lin, Zhou, Shen, Luo, Sun, Cao, Fu, Xu, & Ji, 2022) is a method to generate pseudo-labels for the student model through the teacher model. In order to improve the generalization of the student model, (Amirkhani, Khosravian, Masih-Tehrani, & Kashiani, 2021) proposed a multi-teacher knowledge distillation method to make the student model receive the knowledge aggregation of multiple teachers. Another conventional method is unsupervised regularization which incorporates unlabeled data into the training process by generating an unsupervised signal (such as an unsupervised loss function). And it was often used in recent semi-supervised medical image segmentation methods, such as adversarial learning, collaborative training, consistency regularization, etc. (Jiao, Zhang, Ding, Cai, & Zhang, 2022). For example, (Li, Zhang, & He, 2020) proposed a shape-aware discriminative model for distinguishing between unlabeled and labeled data segmentation result distance transform map, against the segmentation model. (Zhao et al., 2022) proposed a context-aware network and designed a discriminator with an attention mechanism to regulate model learning from unlabeled data. (Luo, Hu, Song, Wang, & Zhang, 2022) introduced cross co-training between CNN and Transformer, which implicitly encourages consistency and complementarity between the two different networks. Among them, consistency regularization is the most commonly used semi-supervised learning method. In this paper, a novel semi-supervised learning framework based on consistency regularization is proposed.

2.2. Consistency regularization

Consistency learning is to enhance the invariance of the predicted outputs by adding different perturbations to the input images. The consistency of different outputs can be calculated by Kullback-Leibler (KL) divergence, mean square error (MSE), Jensen Shannon (JS) divergence, etc. The mean teacher is a commonly used semi-supervised learning model (Tarvainen and Valpola, 2017). It creates data consistency by adding noise to the inputs and updates the parameters of the teacher model using exponential moving averages. Based on the above model, Yu et al., (Yu, Wang, Li, Fu, & Heng, 2019) proposed UA-MT, which used MC-Dropout (Kendall and Gal, 2017) to filter the unreliable information generated by the teacher model and provide reliable guidance to the student model. (Luo et al., 2021) considered the relationship between different tasks then studied the consistency between the segmentation task and the SDM regression task and proposed a DTC network. (Wu, Xu, Ge, Cai, & Zhang, 2021) used two decoders with different upsampling (trilinear interpolation and deconvolution) and used recurrent soft pseudo-labels to form a mutual consistency model (MC-Net). They also proposed a three decoders version called MC-Net+ (Wu et al., 2022), which effectively utilized the information differences in different methods for cycle consistency learning. (Liu et al., 2022) used the class information on the data, two classification models were designed and constructed after the segmentation model, the classification results of different models were compared, and a contrastive consistency loss was proposed. (Huang, Chen, Chen, Lu, & Zou, 2023) used two of three V-Nets as auxiliary models to form complementary consistency.

Although the above methods achieved promising results in semi-supervised left atrium segmentation, the semi-supervised learning models may be overfit to labeled samples, resulting in low model generalization. We believe that adding more consistency to unlabeled data can enable the model to learn more information about unlabeled

samples, thereby improving the generalization ability of the model.

2.3. Virtual adversarial training

Adversarial training (Goodfellow, Shlens, & Szegedy, 2014) is a way to enhance the robustness of neural networks. In the process of adversarial training, the samples will be mixed with some small perturbation, and then the neural network will adapt to this change, to be robust to adversarial samples. Virtual adversarial training (VAT) is adversarial training for semi-supervised learning (Chen, Xu, He, & Ban, 2022; Liu & Zheng, 2022; Miyato et al., 2019). And (Li et al., 2021; Liu et al., 2022) used VAT to add feature perturbations for Semi-supervised image segmentation. However, the decoders used by their VAT are all single-level structures. For the multi-layer feature and skip connection of V-Net, the structure of the VAT decoder needs to be redesigned. The multi-layer structure has richer feature information. Based on the above analysis, we proposed a multi-consistency semi-supervised learning approach and multi-layer feature perturbation based on VAT.

3. Method

In this section, we will introduce the proposed MLC semi-supervised learning model in detail. We elaborate on the model structure, the algorithms, and the loss functions involved. The evaluation metrics used in our work are also described.

3.1. Network architecture

The overall of network architecture is a variant of V-Net (Milletari, et al., 2016) and consists of a shared encoder and three decoders. The three decoders are the segmentation task decoder, the regression task decoder, and the auxiliary decoder, as shown in Fig. 1. Of these, the segmentation task decoder is used to obtain the expected segmented image. This is the main decoder and will be used for evaluation and testing. The SDM regression task decoder is used to obtain the SDM corresponding to the segmented image. SDM lets each voxel indicate its signed distance to the nearest boundary of the target object, which is capable of obtaining geometric activity contours and distance information and, when combined with the CNN, can guide the deep learning model to learn more information about the target contours (Dangi, Linte, & Yaniv, 2019; Park, Florence, Straub, Newcombe, & Lovegrove, 2019; Sukumar and Srivastava, 2022). The auxiliary decoder is used for further consistency on unlabeled samples, forming feature consistency regularization and data consistency regularization with the main segmentation decoder.

In this network structure, the supervised learning part is jointly guided by the original label data and the transformed SDM. The binary segmentation label generates an SDM with target shape contour information through a signed distance function (SDF), which then serves as the label data for the SDM regression task decoder. The formula is shown as follows:

$$SDF(x) = \begin{cases} -inf\|x-y\|_2, x \in S_{in}, y \in \partial S \\ 0, x \in \partial S, y \in \partial S \\ +inf\|x-y\|_2, x \in S_{out}, y \in \partial S \end{cases} \quad (1)$$

where x, and y respectively represent two different voxels in the label mask, the ∂S is the object boundary, S_{in} and S_{out} denote the interior of the object and the exterior of the object respectively.

In this network, the parameters of the shared encoder, main segmentation decoder, and SDM decoder are updated by back propagation. Considering the similarity of the tasks of the main segmentation decoder and auxiliary and the increase in the number of learning parameters carried by the multiple decoders, we used an exponential moving average (EMA) for the auxiliary decoder to update the parameters. The EMA is shown as formula (2), where α is the momentum, $\theta^{'}$ is the parameters of the auxiliary decoder, θ is the parameters of the main segmentation decoder, and t is the number of iterations. In this work, α is set to 0.99.

