Medical Image Analysis 79 (2022) 102447

Contents lists available at ScienceDirect

Medical Image Analysis

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

Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning

Kaiping Wang[a,1], Bo Zhan[a,1], Chen Zu[b], Xi Wu[c], Jiliu Zhou[a,c], Luping Zhou[d], Yan Wang[a,*]

[a] School of Computer Science, Sichuan University, Chengdu, China
[b] Department of Risk Controlling Research, JD.COM, China
[c] School of Computer Science, Chengdu University of Information Technology, China
[d] School of Electrical and Information Engineering, University of Sydney, Australia

ARTICLE INFO

Article history:
Received 9 September 2021
Revised 16 March 2022
Accepted 1 April 2022
Available online 8 April 2022

Keywords:
Semi-supervised segmentation
Mean teacher
Multi-task learning
Tripled-uncertainty
Contrastive learning

ABSTRACT

Due to the difficulty in accessing a large amount of labeled data, semi-supervised learning is becoming an attractive solution in medical image segmentation. To make use of unlabeled data, current popular semi-supervised methods (e.g., temporal ensembling, mean teacher) mainly impose data-level and model-level consistency on unlabeled data. In this paper, we argue that in addition to these strategies, we could further utilize auxiliary tasks and consider task-level consistency to better excavate effective representations from unlabeled data for segmentation. Specifically, we introduce two auxiliary tasks, i.e., a foreground and background reconstruction task for capturing semantic information and a signed distance field (SDF) prediction task for imposing shape constraint, and explore the mutual promotion effect between the two auxiliary and the segmentation tasks based on mean teacher architecture. Moreover, to handle the potential bias of the teacher model caused by annotation scarcity, we develop a tripled-uncertainty guided framework to encourage the three tasks in the student model to learn more reliable knowledge from the teacher. When calculating uncertainty, we propose an uncertainty weighted integration (UWI) strategy for yielding the segmentation predictions of the teacher. In addition, following the advance of unsupervised learning in leveraging the unlabeled data, we also incorporate a contrastive learning based constraint to help the encoders extract more distinct representations to promote the medical image segmentation performance. Extensive experiments on the public 2017 ACDC dataset and the PROMISE12 dataset have demonstrated the effectiveness of our method.

1. Introduction

Segmentation is a basic yet essential task in the realm of medical image processing and analysis. Conventional manual delineation of regions of interest (ROI) comes at the cost of much time. In addition, the quality of manual segmentation heavily relies on the clinical experiences of physicians. To enable clinical efficiency and obtain reliable segmentation results, many researchers are dedicated to studying automatic segmentation techniques. Benefiting from the advancement of deep learning, convolutional neural network (CNN) and its different variants have shown powerful image processing capabilities and have been widely used in automatic segmentation. Particularly, the proposals of Fully Convolutional Networks (FCN) (Long et al., 2015) and U-net (Ronneberger et al., 2015) have achieved a quantum leap and laid a solid foundation in the field of automatic medical image segmentation. Based on them, numerous works have been developed to further improve the segmentation performance. For example, considering the importance of boundaries in segmentation, Fang et al. (2019) and Zhang et al. (2019) explicitly encouraged the model to discriminate the boundaries of the target, thus refining the segmentation performance. Cheng et al., 2020 introduced the direction field to implicitly restrict the shape of segmentation result. Tang et al. (2022) proposed a unified end-to-end medical image segmentation framework combining uncertainty estimation and multi-scale feature extraction. The above methods have obtained excellent performance and generalization. However, benefits and costs always come together. Their success heavily depends on a large amount of annotated data which are hard to be available in the real world, especially in the field of medical image segmentation due to their expensive and time-consuming nature. To alleviate annotation scarcity, a feasible approach is to take semi-supervised learning (Kervadec et al., 2019; Bortsova et al., 2019)

* Corresponding author.
E-mail address: wangyanscu@hotmail.com (Y. Wang).
[1] The first two authors contribute equally to this work.

https://doi.org/10.1016/j.media.2022.102447
1361-8415/

which leverages both labeled and unlabeled data to effectively train the deep network.

Considerable efforts have been devoted to the semi-supervised segmentation community, which can be broadly categorized into two popular groups. The first group refers to those methods trying to predict pseudo labels on unlabeled images and mixing them with ground truth labels to provide additional training information (Zheng et al., 2020a; Park et al., 2018; Zheng et al., 2020b). However, the segmentation results of such self-training based method are susceptible due to the uneven quality of the predicted pseudo labels. The second group of semi-supervised segmentation methods lies in the consistency regularization, that is, encouraging the segmentation predictions to be consistent under different perturbations at data level and model level for the same input. Specifically, the data-level consistency requires the agreement between the results obtained by applying different disturbances on the same input, while the model-level consistency needs to ensure the consensus of the results through different models regarding the same input. A typical data-level example is the Π-model (Laine and Aila, 2016) which minimizes the distance between the results of two forward passes with different regularization strategies. Given that each result is based on a single evaluation of the network, it could be noisy and unstable during the training stage. To improve the stability, a temporal ensembling model (Laine and Aila, 2016) is further proposed based on the Π-model, which aggregates the exponential moving average (EMA) predictions and encourages the consensus between the ensembled predictions and current predictions for unlabeled data. To accelerate the training and enable the online learning, mean teacher (Tarvainen and Valpola, 2017) further improves the temporal ensembling model by enforcing prediction consistency between the current training model (i.e., the student model) and the corresponding EMA model (i.e., the teacher model) with different noise perturbations added at the input, taking both data-level and model-level consistency into consideration. Nevertheless, unreliable results from the teacher model may mislead the student model, thus deteriorating the whole training. Researchers therefore incorporated the uncertainty map into the mean teacher model, forcing the student to learn high confidence predictions from the teacher model (Yu et al., 2019; Wang et al., 2020; Hu et al., 2021). Our research falls in the second group of semi-supervised approaches.

On the other hand, different from the above semi-supervised segmentation methods which mainly focus on the consistency under the disturbances at data level or model level, there are also research works (Luo et al., 2020; Chen et al., 2019; Li et al., 2020; Jia et al., 2021) that explore to improve segmentation from another perspective: multi-task learning. These methods jointly train multiple related tasks in a single model through a shared encoder. For instance, Chen et al. (2019) applied a reconstruction task to assist the segmentation of medical images. Li et al. (2020) proposed a shape-aware semi-supervised model by incorporating a signed distance map generation task to enforce a shape constraint on the segmentation result. By exploring the task-level consistency between the main and the auxiliary tasks, the learned segmentation model could bypass the overfitting problem and learn more representative features from unlabeled data to boost the segmentation performance.

In addition to the semi-supervised learning, self-supervised learning is also an alternative to solve the label scarcity problem, which aims to pretrain a model on unlabeled data to learn effective feature representations that are beneficial for the downstream tasks (Chen et al., 2020a). As a significant self-supervised learning method, contrastive learning has recently shown promising results on several computer vision tasks, such as image classification (Chen et al., 2020a; Hassani and Khasahmadi, 2020; Shi et al., 2022), image segmentation (Pandey et al., 2021, Sun et al., 2022), image translation (Park et al., 2020). Generally speaking, contrastive learning requires to couple the feature of the current sample, i.e., the query, with the features of its variant as well as other samples, i.e., the keys, thereby constructing the positive pair and the negative pairs (Xiao et al., 2020). Then, a model is trained to correctly distinguish the positive pair from large amounts of negative ones. By this way, the pretrained model is able to yield effective visual representations, which could further benefit the learning of downstream tasks.

In this paper, inspired by the success of semi-supervised learning, contrastive learning as well as multi-task learning, we propose a novel end-to-end semi-supervised mean teacher model guided by tripled-uncertainty maps from three highly related tasks. Concretely, considering that the performance of semi-supervised learning is highly dependent on the extraction of effective feature representations from unlabeled data (Rebuffi et al., 2020), we utilize the multi-task learning and the contrastive learning to strengthen the representation ability of the encoder and transfer the learned representations to the medical segmentation task. For multi-task learning, apart from the segmentation task, we bring in two additional auxiliary tasks, i.e., a foreground and background reconstruction task and a signed distance field (SDF) prediction task. The reconstruction task can help the segmentation network capture more semantic information, while the predicted SDF describes the signed distance of a corresponding pixel to its closest boundary point after normalization, thereby constraining the global geometric shape of the segmentation result. Following the spirit of mean teacher architecture, we build a teacher model and a student model, each targeting at all three tasks above. Additionally, to tackle the unreliability and noise in teacher model predictions, we impose uncertainty constraints on all the three tasks, expecting the student model can learn as much accurate information as possible. For contrastive learning, we follow the unsupervised contrastive learning mechanism and impose consistency constraint on the features from encoders straightforwardly, thus promoting the encoders to learn sample specific characteristics.

Our main contributions are four-fold: (1) We inject the spirit of multi-task learning into mean teacher architecture, so that the segmentation task could also benefit from the enhanced semantic and geometric shape information by mining the correlations among the segmentation task, the reconstruction task, and the SDF prediction task. In this manner, our mean teacher model simultaneously takes account of the data-, model- and task-level consistency to better leverage unlabeled data for segmentation. (2) We impose the uncertainty estimation on all tasks and develop a tripled-uncertainty to guide the student model to learn more reliable predictions from the teacher model. (3) Current approaches tend to generate uncertainty maps by averaging the results from multiple Monte Carlo (MC) samplings, neglecting the discrepancy of different results. In contrast, we propose an uncertainty weighted integration (UWI) strategy to assign different weights for different sampling results, generating a more accurate segmentation prediction. (4) Pure semi-supervised learning mainly focuses on the regularization of one sample for each forward. In contrast, contrastive learning involves the relationship modeling of multiple samples, excavating a more representative data distribution. Therefore, we integrate the contrastive constraint into our semi-supervised model, assisting the encoders to learn more distinct representations. The experimental results obtained on both the public 2017 ACDC dataset (Bernard et al., 2018) and PROMISE12 dataset (Litjens et al., 2014) demonstrate the advancement of our method and the effectiveness of its critical modules.

Please note that the preliminary version of this work was presented earlier at 2021 Medical Image Computing and Computer-Assisted Intervention (MICCAI) (Wang et al., 2021). This paper extends the conference version from the following aspects: (1) In-

troduction: we refined the research background to make the motivation of the research more rigorous and compelling; (2) Related work: we added a related work section to give a thorough review of relevant researches; (3) Methodology: we further imposed the contrastive constraint on the encoder to enhance its feature representation ability by keeping the features from the same image close to each other while the features from different images away from each other. In addition, we presented a pseudo code in Algorithm 1 to clearly show the training procedure of our model; (4) Discussion: we not only analyzed and discussed the results of the comparison and ablation experiments, but also summarized the limitations of our method and indicated the future research directions.

The rest of this paper is organized as follows. We briefly review the relevant works in Section 2. The architecture and key algorithms of our method are detailed in Section 3. The datasets and experimental results are presented in Section 4. In Section 5, we discuss the results, the limitations and the future improvements of our current method. In the end, the whole paper is concluded in Section 6.

2. Related work

2.1. Multi-task learning

Multi-task learning is a widely used strategy which aims to jointly learn multiple related tasks, so that the knowledge in a task can be leveraged by others, leading to better generalizability (Zhang and Yang, 2021). In deep neural networks, multi-task learning is often implemented by a parameter-shared architecture and several individual heads for different tasks. Generally, there are two forms of parameter sharing strategies, i.e., soft parameter sharing and hard parameter sharing (Ruder, 2017).

The soft parameter sharing methods tend to design individual architectures for different tasks, and impose the constraint on corresponding parameters of different architectures to encourage them to be similar. Yang and Hospedales (2016) proposed to regularize the similarity of parameters of different networks by trace norm. Duong et al. (2015) presented two language parser networks and utilized the L2 norm to keep the distance between the parameters and weights of two networks to stay close. Similarly, Guo et al. (2018) introduced a multi-level soft sharing strategy with parallel training to complete the sentence simplification task. They believed that part of parameters is layer-specific, which will be kept private during training. Conversely, the related parameters of different networks are regularized by L2 norm. Despite the progress in performance, the soft parameter sharing strategy could suffer from a larger increase in the number of parameters when the involved tasks expand (Sun et al., 2020).

In contrast, the hard parameter sharing strategy is a more popular approach, which can easily integrate several tasks into a single network architecture by sharing the encoder parameters. Without building multiple separate networks, the hard parameter sharing strategy ties the parameters at lower layers of all the tasks together, thus suppressing the growth of parameters number and greatly mitigating the risk of overfitting compared to the soft parameter sharing. A famous practice is the two-stage object detection (Girshick et al., 2014; Girshick, 2015), in which two task heads are equipped based on a shared encoder to predict the bounding box position and the class probability, respectively. In addition, the segmentation task has also benefited from this hard parameter sharing strategy and achieved decent performance (Duan et al., 2019; Zhang et al., 2021a; Amyar et al., 2020; Bischke et al., 2019; Murugesan et al., 2019; Song et al., 2020). For instance, Duan et al. (2019) proposed a multi-task deep neural network which simultaneously predicts the segmentation and anatomical landmarks for cardiac magnetic resonance volumes. Zhang et al. (2021a) integrated the cancer segmentation task and the classification task for automatic gastric tumor segmentation. Liu et al. (2019) explicitly explored the relationship between the lung nodule benign-malignant classification task and the attribute regression task. Due to the simplicity and effectiveness of the hard parameter sharing strategy, we also choose it to build our multi-task architecture.

2.2. Contrastive learning

Recently, contrastive learning has gained widespread attention in the computer vision community. One critical issue to conduct contrastive learning is how to effectively construct the positive pair and negative pairs. A generic and feasible solution is to apply different augmentation transformations to the same input sample, thus resulting in different views. Subsequently, the views from the same input sample are tied as a positive pair and those from other samples form negative pairs. After that, a deep model is trained to discriminate the positive and negative pairs, thus learning the intrinsic features hidden in various samples. Following this, Hjelm et al. (2018) selected both the local features and the global representations from a mini-batch to derive the positive and negative pairs. However, recent study indicated that the performance could be influenced by the number of negative pairs in the training batch (Kalantidis et al., 2020). To this end, Wu et al. (2018) proposed to utilize a memory bank to store the previous features so that the quantity of negative pairs can be enlarged. More extremely, SimCLR (Chen et al., 2020a) directly took a large enough batch size to hold more negative samples and introduced a non-linear projection to better learn the representations. Yet, the memory bank tends to preserve the early features of the network which have little contribution to the contrastive learning during the network evolution. Also, a larger batch means more computational resources need consuming. To alleviate this problem, He et al. (2020) proposed a Momentum Contrast (MOCO) model which maintains a smaller queue to store previous features and remove the out-of-date ones, ensuring the consistency of the stored features and reducing the computational consumption. Moreover, Khosla et al. (2020) brought the prior supervision signal into the contrastive learning and regularized the distance of different classes rather than the randomly selected instances, avoiding the possible false negative pairs.