$$\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t \quad (2)$$

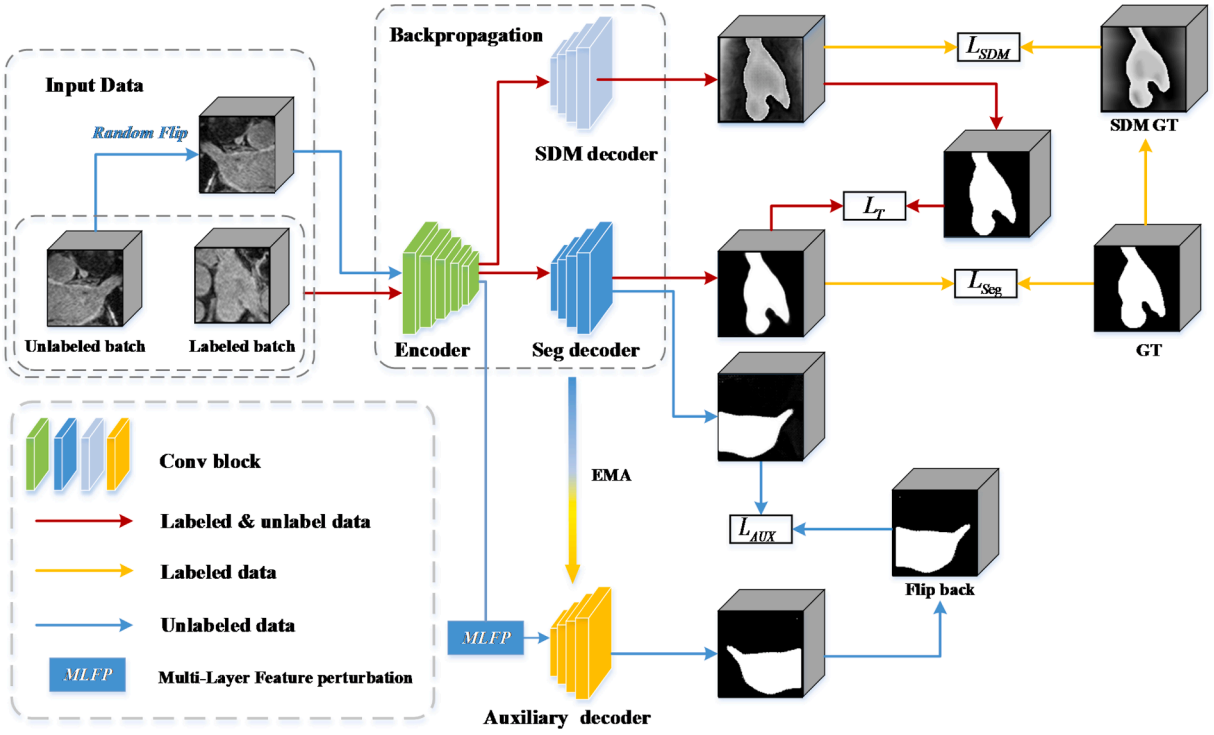


Fig. 1. Proposed Multi-Level Consistency Semi-Supervised Learning Network. The network consists of a shared encoder(green) and three decoders, the three decoders are the main segmentation task decoder(deep blue), the SMD regression task decoder(light blue), and the auxiliary decoder(yellow).

3.2. Multi-level consistency

Our proposed multi-level consistency semi-supervised learning framework combines task consistency, data consistency, and feature consistency. As shown in the data flow in Fig. 1, task level consistency acts on labeled and unlabeled data (red arrows). The SDM for regression task output is inversely transformed, and the MSE is calculated with the output of the segmentation task to obtain the task level consistency loss result. According to the deduction results provided by (Luo et al., 2021), the SDF inversion is shown as formula (3), where z denotes SDM outputs and k takes -1500 in our framework.

$$SDF^{-1}(z) = sigmoid(k{\cdot}z) \tag{3}$$

In order to learn more information about unlabeled data, we used an auxiliary decoder composed of consistency regularization with the main segmentation decoder. As shown in the blue arrow in Fig. 1, this part only learns unlabeled samples. Here, we added data perturbation and feature perturbation, which can be seen as superimposing data consistency and feature consistency at the same time. In terms of data consistency, we randomly flip (horizontally or vertically) the unlabeled input data to form unlabeled sample pairs, which are fed into the main segmentation decoder and the auxiliary decoder after the shared encoder. After the output of the auxiliary decoder is flipped back to the orientation of the original image, the MSE loss is computed with the output of the main segmentation decoder. For feature consistency, we add perturbations to the features before passing them into the auxiliary decoder. Combined with the multi-layer feature fusion structure of the V-Net skip connection, we design a multi-layer feature perturbation based on VAT.

3.3. Multi-layer feature perturbation

For the model, it is hoped that the predicted results of the same sample with perturbation and without perturbation are consistent. As for the perturbation, it maximizes the difference between the two predictions in terms of perturbation as possible, so it must be found a r_{adv} that makes the difference between the two predictions largest. VAT (Miyato, et al., 2019) encourages models to reduce such gaps, thereby increasing robustness. In adversarial training, noise is not directly added to the input, but the gradient direction of model optimization is perturbed. The calculation of is r_{adv} shown in the formula (4) and (5). Where d is an initial random noise, $p(\cdot)$ is the predictive model which is the auxiliary decoder in our work, x is input, y is expected target, $\widehat{\theta}$ is the parameters of prediction model, $D(\cdot)$ is a discriminant function like KL-divergence, $\xi and \in$ considered as hyperparameters.

$$g = \nabla_r D[p(y|x,\widehat{\theta}), p(y|x+r,\widehat{\theta})]|_{r=\xi d} \tag{4}$$

$$r_{adv} \approx \in \frac{g}{\|g\|_2} \tag{5}$$

VAT is an effective way to generate adversarial feature perturbation. The V-Net used in our model is an encoder-decoder structure with skip connections which is to fuse feature information, so as to fuse deep semantic information and shallow semantic information. Based on this, we designed a multi-layer feature perturbation method with VAT, as shown in Fig. 2. The multi-layer features from the encoder are fused with the features of each layer of the decoder after adding perturbations. Such multi-layer feature perturbation can improve feature richness and avoid overfitting. The process of generating multi-layer feature perturbation vector is shown in the Fig. 3. The specific algorithm is shown in Algorithm 1.

Algorithm 1: Multi-layer feature perturbation based on VAT

Input: Multi-layer features $fs = \{fs_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i\times D_i} | i = 1,2,\cdots,N\}$ where N is the number of feature layers, decoder $\mathbb{D}(\cdot)$, ξ, $\in$

Output: perturbation value r_{adv}

1. Generate a random unit vector list $\boldsymbol{d} = \{d_i | i = 1,2,\cdots,N\}$ which has the same shape as fs
2. $\boldsymbol{r} = \xi \boldsymbol{d}$
3. $\boldsymbol{g} = \nabla_r KL(\mathbb{D}(\boldsymbol{fs}), \mathbb{D}(\boldsymbol{fs}+\boldsymbol{r}))$
4. $r_{adv} = \left[\in \frac{g_i}{\|g_i\|_2} forg_i ing\right]$
5. **return** r_{adv}

3.4. Loss function

Formally, we denote the labeled dataset as $\mathscr{D}_l = \{x_n^l, y_n\}_{n=1}^N$ and unlabeled dataset as $\mathscr{D}_u = \{x_m^u\}_{m=1}^M$, $x = x^l \cup x^u$ where $x \in \mathbb{R}^{H\times W\times D}$ are the input volumes, $y \in \{0,1\}^{H\times W\times D}$ are the annotations. The loss function in the network training process is divided into supervised loss and consistency loss. Among them, there are supervised losses including segmentation loss $\mathscr{L}_{seg}$ and SDM loss $\mathscr{L}_{SDM}$, as shown in the formula (6).The definition of $\mathscr{L}_{seg}$ and $\mathscr{L}_{SDM}$ are shown in formula (7) and formula (8). $\mathbb{E}(\cdot)$ is encoder, $\mathbb{D}_{seg}(\cdot)$ is segmentation task decoder, $\mathbb{D}_{sdm}(\cdot)$ is

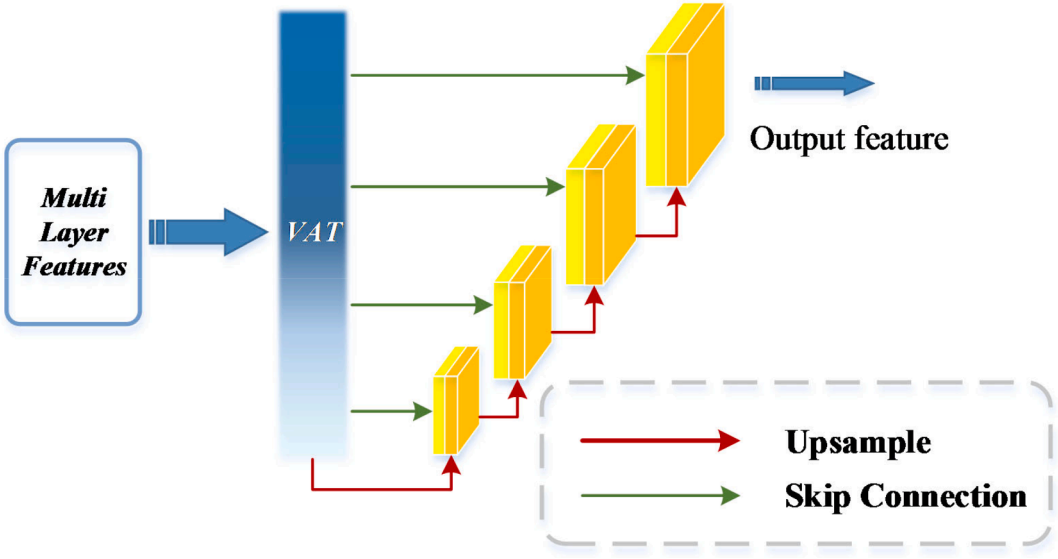


Fig. 2. Multi-layer feature perturbation. The multi-layer features from the encoder are fused with the features of each layer of the decoder after adding perturbations.