2.3. Semi-supervised segmentation

Semi-supervised segmentation aims to leverage the large proportion of unlabeled data to provide extra information for the training of limited labeled data, thus extracting more generalized representations for segmentation and alleviating the possible overfitting problem.

A typical semi-supervised segmentation strategy can be categorized as the pseudo labelling which exploits a pretrained model to generate pseudo labels for unlabeled data to expand the labeled dataset. For instance, Bai et al. (2017) adopted an alternate update manner to optimize network parameters and pseudo labels with a conditional random field to refine the segmentation results. Peng et al. (2020) proposed a co-training method which trains two models simultaneously and encourages their predictions on unlabeled data to provide pseudo-supervised signals to each other. Chen et al. (2020b) believed that training a model with both strong label and pseudo label spaces could lead to disordered backpropagation. As a result, they designed an approach with two branches

for strong labeled and pseudo labeled data, and used the reliable information extracted from the pseudo labels to assist the strong label learning. However, for these methods, the quality of pseudo labels is difficult to be guaranteed, which may introduce much noise to the model and cause inferior performance.

Generative adversarial network (GAN) not only shines in the field of medical image synthesis (Wang et al., 2018a; Wang et al., 2018b; Zhan et al., 2021; Luo et al., 2021; Li et al., 2022), but is also popular in the field of semi-supervised segmentation. For instance, Zhang et al. (2017) introduced a discriminator to distinguish whether the segmentation prediction is from labeled or unlabeled data, thus narrowing the distribution gap between the labeled and unlabeled predictions. Nie et al. (2018) proposed an adversarial confidence network to assist the segmentation model to learn the knowledge from unlabeled data. Li et al. (2020) adopted the adversarial learning to encourage the signed distance maps from the labeled and unlabeled predictions to be close, so that the shape information of unlabeled data can be fully utilized.

More recently, the consistency based semi-supervised methods have been favored by many researchers due to their powerful performance (Kervadec et al., 2019; Luo et al., 2020; Chen et al., 2019). Generally speaking, these methods attempt to devise suitable auxiliary tasks and impose the consistency regularization on the outputs of the main and auxiliary tasks. For example, Kervadec et al. (2019) incorporated a simpler image-level feature regression task, such as predicting the size of target, and encouraged its output to be close to that of the main segmentation task. The consistency constraint between two tasks helps to supplement the segmentation with more useful information from unlabeled data. Similarly, Luo et al. (2020) explored the unlabeled data by ensuring the consistency between the segmentation output and the signed distance map prediction. In addition, the model-level and data-level consistency has also been widely concerned. (Laine and Aila, 2016) proposed a Π-model to encourage the data-level consistency between the predictions under different perturbations. Furthermore, to improve the training stability, they presented a temporal ensembling model based on the Π-model, which integrates the previous outputs as the target of regularization. Inspired by this, (Tarvainen and Valpola, 2017) devised a mean teacher model which considers to ensemble the exponential moving average (EMA) of previous parameters rather than outputs, and ensured the model-level consistency between the student model and the ensembled teacher model. To alleviate the noise from the teacher model, uncertainty estimation (Zheng et al., 2020a; Wang et al., 2020; Hu et al., 2021; Sedai et al., 2019) is also widely applied to enforce the student model to learn more reliable knowledge from the teacher model. For example, Hu et al., (2021) utilized the uncertainty estimation to weight the predictions of mean teacher and ensured a reliable semi-supervised nasopharyngeal carcinoma segmentation.

In addition to the above approaches, there are also some other works handling the unlabeled data from the perspectives of entropy minimization (Kalluri et al., 2019; Hang et al., 2020), mutual information regularization (Peng et al., 2020; Xi, 2019) and graph construction (Huang et al., 2021; Zhang et al., 2021b). For instance, Hang et al. (2020) introduced the principle of entropy minimization into the student model to produce high-confident predictions of unlabeled data. Peng et al. (2020) regularized the mutual information on the predictions of the original and transformed unlabeled images to boost the segmentation performance. Huang et al. (2021) employed the bilateral graph convolution to capture the long-range dependencies and refine the visual representations in the semantic segmentation task and the boundary detection task.

In this work, we put emphasis on the continuous exploration of the mean teacher architecture.

3. Methodology

An overview of our proposed network is illustrated in Fig. 1, consisting of a teacher model and a student model, following the idea of mean teacher. The student model is the target model to be trained, and it assigns the exponential moving average (EMA) of its weights to the teacher model at each step of training. On the other hand, the predictions of the teacher model would be viewed as additional supervisions for the student model to learn. These two models adopt a similar encoder-decoder structure, where the encoder is shared among different tasks while the decoders are task-specific. In our problem setting, we are given a training set containing N labeled data and M unlabeled data, where $N \ll M$. The labeled set is defined as $\mathcal{D}^l = \{\mathbf{X}^i, \mathbf{Y}^i\}_{i=1}^{N}$ and the unlabeled set as $\mathcal{D}^u = \{\mathbf{X}^i\}_{i=N+1}^{N+M}$, where $\mathbf{X}^i \in \mathbb{R}^{H \times W}$ is the intensity image, $\mathbf{Y}^i \in \{0, 1\}^{H \times W}$ is the corresponding segmentation label. For the labeled data, we can also obtain the ground truth of signed distance field (SDF) $\mathbf{Z}^i \in \mathbb{R}^{H \times W}$ from $\mathbf{Y}^i$ via the SDF function in (Xue et al., 2020) for the SDF prediction task. Similarly, for the reconstruction task, the ground truths of foreground and background $\mathbf{G}^i \in \mathbb{R}^{2 \times H \times W}$ can be obtained by $\mathbf{Y}^i \odot \mathbf{X}^i$ and $(1 - \mathbf{Y}^i) \odot \mathbf{X}^i$ where $\odot$ refers to element-wise multiplication. Our goal is to leverage the contrastive learning to strengthen the representative ability of encoders, through which semantic information and geometric shape knowledge can be fully learned from the two auxiliary tasks, i.e., the SDF prediction task and the foreground and background reconstruction task, to enhance the segmentation performance on both $\mathcal{D}^l$ and $\mathcal{D}^u$. In summary, given $\mathcal{D}^l$ and $\mathcal{D}^u$, the student model is optimized by minimizing 1) the supervised segmentation loss $\mathcal{L}_S$ on labeled data $\mathcal{D}^l$, 2) the inter-model consistency loss $\mathcal{L}_{\text{CSL}}^{\text{model}}$ between the student model and teacher model on both $\mathcal{D}^l$ and $\mathcal{D}^u$, 3) the inter-task consistency loss $\mathcal{L}_{\text{CSL}}^{\text{task}}$ among different tasks on $\mathcal{D}^l$ and $\mathcal{D}^u$, and 4) the contrastive loss $\mathcal{L}_{\text{CTL}}^{\text{model}}$ between the student model and teacher model on both $\mathcal{D}^l$ and $\mathcal{D}^u$. In addition, the tripled-uncertainty maps with respect to the three tasks generated by the teacher model guide the student model to learn more reliable predictions from the teacher model. Moreover, to enhance the robustness of our mean teacher model, different perturbations ξ and ξ' are fed into the student and teacher model, respectively. More details are introduced in the subsequent sections.

3.1. Student model

U-net (Ronneberger et al., 2015) is a significant encoder-decoder style framework which was originally proposed to be used in the field of medical segmentation analysis. Due to its high performance and excellent generalization, our student model employs U-net as the backbone of three tasks, i.e., segmentation task, reconstruction task and SDF prediction task. Note that the encoder of the student model is shared by the three tasks with the same parameters while the parameters in the task-specific decoders are different to fit the different tasks. In this manner, the encoder is forced to capture features related to the semantic information and geometric shape information with the optimization of three highly relevant tasks, leading to a low disparity between the output segmentation result and the ground truth. In addition, considering that the reconstruction task aims to learn effective high-level semantic representations, it needs to condense the spatial features of the input. However, the skip connections adopted in U-net allow the network to directly copy features from early layers without dimensionality reduction, which could lead this task invalid. Therefore, following (Chen et al., 2019), we drop the skip connection in the reconstruction task.

Concretely, the encoder consists of four convolutional blocks with different feature scales. In the first three blocks, each one

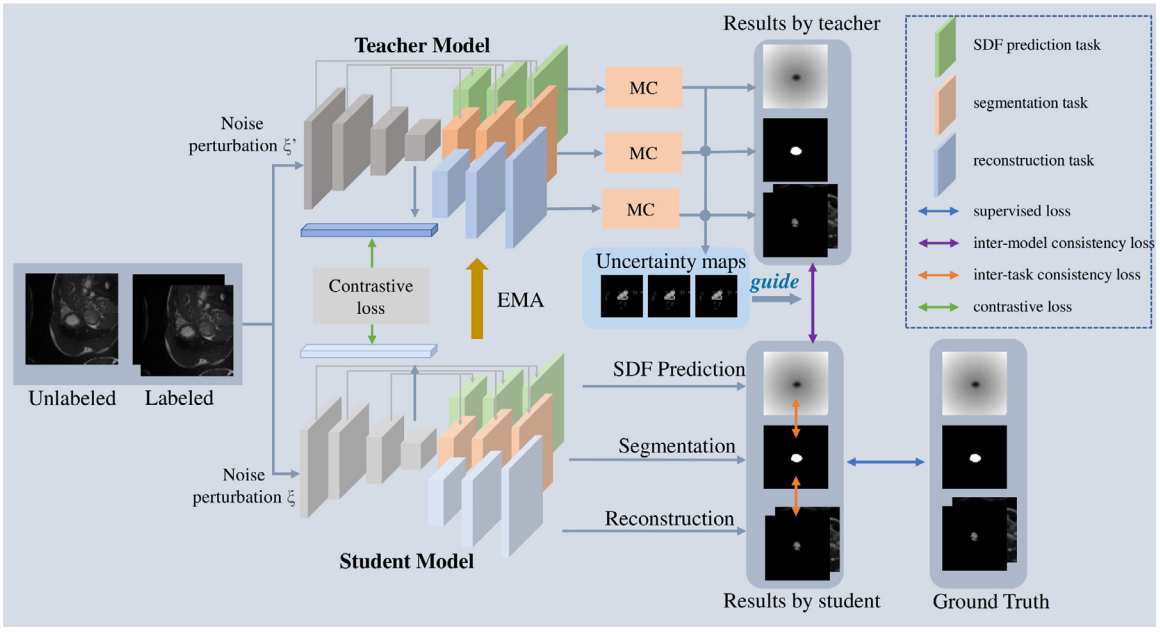

Fig. 1. Overview of our tripled-uncertainty guided mean teacher model.

consists of two convolutional layers with kernel size of 3 × 3, followed by a batch normalization layer and a ReLU activation layer. The stride of the second convolutional layer is set as 2 to down-sample the feature maps, for acquiring condensed spatial information. For the fourth block, two 3 × 3 convolutional layers with stride of 1 are employed to obtain the feature representations of the encoder. Concerning the three task-specific decoders, each of them contains three up-sampling blocks and a task-specific head. The structures of up-sampling blocks are symmetrical to those of the first three convolutional blocks in the encoder. The segmentation head is equipped with softmax activation while the reconstruction head and the SDF head utilize sigmoid activation.

Given an input image $\mathbf{X}^i$, the three tasks can generate the segmentation result $\tilde{\mathbf{Y}}_S^i$, the reconstruction result $\tilde{\mathbf{G}}_S^i$, and the SDF result $\tilde{\mathbf{Z}}_S^i$, as follows:

$$\tilde{\mathbf{Y}}_S^i = f_{seg}\left(\mathbf{X}^i; \theta_{seg}, \xi\right),$$

$$\tilde{\mathbf{G}}_S^i = f_{rec}\left(\mathbf{X}^i; \theta_{rec}, \xi\right),$$

$$\tilde{\mathbf{Z}}_S^i = f_{sdf}\left(\mathbf{X}^i; \theta_{sdf}, \xi\right), \tag{1}$$

where f_{seg}, f_{rec}, f_{sdf} represent the segmentation network, the reconstruction network and the SDF prediction network with corresponding parameters θ_{seg}, θ_{rec}, θ_{sdf}, and ξ is the noise perturbation of the student model.

3.2. Teacher Model

The network of our teacher model is reproduced from the student model, yet they have different ways for updating parameters. The student model updates its parameters $\boldsymbol{\theta} = \{\theta_{seg}, \theta_{rec}, \theta_{sdf}\}$ by gradient descent while the teacher model updates its parameters $\boldsymbol{\theta'} = \{\theta'_{seg}, \theta'_{rec}, \theta'_{sdf}\}$ as the EMA of the student model parameters $\boldsymbol{\theta}$ in different training steps. In particular, at training step t, the parameters of the teacher model, i.e., $\boldsymbol{\theta_t}'$, are updated according to:

$$\boldsymbol{\theta}_{\boldsymbol{t}}^{'} = \tau\boldsymbol{\theta}_{\boldsymbol{t-1}}^{'} + (1-\tau)\boldsymbol{\theta}, \tag{2}$$

where τ is the coefficient of EMA decay to control the updating rate.

Moreover, as there is no label on $\mathcal{D}^u$, the results of the teacher model would be utilized as pseudo labels for student model. However, the possible undesirable biases in the results may bring noise and mislead the student model, thus causing inferior segmentation performance. To minimize such unreliability, we bring in the uncertainty estimation in the teacher model to guide the student to learn more reliable knowledge. Specifically, we perform K times forward passes with Monte Carlo (MC) dropout, thus obtaining K preliminary results of all the tasks with regard to the input $\mathbf{X}^i$, i.e.,$\{\tilde{\mathbf{Y}}_T^{ij}\}_{j=1}^K$, $\{\tilde{\mathbf{G}}_T^{ij}\}_{j=1}^K$, and $\{\tilde{\mathbf{Z}}_T^{ij}\}_{j=1}^K$. Subsequently, the results of teacher model and corresponding uncertainty maps can be obtained from these preliminary results.