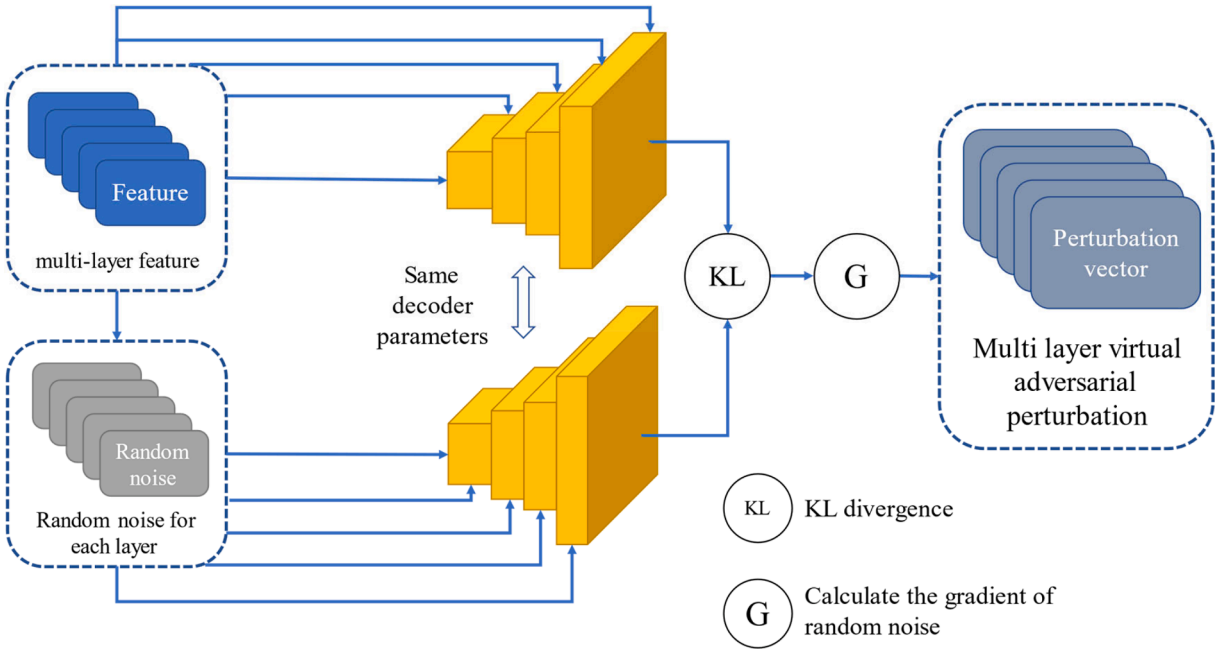


Fig. 3. Multi-layer feature perturbation based on VAT detailed process.

SDM regression task decoder, $\mathbb{D}_{aux}(\cdot)$ is the auxiliary decoder. β is set to 0.3 in this work.

$$\mathscr{L}_{sup} = \mathscr{L}_{seg} + \beta \mathscr{L}_{SDM} \tag{6}$$

$$\mathscr{L}_{seg}(x^l, y) = \sum_{x_i^l, y_i \in \mathscr{D}_l} \left(1 - Dice\left(\mathbb{D}_{seg}\left(\mathbb{E}\left(x_i^l\right)\right), y_i\right)\right) \tag{7}$$

$$\mathscr{L}_{SDM}(x^l, y) = \sum_{x_i^l, y_i \in \mathscr{D}_l} \left\| \mathbb{D}_{sdm}\left(\mathbb{E}\left(x_i^l\right)\right) - SDF(y_i) \right\|^2 \tag{8}$$

The consistency loss $\mathscr{L}_{unsup}$ consists of two parts, one is the task loss $\mathscr{L}_T$ between the segmentation decoder and the regression task decoder, and another one is the loss $\mathscr{L}_{AUX}$ between the main segmentation decoder and the auxiliary decoder. The definitions of both losses are shown in formula (10) and formula (11). $\mathscr{T}(\bullet)$ and $\mathscr{T}^{-1}(\cdot)$ denote random flip transformation and its corresponding inverse transformation.

$$\mathscr{L}_{unsup} = \mathscr{L}_T + \mathscr{L}_{AUX} \tag{9}$$

$$\mathscr{L}_T(x) = \sum_{x_i \in (\mathscr{D}_l \cup \mathscr{D}_u)} \left\| \mathbb{D}_{seg}(\mathbb{E}(x_i)) - SDF^{-1}(\mathbb{D}_{sdm}(\mathbb{E}(x_i))) \right\|^2 \tag{10}$$

$$\mathscr{L}_{AUX}(x^u) = \sum_{x_i^u \in \mathscr{D}_u} \left\| \mathbb{D}_{seg}\left(\mathbb{E}\left(x_i^u\right)\right) - \mathscr{T}^{-1}\left(\mathbb{D}_{aux}\left(\mathbb{E}\left(\mathscr{T}\left(x_i^u\right)\right)\right)\right) \right\|^2 \tag{11}$$

The overall loss function of the model is shown in formula (12), where γ is a hyperparameter. According to the ablation study, the model works best when we set it to 0.9. The effect of the value of γ on the performance of the model will be discussed in 5.1. λ is a Gaussian warming-up function with respect to iterations t, shown in formula (13), t_{max} is the maximum iteration number.

$$\mathscr{L}_{total} = \gamma \mathscr{L}_{sup} + \lambda \mathscr{L}_{unsup} \tag{12}$$

$$\lambda(t) = e^{\left(-5\left(1 - \frac{t}{t_{max}}\right)^2\right)} \tag{13}$$

The semi-supervised segmentation training algorithm based on multi-level consistency is shown in Algorithm 2.

Algorithm 2: Semi-supervised training through Multi-level consistency

Input: $x \in \{\mathscr{D}_l \cup \mathscr{D}_u\}$, $y \in \mathscr{D}_l$
Output: MLC model's parameter θ_e for encoder, θ_{seg} for segmentation decoder, θ_{sdm} for SDM decoder, θ_{aux} for auxiliary decoder
1. initialize θ_e, θ_{seg}, θ_{sdm}, θ_{aux}
2. $\mathbb{E}(\cdot)$ is the encoder with parameter θ_e
3. $\mathbb{D}_{seg}(\cdot)$ is the segmentation decoder with parameter θ_{seg}
4. $\mathbb{D}_{sdm}(\cdot)$ is the SDM decoder with parameter θ_{sdm}
5. $\mathbb{D}_{aux}(\cdot)$ is the auxiliary decoder with parameter θ_{aux}
6. $\mathscr{T}(\bullet)$ is random flip transform
7. $\mathscr{T}^{-1}(\cdot)$is the inverse transformation relative to $\mathscr{T}(\bullet)$
8. while *do not reach the maximum iterations* **do**
9. Sample batch $b_l = (x^l, y) \in \mathscr{D}_l$, $b_l = x^u \in \mathscr{D}_u$, $b = b_l \cup b_u$, $x = x^l \cup x^u$, where b_l, b_u, b denote labeled batch, unlabeled batch and total batch respectively
10. Generate SDM ground truth $SDF(y)$
11. Compute dual task prediction $\mathbb{D}_{seg}(\mathbb{E}(x_i))$ and $\mathbb{D}_{sdm}(\mathbb{E}(x_i))$, $i \in \{1, \cdots, |b|\}$
12. Generate random flip map of unlabeled data $\mathscr{T}(x_i^u)$, $i \in \{1, \cdots, |b_u|\}$
13. Compute unlabeled data' s multi-layer features $mlf = \mathbb{E}(\mathscr{T}(x_i^u))$
14. Add perturbation to mlf through VAT
15. Decode mlf $\mathbb{D}_{aux}(mlf)$
16. Get the inverse transformation of random flip $\mathscr{T}^{-1}(\mathbb{D}_{aux}(mlf))$
17. $\mathscr{L}_{seg}(x^l, y) = \frac{1}{|b_l|}\sum_{x_i^l, y_i \in b_l}(1 - Dice(\mathbb{D}_{seg}(\mathbb{E}(x_i^l)), y_i))$
18. $\mathscr{L}_{SDM}(x^l, y) = \frac{1}{|b_l|}\sum_{x_i^l, y_i \in b_l} \| \mathbb{D}_{sdm}(\mathbb{E}(x_i^l)) - SDF(y_i) \|^2$
19. $\mathscr{L}_T(x) = \frac{1}{|b|}\sum_{x_i \in b} \| \mathbb{D}_{seg}(\mathbb{E}(x_i)) - SDF^{-1}(\mathbb{D}_{sdm}(\mathbb{E}(x_i))) \|^2$
20. $\mathscr{L}_{AUX}(x^u) = \frac{1}{|b_u|}\sum_{x_i^u \in b_u} \| \mathbb{D}_{seg}(\mathbb{E}(x_i^u)) - \mathscr{T}^{-1}(\mathbb{D}_{aux}(mlf)) \|^2$
21. $\mathscr{L}_{sup} = \mathscr{L}_{seg} + \beta \mathscr{L}_{SDM}$
22. $\mathscr{L}_{unsup} = \mathscr{L}_T + \mathscr{L}_{AUX}$
23. $\mathscr{L}_{total} = \gamma \mathscr{L}_{sup} + \lambda_t \mathscr{L}_{unsup}$
24. Compute gradient of $\mathscr{L}_{total}$ and update parameters θ_e, θ_{seg}, θ_{sdm} by back propagation
25. update parameter θ_{aux} by exponential move average
26. end
27. return θ_e, θ_{seg}, θ_{sdm}, θ_{aux}