Traditional uncertainty-based segmentation methods, like (Yu et al., 2019; Wang et al., 2020; Sedai et al., 2019), always generate the final integrated result by averaging the preliminary results simply, neglecting the specificity of them. Differently, for the main segmentation task, we innovatively design an uncertainty weighted integration (UWI) strategy to assign different weights for different sampling results. The process of the proposed UWI is illustrated in Fig. 2. First, the K preliminary segmentation results $\{\tilde{\mathbf{Y}}_T^{ij}\}_{j=1}^K$ represented by the green grids are received as input. Considering the diverse uncertainty among these results, we then calculate the uncertainty maps $\{U_{\text{seg}}^{ij} = -\sum_{c\in C}\hat{\mathbf{Y}}_T^{ijc}\log_C\hat{\mathbf{Y}}_T^{ijc}\}_{j=1}^K$ for each preliminary result, where C is the number of classes to be segmented and is set to 2 here, c is the c-th class. By doing so, the value range of uncertainty maps is between 0 and 1, and a larger value represents a higher degree of uncertainty. Since these uncertainty maps imply the uncertainty areas in the teacher model learning, we further construct K confidence maps by $\{C_{\text{seg}}^{ij} = 1 - U_{\text{seg}}^{ij}\}_{j=1}^K$, where each pixel corresponds to a vector with length of K. The values of the vector are then normalized to $[0, 1]$ by applying a softmax operation, which can be expressed as $e^{\{\{C_{\text{seg}}^{ij}\}_{j=1}^K\}}/\sum_{j=1}^K e^{\{C_{\text{seg}}^{ij}\}}$. Afterwards, we obtain K weight maps $\{\mathbf{W}_{\text{seg}}^{ij}\}_{j=1}^K$ corresponding to K preliminary results, where pixels with higher weights (lower uncertainty) would contribute more to the final results, thus guiding the teacher model to heed the areas

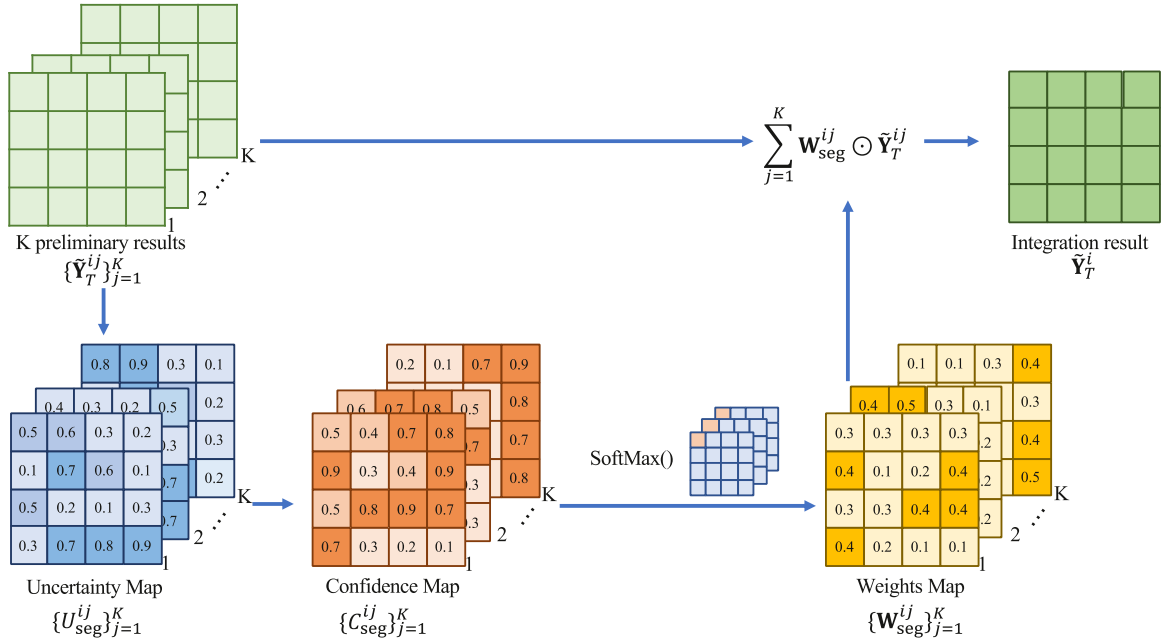

Fig. 2. The calculation flow of Uncertainty Weighted Integration (UWI).

across results with high confidence. Finally, the integrated segmentation prediction $\tilde{\mathbf{Y}}_T^i$ can be derived by $\tilde{\mathbf{Y}}_T^i = \sum_{j=1}^K \mathbf{W}_{\text{seg}}^{ij} \odot \tilde{\mathbf{Y}}_T^{ij}$.

As for the other two auxiliary tasks, it is noteworthy that they predict the real regression values rather than the probabilistic values as the segmentation task. Accordingly, the entropy is unsuitable for the uncertainty estimation for them. Therefore, we obtain the aggregated results directly by averaging operation, that is, $\tilde{\mathbf{G}}_T^i = \frac{1}{K}\sum_{j=1}^K \tilde{\mathbf{G}}_T^{ij}$, $\tilde{\mathbf{Z}}_T^i = \frac{1}{K}\sum_{j=1}^K \tilde{\mathbf{Z}}_T^{ij}$. For the same reason, we utilize the variance instead of entropy as the uncertainty of the aggregated results of these two auxiliary tasks by following (Kendall and Gal, 2017). To sum up, leveraging the aggregated results of three tasks, we can acquire tripled-uncertainty maps of all the tasks by:

$$
\begin{aligned}
U_{\text{seg}} &= -\sum_{c\in C} \tilde{\mathbf{Y}}_T^{ic} \log_C \tilde{\mathbf{Y}}_T^{ic}, \\
U_{\text{rec}} &= \frac{1}{K}\sum_{j=1}^K \left(\tilde{\mathbf{G}}_T^{ij} - \tilde{\mathbf{G}}_T^i\right)^2, \\
U_{\text{sdf}} &= \frac{1}{K}\sum_{j=1}^K \left(\tilde{\mathbf{Z}}_T^{ij} - \tilde{\mathbf{Z}}_T^i\right)^2.
\end{aligned} \tag{3}
$$

With the tripled-uncertainty guidance, the student model can alleviate the adverse effect of the misleading information and learn more trustworthy knowledge from the teacher model.

3.3. Objective functions

As aforementioned, the objective function is composed of four aspects: 1) Supervised loss $\mathcal{L}_{\text{S}}$ on labeled data $\mathcal{D}^l$; 2) Inter-model consistency loss $\mathcal{L}_{\text{CSL}}^{\text{model}}$ between the student model and the teacher model on both $\mathcal{D}^l$ and $\mathcal{D}^u$; 3) Inter-task consistency loss $\mathcal{L}_{\text{CSL}}^{\text{task}}$ among different tasks in the student model on $\mathcal{D}^l$ and $\mathcal{D}^u$; 4) Inter-model contrastive loss $\mathcal{L}_{\text{CTL}}^{\text{model}}$ between the output features produced by the student encoder and teacher encoder on $\mathcal{D}^l$ and $\mathcal{D}^u$.

Specifically, $\mathcal{L}_{\text{S}}$ is the weighted sum of the supervised losses on three tasks, i.e., $\mathcal{L}_{\text{S}}^{\text{seg}}$, $\mathcal{L}_{\text{S}}^{\text{rec}}$, $\mathcal{L}_{\text{S}}^{\text{sdf}}$, and can be formulated as:

$$
\mathcal{L}_{\text{S}} = \mathcal{L}_{\text{S}}^{\text{seg}} + \alpha_1 \mathcal{L}_{\text{S}}^{\text{rec}} + \alpha_2 \mathcal{L}_{\text{S}}^{\text{sdf}}, \tag{4}
$$

where $\mathcal{L}_{\text{S}}^{\text{seg}}$ uses Dice loss following (Milletari et al., 2016), $\mathcal{L}_{\text{S}}^{\text{rec}}$ and $\mathcal{L}_{\text{S}}^{\text{sdf}}$ use mean squared error (MSE) loss, α_1 and α_2 are coefficients for balancing the loss terms.