3.5. Metrics

In this work, we used four evaluation metrics to assess the model performance, which are the Dice coefficient (Dice, 1945), Jaccard coefficient (Liu et al., 2012), average surface distance (ASD) (Heimann

et al., 2009), and Hausdorff distance (95HD) (Gerig, Jomier, & Chakos, 2001) respectively. The Dice coefficient and Jaccard coefficient are used to measure the similarity between the two sets, and the value range is [0 %, 100 %]. The formulas are shown as follows:

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \times 100\% \tag{14}$$

$$\text{Jaccard} = \frac{|P \cap G|}{|P \cup G|} \times 100\% \tag{15}$$

where P is the set of a voxel of the predicted target, and G is the set of pixel points of the target label.

The ASD and 95HD are used to measure the distance between the two sets, and the value range is $[0, +\infty)$. The formulas are shown as follows:

$$ASD = \frac{1}{S(P) + S(G)} \left(\sum_{s_P \subseteq s(P)} d(S_P, S(G)) + \sum_{s_G \subseteq s(G)} d(S_G, S(P)) \right) \tag{16}$$

$$95HD = 0.95 \times \max\left\{ max_{g \subseteq G}\left\{ min_{p \subseteq P} d(g,p) \right\}, max_{p \subseteq P}\left\{ min_{g \subseteq G} d(g,p) \right\} \right\} \tag{17}$$

where $s(\cdot)$ denotes the set of voxels of the predicted surface, and $d(\cdot)$ denotes the shortest Euclidean distance of a voxel to another surface.

4. Experiments and results

In this section, we presented experimental results and compared them with other SOTA methods on four evaluation metrics. Robustness analysis and visualization analysis are carried out at the same time, which fully demonstrates the effectiveness of our method.

4.1. Dataset

The dataset for this work comes from the STACOM 2018 Left Atrial Segmentation Challenge (Xiong, et al., 2021). (https://atriaseg2018.cardiacatlas.org/). The dataset mostly was provided by the University of Utah (NIH/NIGMS Center for Integrative Biomedical Computing (CIBC)), with the remainder from various other institutes, including 154 3D late gadolinium-enhanced magnetic resonance images (LGE MRI) from 60 patients with atrial fibrillation. 100 samples of them are the training set. The spatial resolution of the 3D LGR MRI scan is 0.625 × 0.625 × 0.625 mm^3, and the spatial size is 576 × 576 × 88 or 640 × 640 × 88. For a fair comparison, we divided the 100 samples into 80 training samples and 20 verification samples using the same data partition method as previous works (Huang, et al., 2023; Li, et al., 2020; Luo et al., 2021; Wu et al., 2022; Wu et al., 2021; Yu et al., 2019).

4.2. Implementation details

We implemented the proposed framework under python 3.9.12 and torch 1.11.0 using an NVIDIA RTX 3090 GPU and an Intel XEON Gold 6226R CPU. During the training process, we trained a total of 15,000 iterations, the learning rate was constant at 0.1, the batch size was 4, and the random number seed was 1337. SGD optimizer was used for the network training with a momentum of 0.9. And we randomly cropped the 3D volume into 112 × 112 × 80 as the input of the network. Random flip and random rotation were used in the image preprocessing stage.

Since the model receives a size of 112 × 112 × 80, during training, and the real MRI volume is larger than this size, we used a sliding window strategy with a stride of 18 × 18 × 4 and a patch size of 112 × 112 × 80 to obtain the final result in the inference phase. For the overlapping parts of the sliding windows, we take the average. In order to deal with some fragmented blocks that may appear in the segmentation results, we used non-maximum suppression (NMS)(S. Li, et al., 2020) to retain the largest connected area in the post-processing stage, as shown in Fig. 4.

4.3. Comparative experiments

Our model is experimented with 20 % labeled data and compared with several six SOTA methods including UA-MT (Yu et al., 2019), SASSNet (S. Li, et al., 2020), DTC (Luo et al., 2021), MC-Net (Yicheng Wu et al., 2021), MC-Net+ (Y. Wu et al., 2022), and CC-Net (H. Huang, et al., 2023). And all compared methods were implemented on the optimal hyperparameter settings as section 4.2. The experiment results are shown in Table 1. The first two rows of the table show the performance of the fully supervised V-Net at 100 % labels and 20 % labels. Note that, the model complexities, i.e., the number of parameters (Para.) and multiply-accumulate operations (MACs), are measured during the model training. From the results in the Table 1, it can be seen that our proposed method is significantly better than the SOTA methods in four evaluation metrics. The valuation metrics Dice, Jaccard, ASD, and 95HD are 91.69 %, 84.71 %, 1.43 voxel, and 5.44 voxel respectively. The performance even outperforms a fully supervised V-Net.

In addition, we trained the model with five different labeled ratios (10 %, 20 %, 30 %, 40 %, 50 %, and 60 %) and compared it with the two current best methods (MC-Net+ (Y. Wu et al., 2022) and CC-Net (Huang, et al., 2023)) in this filed. As shown in Fig. 5. Note that, with a 10 % label, our results are not better than CC-Net. Considering that CC-Net uses three full-size V-Nets for training, it can be seen from the Table 1 that CC-Net has a higher computational complexity, so our model still has advantages. Except this, our model outperforms the other two methods at other annotation rates (from 20 % to 60 %). At the same time, it can be seen from Fig. 5 that these semi-supervised learning methods have the highest benefit when the labeling rate is 20 %. The performance of the model is poor at a labeling rate of 10 %, and the improvement of the model is small with a labeling rate of 30 % and above.

4.4. Robustness analysis

To further validate the robustness of the model, we used boxplots to compare the distribution of DTC (Luo et al., 2021), MC-Net+ (Wu et al., 2022), CC-Net (Huang, et al., 2023), and the proposed model in the segmentation results of 20 test cases, as shown in Fig. 6. And it can also be clearly seen that the segmentation results obtained by our method are not only superior to the average and median in the four evaluation metrics but also have a relatively compact data distribution, which is significantly improved compared with the other three methods. Therefore, it can be considered that our method has a better robust performance on the task of LA segmentation.

In addition, we conduct experiments on another LGE-MRI left atrial segmentation dataset which is provided by LAScarQS 2022: left atrium and scar quantification and segmentation challenge (Li, Zimmer, Schnabel, & Zhuang, 2021, 2022a,b) (https://zmiclab.github.io/projects/lascarqs22/index.html). In the second task of this challenge, 130 cases of 3D cardiac MRI were provided. We randomly split into 104 training samples and 26 testing samples (4:1). During training, 20 samples are used as labeled data (about 20 % labeling rate), and the relevant hyperparameter settings are consistent with 4.2.

We compared the performance of DTC (Luo et al., 2021), MC-Net+ (Wu et al., 2022), CC-Net (Huang, et al., 2023) on this dataset. The valuation metrics of Dice, Jaccard, ASD, and 95HD of our model are 89.92 %, 81.76 %, 1.14 voxel, and 5.63 voxel respectively. As can be seen from the Table 2, our three metrics in Dice Jaccard and 95HD are significantly better than other methods. At the same time, it can be seen from the Fig. 7, the Dice box plot of the 26 test samples in the data set that our result data distribution is more concentrated. In all, the proposed model is also suitable for another datasets segmentation, which can provide robust and good segmentation results.

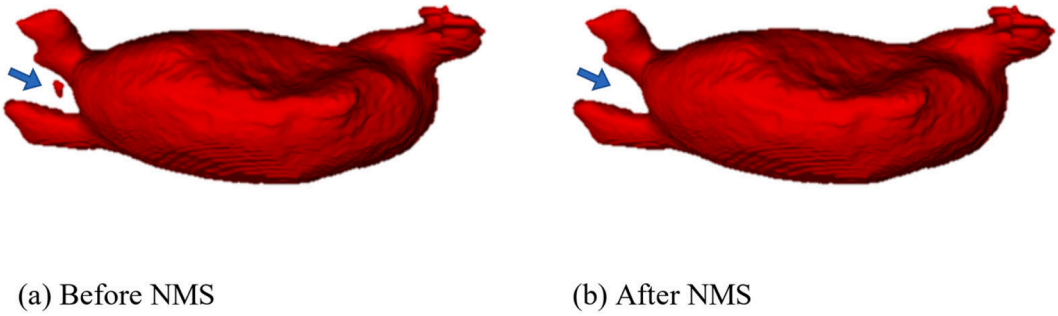


Fig. 4. NMS used in inference phase.

Table 1
Comparative experiment results under 20% labeled data.

Methods	Scans used		Metrics				Complexity	
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	ASD[voxel]↓	95HD[voxel]↓	Para.(M)	MACs(G)
V-Net	80	0	91.62	84.6	1.64	5.40	9.45	188.73
V-Net	16	0	86.96	77.31	3.22	11.85	9.45	188.73
UA-MT (Yu et al., 2019)	16	64	89.31	80.86	1.97	7.49	18.0.90	377.46
SASSNet (Li, et al., 2020)	16	64	89.73	81.52	1.88	7.07	20.47	250.51
DTC(Luo et al., 2021)	16	64	90.06	82.07	1.74	7.12	9.45	188.73
MC-net (Wu et al., 2021)	16	64	90.92	83.45	1.86	6.52	12.35	380.39
MC-net+ (Wu et al., 2022)	16	64	91.07	83.67	1.67	5.84	15.25	572.23
CC-Net (Huang, et al., 2023)	16	64	91.34	84.14	1.60	5.61	28.35	566.19
Ours	16	64	**91.69**	**84.71**	**1.43**	**5.44**	13.59	519.77

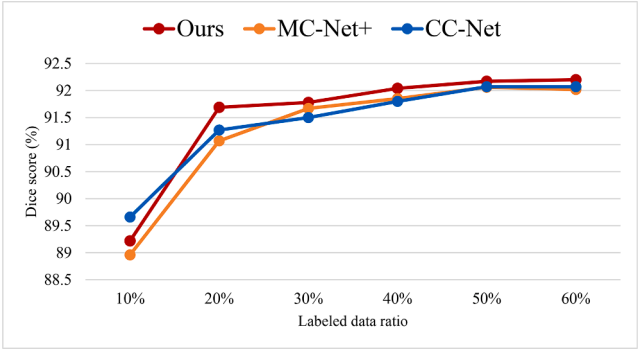


Fig. 5. Comparison results with different labeled rates of MC-Net + and CC-Net.

4.5. Visualization analysis

To further compare the segmentation performance of the models, we show 2D slices and 3D reconstructions of segmentation results on the STACOM dataset, as illustrated in Fig. 8. The visualization analysis used ITK-SNAP (Yushkevich et al., 2006) (https://www.itksnap.org/) to reconstruct the segmentation results into 3D graphics. Compared the segmentation results obtained by our model with these results by using UA-MT (Yu et al., 2019), SASSNet (Li, et al., 2020), DTC (Luo et al., 2021), MC-Net+ (Wu et al., 2022), and CC-Net (Huang, et al., 2023). It can be found that our model can avoid regions that are not expected to be segmented well (blue arrows, orange arrows) in 2D slice results. Moreover, from the 3D segmentation results, it can be seen that the segmentation results of our model have a fuller structure than other models at the pulmonary vein (blue circle). In addition, in the two 3D cases by using MC-Net + and CC-Net, the segmentation results have some defects. As the area indicated by the blue arrow, the SASS and DTC protrude from the left atrial cavity in case 1. In this case 2, DTC lacked the left atrial appendage, and there were depressions in the left atrial cavity of MC-NET + and CC-Net. On the whole, the segmentation results of our model are closer to the ground truth as a whole, and the segmentation effect is better.

4.6. Comparison with methods from the 2018 challenge

Recently, data from 54 cases evaluated in the 2018 left atrial segmentation challenge have become accessible and the results of the challenge have been published in (Xiong et al., 2021). We compared the current SOTA method for semi-supervised left atrium segmentation on these 54 samples (using 20 % labeled samples, a total of 80 training samples to obtain the pre-trained model). At the same time, we selected several top-ranking methods in the challenge for comparison, such as two-stage 3D U-Net segmentation with ranking 1 (Xia, Yao, Hu, & Hao, 2019), 3D U-Net with additional residual connections, dense connections and dilated convolutions with ranking 2 (Huang, 2018), 2D Dilated ResNet with spatial pyramid pooling with ranking 3 (Bian, Yang, Ma, Zheng, Liu, Nezafat, Heng, & Zheng, 2019) and 3D U-Net with distance maps with ranking 10 (Jia, Despinasse, Wang, Delingette, Pennec, Jaïs, Cochet, & Sermesant, 2019). However, these methods are all fully supervised. It can be seen from Table 3, the results that the semi-supervised method still has a lot of gaps compared with these excellent fully supervised methods. Given that the semi-supervised methods only use 20 % of the annotation rate, this gap is predictable. However, our method and MC-Net + and CC-Net are comparable to the tenth place, which also shows that the current semi-supervised segmentation methods are promising. Our method overall outperforms other semi-supervised methods on these 54 samples, and it can be seen that our method overall is superior to other SOTA semi-supervised left atrium segmentation methods.

5. Discussion

5.1. Ablation study

In order to verify the effectiveness of the model we proposed, we designed a total of four experimental schemes: (1) Dual Task Decoder (DTD), (2) Triple decoder with only data perturbation (TrD + DP), (3)