For the same input from $\mathcal{D}^l$ or $\mathcal{D}^u$, since the teacher model is an ensembling of the student model, the outputs of both models on three tasks should be identical. Therefore, we employ the inter-model consistency loss $\mathcal{L}_{\text{CSL}}^{\text{model}}$ to constrain this condition as follows:

$$
\begin{aligned}
\mathcal{L}_{\text{CSL}}^{\text{model}} &= \mathcal{L}_{\text{CSL}}^{\text{seg}} + \mu_1 \mathcal{L}_{\text{CSL}}^{\text{rec}} + \mu_2 \mathcal{L}_{\text{CSL}}^{\text{sdf}}, \\
\mathcal{L}_{\text{CSL}}^{\text{seg}} &= \frac{1}{N+M}\sum_{i=1}^{N+M} \exp\left(-U_{\text{seg}}\right) \odot \left(\tilde{\mathbf{Y}}_S^i - \tilde{\mathbf{Y}}_T^i\right)^2, \\
\mathcal{L}_{\text{CSL}}^{\text{rec}} &= \frac{1}{N+M}\sum_{i=1}^{N+M} \exp\left(-U_{\text{rec}}\right) \odot \left(\tilde{\mathbf{G}}_S^i - \tilde{\mathbf{G}}_T^i\right)^2, \\
\mathcal{L}_{\text{CSL}}^{\text{sdf}} &= \frac{1}{N+M}\sum_{i=1}^{N+M} \exp\left(-U_{\text{sdf}}\right) \odot \left(\tilde{\mathbf{Z}}_S^i - \tilde{\mathbf{Z}}_T^i\right)^2,
\end{aligned} \tag{5}
$$

where the tripled-uncertainty U_{seg}, U_{rec}, U_{sdf} are used as weight maps to encourage the student model to learn meaningful information from the teacher model, and μ_1, μ_2 are balancing coefficients.

Similarly, owing to the shared encoder, the results of three tasks are supposed to be consistent in semantic level for the same input. Based on this, we devise the inter-task consistency loss $\mathcal{L}_{\text{CSL}}^{\text{task}}$ to narrow the gap between $\tilde{\mathbf{Y}}_S^i$ and $\tilde{\mathbf{G}}_S^i$, $\tilde{\mathbf{Z}}_S^i$. Accordingly, $\mathcal{L}_{\text{CSL}}^{\text{task}}$ is formulated as:

$$
\mathcal{L}_{\text{CSL}}^{\text{task}} = \frac{1}{N+M}\sum_{i=1}^{N+M} \left(\left(\tilde{\mathbf{Z}}_S^i - SDF\left(\tilde{\mathbf{Y}}_S^i\right)\right)^2 + \left(\tilde{\mathbf{G}}_S^i - Mask\left(\tilde{\mathbf{Y}}_S^i, \mathbf{X}^i\right)\right)^2\right), \tag{6}
$$

where $SDF(\tilde{\mathbf{Y}}_S^i)$ converts $\tilde{\mathbf{Y}}_S^i$ to the domain of SDF following the function in (Xue et al., 2020), and $Mask(\tilde{\mathbf{Y}}_S^i, \mathbf{X}^i)$ is the concatenation of $\tilde{\mathbf{Y}}_S^i \odot \mathbf{X}^i$ and $(1 - \tilde{\mathbf{Y}}_S^i) \odot \mathbf{X}^i$.

In addition to the output-level constraints $\mathcal{L}_{\text{S}}$, $\mathcal{L}_{\text{CSL}}^{\text{model}}$ and $\mathcal{L}_{\text{CSL}}^{\text{task}}$, we also impose a contrastive learning based loss $\mathcal{L}_{\text{CTL}}^{\text{model}}$ on the

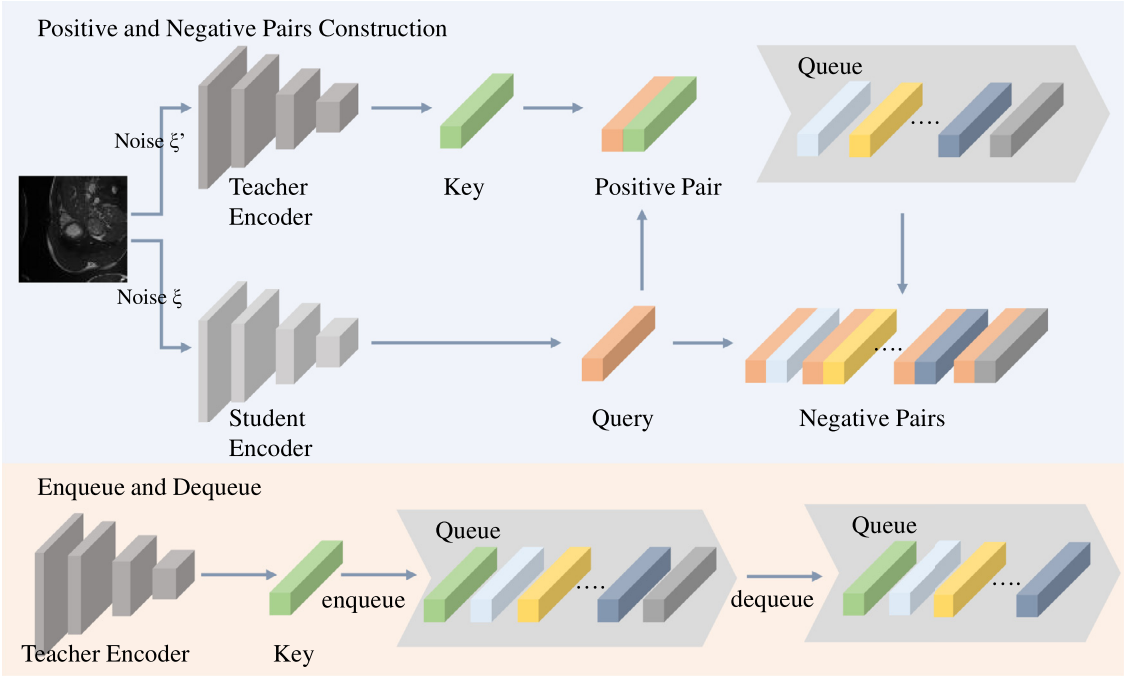

Fig. 3. The construction of positive and negative pairs for the contrastive loss.

output features of encoders to enhance their representation ability. The construction of the positive pair and the negative pairs is depicted in the upper part of Fig. 3. Concretely, we treat the feature representation of student encoder as a query and the feature representation of teacher encoder as a key. The query and key in the current iteration are formed into a positive pair. On the other hand, to build negative pairs, we also maintain a queue to store the keys produced by the teacher encoder from previous iterations. Then the query is coupled with each key in the queue to form multiple negative pairs. The process of enqueue and dequeue are presented in the lower part of Fig. 3. After the loss calculation in each iteration, the current key from the current teacher model will be pushed into the queue while the oldest key will be pushed out. Thus, the queue is updated in a progressive way and the consistency of the stored keys can be ensured. Note that all the feature representations are flattened for follow-up calculations.

Our ultimate goal is to enable our network to discriminate the positive pair and the negative pairs, that is, to group the feature representations in the positive pair closer and diverse the feature representations in the negative pairs far from each other. Herein, we harness the inner product to measure the similarity between the query and the key, and the contrastive loss can be formulated as follows:

$$\mathcal{L}_{\mathrm{CTL}}^{\mathrm{model}} = -\log \frac{\exp\left(f_s \cdot f_t^{+}\right)}{\sum_{f_t^{-}} \exp\left(f_s \cdot f_t^{-}\right)}, \tag{7}$$

where f_s represents the query feature from the student encoder, f_t^{+} indicates the positive key feature from the current teacher encoder, and f_t^{-} represents the negative key features from the queue. "·" is denoted as the inner product operation of vectors.