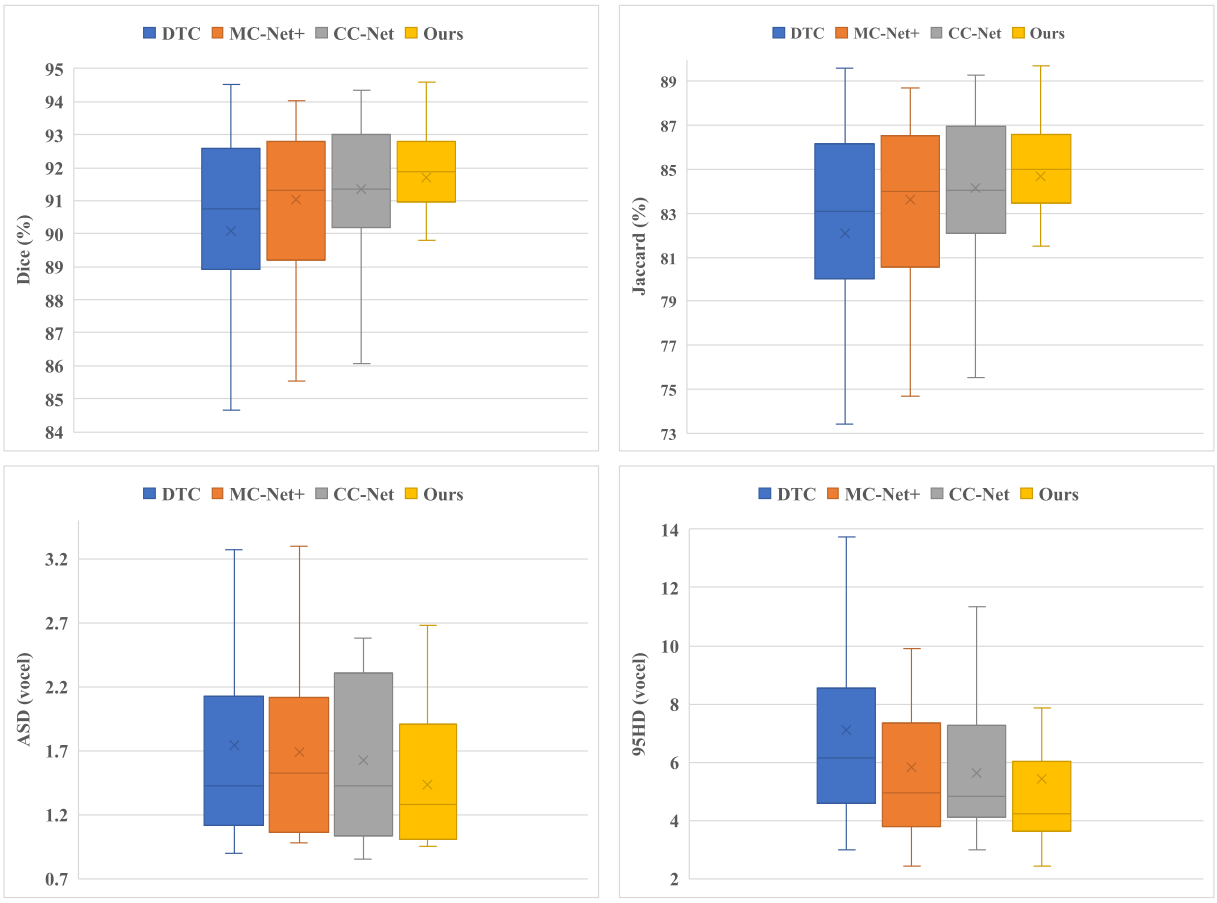


Fig. 6. Boxplots of segmentation results for 20 test cases.

Table 2
Comparative experiment results under 20% labeled data on the LAScarQS 2022 dataset.

Methods	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	ASD[voxel]↓	95HD[voxel]↓
DTC (Luo et al., 2021)	20	84	85.11	75.06	1.52	9.12
MC-net+ (Y. Wu et al., 2022)	20	84	86.28	76.67	1.49	8.57
CC-Net (H. Huang, et al., 2023)	20	84	89.31	80.78	**1.10**	6.73
Ours	20	84	**89.92**	**81.76**	1.14	**5.63**

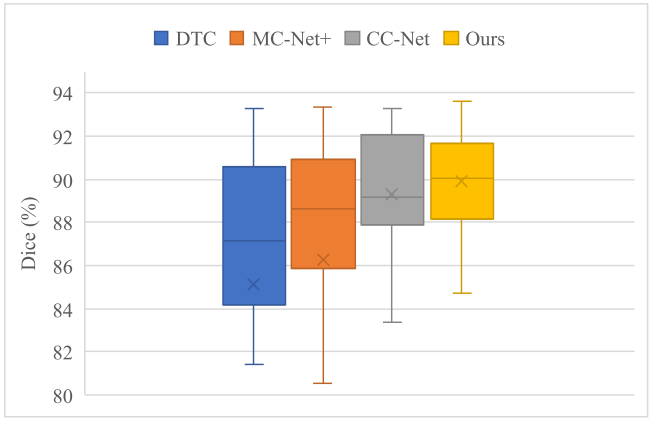


Fig. 7. Boxplot of segmentation Dice score for 26 test cases on the LAScarQS 2022 dataset.

Triple decoder with only feature perturbation (TrD + FP), (4)The network we proposed(MLC). Among them, the DTD includes a segmentation task decoder and an SDM regression task decoder. And the TrD is to add an auxiliary decoder to DTD. Since the task consistency part of our network model is based on DTC (Luo et al., 2021) with decoder decoupling of dual-task outputs, we also compared it in ablation experiments. The results are shown in Table 4, These results confirm the effectiveness of the decoupling decoder and the improvement of model performance after adding data perturbation and feature perturbation.

According to formula (12), in order to explore the influence of supervised loss weight γ on the performance of the model, we took different values of γ (0.3, 0.5, 0.7, 0.9, 1) for ablation experiments. The Dice score and Jaccard score of the experimental results are shown in Fig. 9. When the γ is set to 0.9, the model performed best.

5.2. Limitations and future work

Although our method outperforms most SOTA methods at labeling rates of 20 % and above, there is still some space for improvement. As can be seen in Fig. 5, our model can achieve state-of-the-art performance

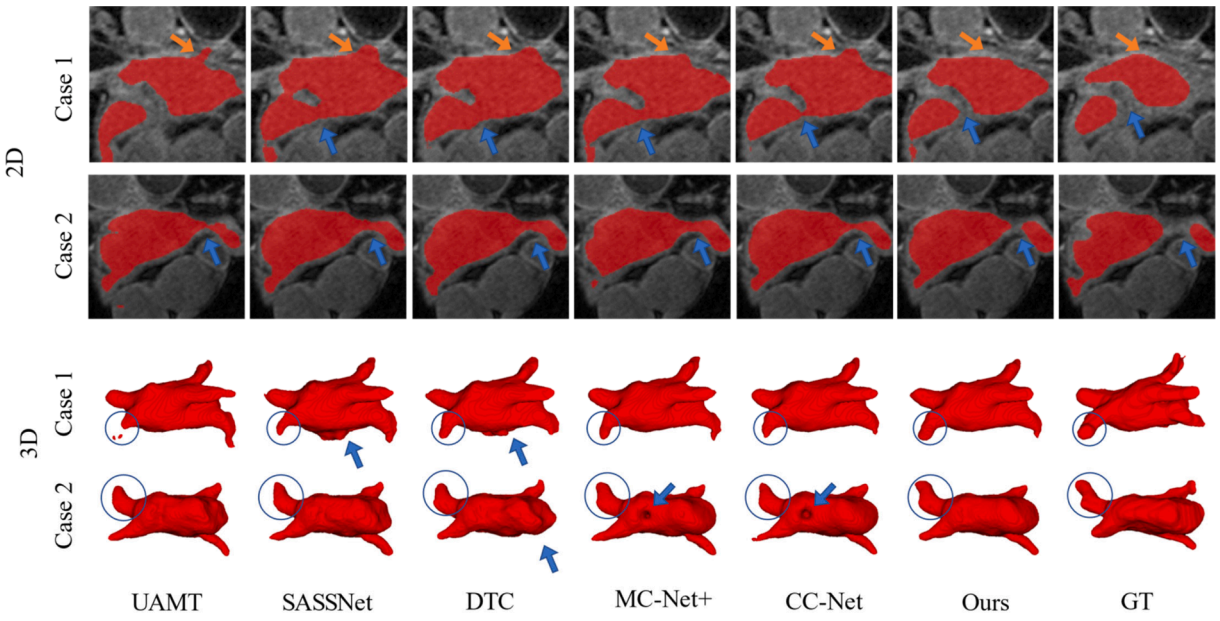


Fig. 8. 2D and 3D visualization of segmentation results on the STACOM 2018 dataset.

Table 3
Comparison of semi-supervised and fully supervised methods on 54 test samples of STACOM 2018 dataset.

Methods		Metrics			
		Dice(%)↑	Jaccard(%)↑	ASD[voxel]↓	95HD[voxel]↓
Full-supervised partial methods for the 2018 challenge	(Xia, et al., 2019) (Rank. 1)	**93.2**	**87.4**	**0.75**	8.89
	(Huang, 2018) (Rank. 2)	93.1	87.2	0.75	**8.50**
	(Bian, et al., 2019) (Rank. 3)	92.6	86.9	0.76	9.21
	(Jia, et al., 2019) (Rank. 10)	90.7	83.2	1.08	10.68
Semi-supervised 16 labeled samples for training	UA-MT (Yu et al., 2019)	88.64	79.84	3.15	11.12
	SASS (Li, et al., 2020)	88.37	79.66	2.41	8.77
	DTC (Luo et al., 2021)	87.77	78.85	2.43	8.96
	MC-net+ (Wu et al., 2022)	90.92	83.47	1.74	6.63
	CC-Net (Huang, et al., 2023)	90.80	83.30	1.52	**5.97**
	Ours	**91.03**	**83.67**	**1.51**	6.12

Table 4
Ablation study results of multi-level consistency.

Methods	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	ASD[voxel]↓	95HD[voxel] ↓
DTC(Luo et al., 2021)	16	64	90.06	82.07	1.74	7.12
DTD	16	64	90.15	82.21	1.60	7.57
TrD + FP	16	64	90.91	83.40	1.43	6.32
TrD + DP	16	64	91.01	83.57	1.94	6.02
MLC	16	64	**91.69**	**84.71**	**1.43**	**5.44**

when the label rate is 20 % and above. But in the case of a 10 % labeling rate, the performance of our model is not as good as the current best methods CC-Net (Huang, et al., 2023). Considering that the auxiliary decoder in MLC uses unlabeled data, we guess that when the label rate is only 10 %, the auxiliary decoder misleads the training of the main segmentation decoder. On the other hand, using VAT (Miyato, et al., 2019) allows the same decoder to have multiple forward propagation operations, which will cause time loss. So, in our model, EMA is used instead of backpropagation for the auxiliary decoder to update parameters to optimize the time consumption in model training.

The reason for the poor performance of the 10 % labeling model analyzed above, several studies (Kim et al., 2022; Li et al., 2021; Wang et al., 2022; Wang, Zhang, Tian, Zhong, Shi, Zhang, & He, 2020; Yu et al., 2019) have demonstrated that uncertainty-guided approaches can deal with such a situation. Our follow-up work will also start with this aspect to optimize the performance of the model and make the model achieve better performance with fewer labeled samples. Meanwhile, we will also try to apply our semi-supervised model to other medical image segmentation tasks.

6. Conclusion

In this paper, we proposed a multi-level consistency (MLC) semi-supervised learning method for left atrium segmentation in 3D MRI images. Three levels of consistency in tasks, data, and features to make the model more fully learn information from unlabeled data and a VAT-based multi-layer feature perturbation is proposed for feature consistency in our work. The proposed network achieves better performance

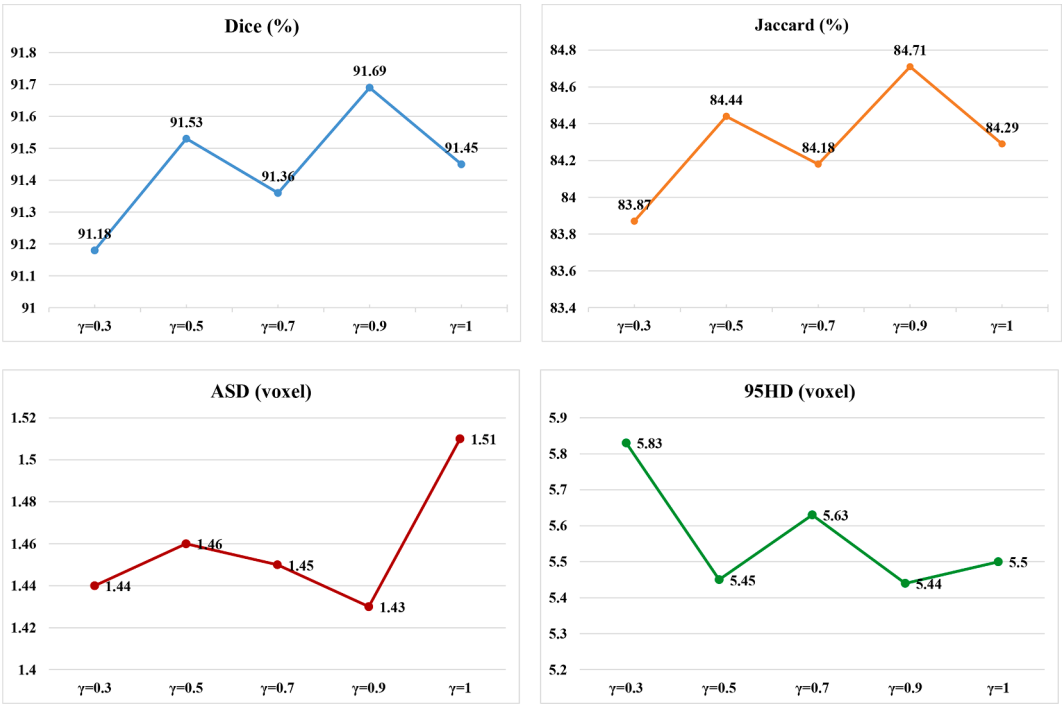


Fig. 9. Performance of multi-level consistency method under different γ values.

and robustness compared with other methods at a 20 % labeling rate. On the STACOM 2018 challenge dataset our model achieves 91.69 %, 84.71 %,1.43 voxel and 5.44 voxel in Dice, Jaccard, ASD and 95HD. On the LAScarQS2022 challenge dataset, our model has better results than other methods, which proves that our method has better generalization ability. Our work provides an alternative for semi-supervised learning of medical image segmentation.

CRediT authorship contribution statement

Zhebin Shi: Writing – review & editing, Methodology, Software. **Mingfeng Jiang:** Supervision, Writing – original draft, Formal analysis, Funding acquisition. **Yang Li:** Writing – review & editing. **Bo Wei:** Data curation. **Zefeng Wang:** Resources, Data curation. **Yongquan Wu:** Validation, Data curation. **Tao Tan:** Validation. **Guang Yang:** Writing – review & editing, 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.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (62011530130 and 62272415), the Key Research and Development Program of Zhejiang Province (2020C03060), the Joint Fund of Zhejiang Provincial Natural Science Foundation (LSZ19F010001), the General Scientific Research Project of the Department of Education of Zhejiang Province (Y202250302), the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), and the UKRI Future Leaders Fellowship (MR/V023799/1).

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Amirkhani, A., Khosravian, A., Masih-Tehrani, M., & Kashiani, H. J. I. A. (2021). Robust semantic segmentation with multi-teacher knowledge distillation. *IEEE Access, 9*, 119049–119066. https://doi.org/10.1109/ACCESS.2021. 3107841

Bian, C., Yang, X., Ma, J., Zheng, S., Liu, Y.-A., Nezafat, R., Heng, P.-A., & Zheng, Y. (2019). Pyramid network with online hard example mining for accurate left atrium segmentation. international workshop on statistical atlases and computational models of the heart. Cham. https://doi.org/10.1007/978-3-030-12029-0_26.

Chen, L.-C., Lopes, R. G., Cheng, B., Collins, M. D., Cubuk, E. D., Zoph, B., Adam, H., & Shlens, J. (2020). Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. https://doi.org/10.1007/978-3-030-58545-7_40.

Chen, Y., Xu, K., He, D., & Ban, X. (2022). Generating robust real-time object detector with uncertainty via virtual adversarial training. *International Journal of Machine Learning and Cybernetics, 13*, 431–445. https://doi.org/10.1007/s13042-021-01416-3

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference*. https://doi.org/10.1007/978-3-319-46723-8_49

da Silva, I. F. S., Silva, A. C., de Paiva, A. C., & Gattass, M. J. E. S. (2022). A cascade approach for automatic segmentation of cardiac structures in short-axis cine-MR images using deep neural networks. *Expert Systems with Applications, 197*, Article 116704. https://doi.org/10.1016/j.eswa.2022.116704

Dangi, S., Linte, C. A., & Yaniv, Z. (2019). A distance map regularized CNN for cardiac cine MR image segmentation. *Med Phys, 46*, 5637–5651. https://doi.org/10.1002/mp.13853

Dice, L. R. J. E. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297–302. https://doi.org/10.2307/1932409

Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference. Cham. https://doi.org/10.1007/978-3-030-87199-4_6.

Gerig, G., Jomier, M., & Chakos, M. (2001). Valmet: A new validation tool for assessing and improving 3D object segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4. https://doi.org/10.1007/3-540-45468-3_62.

Goodfellow, I. J., Shlens, J., & Szegedy, C. J. a. e.-p. (2014). Explaining and Harnessing Adversarial Examples, [preprint], arXiv:1412.6572. https://doi.org /10.48550/arXiv.1412.6572.

Gou, J., Yu, B., Maybank, S. J., Tao, D. J. I. J., & o. C. V.. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*, 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Han, Z., Jian, M., & Wang, G.-G. (2022). ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowledge-Based Systems, 253*, Article 109512. https://doi.org/10.1016/j.knosys.2022.109512

Heimann, T., van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., … Wolf, I. (2009). Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imaging, 28*, 1251–1265. https://doi.org/10.1109/TMI.2009.2013851

Huang, H., Chen, Z., Chen, C., Lu, M. & Zou, Y. (2023). Complementary consistency semi-supervised learning for 3D left atrial image segmentation. Computers in Biology and Medicine, Volume 165, October 2023, 107368.