Finally, the total objective function $\mathcal{L}_{\mathrm{total}}$ can be summarized as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{S}} + \lambda_1 \mathcal{L}_{\mathrm{CSL}}^{\mathrm{model}} + \lambda_2 \mathcal{L}_{\mathrm{CSL}}^{\mathrm{task}} + \lambda_3 \mathcal{L}_{\mathrm{CTL}}^{\mathrm{model}}, \tag{8}$$

where λ_1,λ_2 and λ_3 are the ramp-up weighting coefficients for balancing these four terms.

3.4. Training details

Our network is implemented by Pytorch framework and trained on two NVIDIA GeForce 2070SUPER GPUs with total memory of 16GB. We utilize Adam optimizer to train the whole network for 100 epochs with learning rate of 1e-4 and batchsize of 2. To achieve a balance between the training efficiency and the uncertainty map quality, we perform $K = 8$ times MC dropout in the teacher model. While in the test phase, we turn off the dropout to generate the estimation directly. For updating Eq. (2), we set τ as 0.99 according to (Tarvainen and Valpola, 2017). The setting of hyper-parameters α_1, α_2 in Eq. (4) and μ_1, μ_2 in Eq. (5) are studied in the following experiments. As for λ_1 and λ_2 in Eq. (8), following (Tarvainen and Valpola, 2017), we set them equally as a time-dependent Gaussian warming-up function $\lambda(t) = 0.1 * e^{(-5(1-t/t_{max})^2)}$ where t and t_{max} indicate the current training step and total training steps, respectively. For λ_3, considering that our ultimate goal is to determine classes at pixel level, continuous contrastive learning would force the network to focus more on high-level semantic information at the possible sacrifice of pixel-level details, thus raising an adverse effect to the pixel-level segmentation in later training. Accordingly, we empirically set it to 0.01 for the first 30 epochs and 0 afterwards. Additionally, the size of the queue to store negative keys is set to 100. Note that, the detailed training process of our model is presented in Algorithm 1, and only the shared encoder and the segmentation decoder in the student model are retained for generating segmentation predictions in the testing phase.

4. Experiments and results

4.1. Experimental settings

Dataset: We evaluate our method on the public datasets of 2017 ACDC challenge (Bernard et al., 2018) for cardiac segmentation and PROMISE12 (Litjens et al., 2014) for prostate segmentation. Detailed data statistics regarding the two datasets are listed in Tables 1

Algorithm 1 The core learning algorithm of our method.

input: N labeled training data $\mathcal{D}^l = \{\boldsymbol{X}^i, \boldsymbol{Y}^i, \boldsymbol{Z}^i, \boldsymbol{G}^i\}_{i=1}^{N}$, M unlabeled training data $\mathcal{D}^u = \{\boldsymbol{X}^i\}_{i=N+1}^{N+M}$. $\boldsymbol{X}^i \in \mathbb{R}^{H\times W}$ is the intensity image, $\boldsymbol{Y}^i \in \{0,1\}^{H\times W}$ is the corresponding segmentation label, $\boldsymbol{Z}^i \in \mathbb{R}^{H\times W}$ is the ground truth SDF derived from $\boldsymbol{Y}^i$ via the SDF function in (Xue et al., 2020), $\boldsymbol{G}^i \in \mathbb{R}^{2\times H\times W}$ contains the ground truths of foreground and background obtained by $\boldsymbol{Y}^i \odot \boldsymbol{X}^i$ and $(1-\boldsymbol{Y}^i) \odot \boldsymbol{X}^i$ where $\odot$ refers to element-wise multiplication.
output: The trained student model and teacher model.
initialize: *epoch*=0, *total_epoch*=100, K=8.
while *epoch* < *total_epoch*
  Add noise ξ to the image $\boldsymbol{X}^i$ and input it into the student model, obtaining the output of the student encoder f_s, the output of the segmentation task $\tilde{\boldsymbol{Y}}_S^i$, the output of the reconstruction task $\tilde{\boldsymbol{G}}_S^i$, and the output of the SDF prediction task $\tilde{\boldsymbol{Z}}_S^i$.
  Add noise ξ' to the image $\boldsymbol{X}^i$ and input it into the teacher model, obtaining the output of the teacher encoder f_t^+.
  for all $j \in \{1,\ldots,K\}$ **do** # MC sampling
    Input f_t^+ into the three decoders of teacher model with different dropouts to obtain the output of the segmentation task $\tilde{\boldsymbol{Y}}_T^{ij}$, the output of the reconstruction task $\tilde{\boldsymbol{G}}_T^{ij}$, and the output of the SDF prediction task $\tilde{\boldsymbol{Z}}_T^{ij}$.
  end for
  Integrate K sampling results $\{\tilde{\boldsymbol{Y}}_T^{ij}\}_{j=1}^{K}$ using the proposed UWI strategy.
  Integrate K sampling results $\{\tilde{\boldsymbol{G}}_T^{ij}\}_{j=1}^{K}$ and $\{\tilde{\boldsymbol{Z}}_T^{ij}\}_{j=1}^{K}$ using the average operation.
  Calculate the tripled-uncertainty U_{seg}, U_{rec} and U_{sdf} of the three tasks with Eq. (3).
  if $\boldsymbol{X}^i \in \mathcal{D}^l$
    Calculate the supervised loss $\mathcal{L}_S$ with Eq. (4)
  end if
  Calculate the inter-model consistency loss $\mathcal{L}_{\text{CSL}}^{\text{model}}$ and the inter-task consistency loss $\mathcal{L}_{\text{CSL}}^{\text{task}}$ with Eqs. (5) and (6).
  Couple the query f_s with the current key f_t^+ to construct the positive pair (f_s, f_t^+), and with the previous key f_t^- from the queue to construct the negative pair (f_s, f_t^-).
  Calculate the inter-model contrastive loss $\mathcal{L}_{\text{CTL}}^{\text{model}}$ with Eq. (7).
  Update the student model by gradient descent and the teacher model by EMA of student's parameters.
  Enqueue the current key and dequeue the oldest key.
  epoch=*epoch*+1
end while

and 2. The 2017 ACDC dataset contains 100 subjects belonging to five types of conditions: healthy, previous myocardial infraction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. Each type contains 20 subjects. Our target is to segment the left ventricle region out. For fair training and evaluation, 75 subjects are assigned to the training set, 5 to the validating set and 20 to the testing set. The PROMISE12 dataset has 50 transversal T2-weighted magnetic resonance imaging (MRI) images collected from four centers: Haukeland University Hospital (HK) in Norway, the Beth Israel Deaconess Medical Center (BIDMC) in the US, University College London (UCL) in the United Kingdom and the Radboud University Nijmegen Medical Centre (RUNMC) in the Netherlands. We randomly selected 35 samples as the training set, 5 as the validating set and 10 as the testing set. In order to study the impact of different amounts of labeled and unlabeled data on model performance, we further divide the training set into different combinations of labeled set and unlabeled set, denoted as n/m, where n and m are the numbers of labeled and unlabeled samples, respectively.