Huang, N. (2018). https://www.dropbox.com/s/yyvj4352dax0q26/description_Ning_Huang.pdf?dl=0.

Jia, S., Despinasse, A., Wang, Z., Delingette, H., Pennec, X., Jaïs, P., Cochet, H., & Sermesant, M. (2019). Automatically segmenting the left atrium from cardiac images using successive 3D U-nets and a contour loss. international workshop on statistical atlases and computational models of the heart. Cham. https://doi.org/10.1007/3-540-45468-3_62.

Jiao, R., Zhang, Y., Ding, L., Cai, R., & Zhang, J. J. a. e.-p. (2022). Learning with limited annotations: A Survey on deep semi-supervised learning for medical image segmentation, [preprint], arXiv:2207.14191. https://doi.org /10.48550/arXiv.2207.14191.

Kendall, A., & Gal, Y. J. A. i. n. i. p. s. (2017). What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems, 30.

Kim, J., Min, Y., Kim, D., Lee, G., Seo, J., Ryoo, K., & Kim, S. (2022). ConMatch: Semi-supervised learning with confidence-guided consistency regularization. *European Conference on Computer Vision. Cham.*. https://doi.org/10.1007/978-3-031-20056-4_39

Lee, D., Kim, S., Kim, I., Cheon, Y., Cho, M., & Han, W.-S. (2022). Contrastive regularization for semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPRW56347.2022.00436

Li, S., Zhang, C., & He, X. (2020). Shape-aware semi-supervised 3D semantic segmentation for medical images. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. https://doi.org/10.1007/978-3-030-59710-8_54.

Li, Y., Luo, L., Lin, H., Chen, H., & Heng, P.-A. (2021). Dual-consistency semi-supervised learning with uncertainty quantification for COVID-19 lesion segmentation from CT images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*. https://doi.org/10.1007/978-3-030-87196-3_19

Li, L., Zimmer, V. A., Schnabel, J. A., & Zhuang, X. (2021). AtrialGeneral: Domain generalization for left atrial segmentation of multi-center LGE MRIs. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*. https://doi.org/10.1007/978-3-030-87231-1_54

Li, L., Zimmer, V. A., Schnabel, J. A., & Zhuang, X. (2022a). AtrialJSQnet: A New framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information. *Medical Image Analysis, 76*, Article 102303. https://doi.org/10.1016/j.media.2021.102303

Li, L., Zimmer, V. A., Schnabel, J. A., & Zhuang, X. (2022b). Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: A review. *Medical Image Analysis, 77*, Article 102360. https://doi.org/10.1016/j.media.2022.102360

Lippi, G., Sanchis-Gomar, F., & Cervellin, G. (2021). Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *International Journal of Stroke, 16*, 217–221. https://doi.org/10.1177/1747493019897870

Liu, Y., Cheng, H.-D., Huang, J., Zhang, Y., & Tang, X. J. J. (2012). An effective approach of lesion segmentation within the breast ultrasound image based on the cellular automata principle. *Journal of Digital Imaging, 25*, 580–590. https://doi.org/10.1007/s10278-011-9450-6

Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., & Carneiro, G. (2022). Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR52688.2022.00422

Liu, Y., Wang, W., Luo, G., Wang, K., Li, S. J. C. M. I., & Graphics.. (2022). A contrastive consistency semi-supervised left atrium segmentation model. *Computerized Medical Imaging and Graphics, 99*, Article 102092. https://doi.org/10.1016/j.compmedimag.2022.102092

Liu, P., & Zheng, G. (2022). Handling imbalanced data: uncertainty-guided virtual adversarial training with batch nuclear-norm optimization for semi-supervised medical image classification. *IEEE Journal of Biomedical and Health Informatics, 26*, 2983–2994. https://doi.org/10.1109/JBHI.2022. 3162748

Luo, X., Chen, J., Song, T., & Wang, G. (2021). Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*. https://doi.org/10.1609/aaai.v35i10.17066

Luo, X., Hu, M., Song, T., Wang, G., & Zhang, S. (2022). Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *International Conference on Medical Imaging with Deep Learning*.

McGann, C., Akoum, N., Patel, A., Kholmovski, E., Revelo, P., Damal, K., … Electrophysiology.. (2014). Atrial fibrillation ablation outcome is predicted by left atrial remodeling on MRI. *Circulation: Arrhythmia and Electrophysiology, 7*, 23–30.

Mi, P., Lin, J., Zhou, Y., Shen, Y., Luo, G., Sun, X., Cao, L., Fu, R., Xu, Q., & Ji, R. (2022). Active teacher for semi-supervised object detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org /10.1109/CVPR52688.2022.01408.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016 of Conference). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV). https://doi.org/10.1109/3dv.2016.79.

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 41, 1979-1993. https://doi.org/10.1109/TPAMI.2018.2858821.

Niyas, S., Pawan, S. J., Anand Kumar, M., & Rajan, J. (2022). Medical image segmentation with 3D convolutional neural networks: A survey. *Neurocomputing, 493*, 397–413. https://doi.org/10.1016/j.neucom.2022.04.065

Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. In *2019 IEEE/CVF conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2019.00025

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. https://doi.org/10.1007 /978-3-319-24574-4_28.

Shotton, J., Johnson, M., & Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *2008 IEEE conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2008.4587503

Su, W., & Wang, Z. (2019). Widening residual refine edge reserved neural network for semantic segmentation. *Multimedia Tools and Applications, 78*, 18229–18247. https://doi.org/10.1007/s11042-018-7121-z

Sukumar, N., & Srivastava, A. (2022). Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks. Computer Methods in Applied Mechanics and Engineering, 389, 114333. https://doi.org /10.1016/j.cma.2021.114333.

Tarvainen, A., & Valpola, H. J. A. i. n. i. p. s. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30.

Tong, J.-J., Zhang, P., Weng, Y.-X., & Zhu, D.-H. (2018). Kernel sparse representation for MRI image analysis in automatic brain tumor segmentation. *Frontiers of Information Technology & Electronic Engineering, 19*, 471–480. https://doi.org/10.1631/FITEE.1620342

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., & He, Z. (2020). Double-uncertainty weighted method for semi-supervised learning. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. https://doi.org/10.1007/978-3-030-59710-8_53.

Wang, J., Liu, X., Yin, J., & Ding, P. (2022). DC-net: Dual-Consistency semi-supervised learning for 3D left atrium segmentation from MRI. *Biomedical Signal Processing and Control, 78*. https://doi.org/10.1016/j.bspc.2022.103870

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

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., & Cai, J. (2022). Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis, 81*, Article 102530. https://doi.org/10.1016/j.media.2022.102530

Wu, Y., Xu, M., Ge, Z., Cai, J., & Zhang, L. (2021). Semi-supervised left atrium segmentation with mutual consistency training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*. https://doi.org/10.1007/978-3-030-87196-3_28

Xia, Q., Yao, Y., Hu, Z., & Hao, A. (2019). Automatic 3D atrial segmentation from GE-MRIs using volumetric fully convolutional networks. International Workshop on Statistical Atlases and Computational Models of the Heart. Cham. https://doi.org/10.1007/978-3-030-12029-0_23.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., … Zhao, J. (2021). A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis, 67*, Article 101832. https://doi.org/10.1016/j.media.2020.101832

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering, 35*, 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219

Yu, L., Wang, S., Li, X., Fu, C.-W., & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference*. https://doi.org/10.1007/978-3-030-32245-8_67

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. J. N. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage, 31*, 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

Zeng, Z., Wang, X., Zhang, J., & Wu, Q. (2016). Semi-supervised feature selection based on local discriminative information. *Neurocomputing, 173*, 102–109. https://doi.org/10.1016/j.neucom.2015.05.119

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., & Ooi, B. C. (2022). Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR52688.2022.02001

Zhao, C., Xiang, S., Wang, Y., Cai, Z., Shen, J., & Zhou, S. (2022). Context-aware network fusing transformer and V-Net for semi-supervised segmentation of 3D left atrium. *Expert Systems with Applications, 214*, Article 119105. https://doi.org/10.1016/j.eswa.2022.119105