Evaluation: To quantitatively assess the performance, we use two standard evaluation metrics, i.e., Dice coefficient ($\text{Dice} = \frac{2*|X\cap Y|}{|X|+|Y|}$) and Jaccard Index ($\text{JI} = \frac{|X\cap Y|}{|X\cup Y|}$). Both Dice and JI evaluate the overlap area between the prediction and the ground-truth. Higher scores indicate better segmentation performance for both the metrics. Furthermore, we also visualize the segmentation results to intuitively compare different segmentation methods.

4.2. Investigation of trade-off hyper-parameters

We have four hyper-parameters to balance the loss terms, i.e., α_1, α_2 in Eq. (4) and μ_1, μ_2 in Eq. (5). To investigate their optimal values, we perform experiments on the 2017 ACDC dataset based on the 5-fold cross-validation strategy. Concretely, referring to (Ruan et al., 2021), we first roughly locate α_1, α_2, μ_2 in the interval of [0.2,4], μ_1 in the interval of [0.05,1]. Then, we alternatively search the optimal value of one hyper-parameter and fix the remaining ones until all four hyper-parameters have been traversed. The results are given in Fig. 4.

In Fig. 4(a), we fix α_2 as 1, μ_1 as 0.2, μ_2 as 1. As we can see, the Dice and JI show a positive correlation with α_1 when α_1 increases from 0.2 to 1, but once α_1 exceeds 1, the Dice and JI decrease. Accordingly, we determine α_1 as 1. Next, in Fig. 4(b), we study the impact of different values of α_2 when α_1 is determined as 1, μ_1 is fixed as 0.2, μ_2 is fixed as 1. Evidently, when α_2 arrives at 1, the Dice and JI reach a top performance, indicating an accurate prediction. Similar operations are conducted on the investigation of μ_1 and μ_2, and the results are displayed in Fig. 4(c) and (d), respectively. On the basis of these results, we can determine the optimal selection of these trade-off hyper-parameters, i.e., $\alpha_1 = 1$, $\alpha_2 = 1$, $\mu_1 = 0.2$, $\mu_2 = 1$, which will be applied in the following experiments.

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

Comparison on 2017 ACDC: To demonstrate the superiority of the proposed method in leveraging unlabeled and labeled data, we compare our method with several methods, including U-net (Ronneberger et al., 2015), Curriculum (Kervadec et al., 2019), Mean Teacher (Tarvainen and Valpola, 2017), Uncertainty-aware Mean Teacher (UA-MT) (Yu et al., 2019), Multi-task Attention-based Semi-supervised Learning (MASSL) (Chen et al., 2019), Shape-aware (Xue et al., 2020) and Mutual Information based Semi-supervised Segmentation (MISSS) (Peng et al., 2021). It is worth noting that only U-net is trained in a full-supervised manner with the limited labeled data while others are semi-supervised. Table 3 is a summary of quantitative results on 2017 ACDC dataset in different n/m settings. As observed, our proposed method outperforms all the compared methods with the highest Dice and JI values in all n/m settings. Specifically, compared with the full-supervised U-net, our method can leverage the unlabeled data and largely improve Dice and JI from 60.1%, 47.3% to 86.4%, 77.6%, respectively, when n=5. For semi-supervised methods, UA-MT can be regarded as a prototype which also integrates the uncertainty into the mean teacher model. Compared with UA-MT, our method generates much better results for all the n/m cases. Even for the suboptimal approaches, e.g., Shape-aware and MISSS, our method still boosts the quantitative metrics by 2%-5%, especially when few labeled data is

Table 1
Data statistics of 2017 ACDC dataset.

Condition	Healthy	Previous Myocardial Infraction	Dilated Cardiomyopathy	Hypertrophic Cardiomyopathy	Abnormal Right Ventricle
Number	20	20	20	20	20
Classes	Background, Right Ventricle, Myocardium and Left Ventricle				

Table 2
Data statistics of PROMISE12 dataset.

Center	Field Strength	Endorectal coil	Resolution (in-plane/through-plane in mm)	Manufacturer
HK	1.5T	Yes	0.625/3.6	Siemens
BIDMC	3T	Yes	0.25/2.2-3	GE
UCL	1.5 and 3T	No	0.325-0.625/3-3.6	Siemens
RUNMC	3T	No	0.5-0.75/3.6-4.0	Siemens

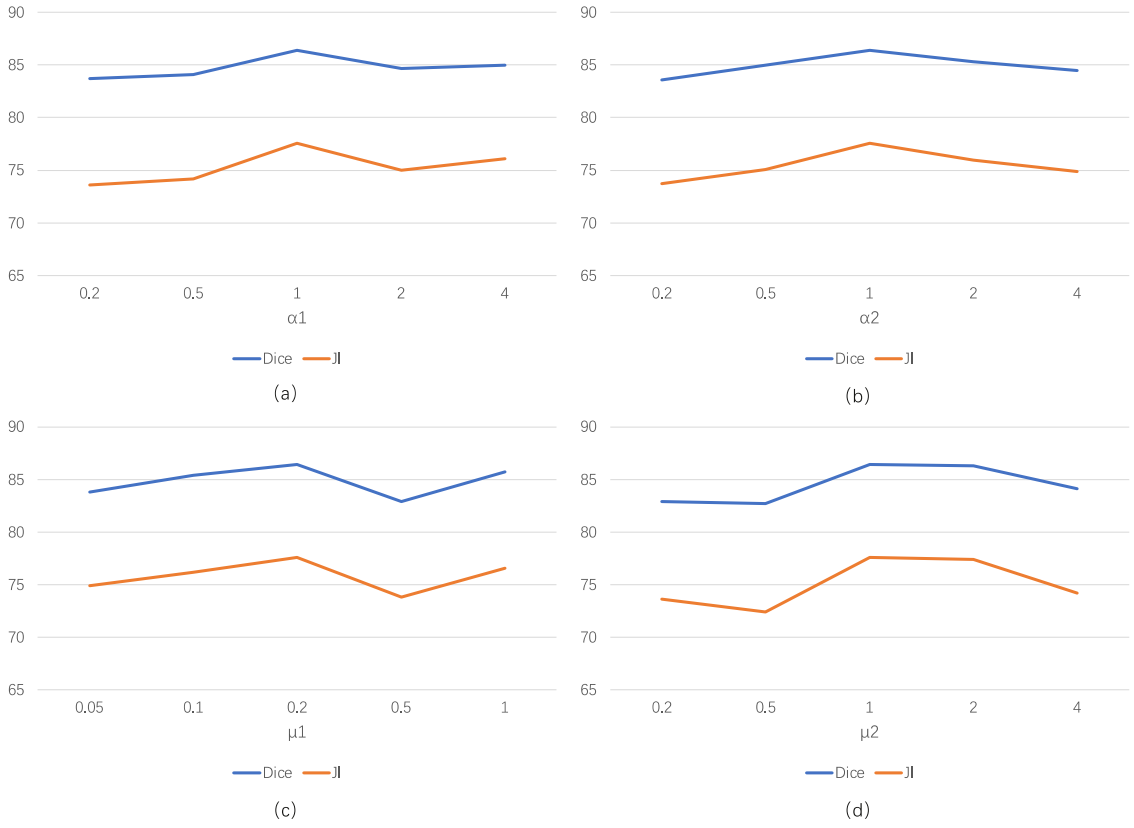

Fig. 4. Investigation on effect of different values of (a) α_1, (b) α_2, (c) μ_1, and (d) μ_2 in terms of Dice and JI.

Table 3
Quantitative comparison results on 2017 ACDC dataset. * means our method is significantly better than the compared method with $p<0.05$ via paired t-test.

n/m	5/70		10/65		20/55	
	Dice[%]	JI[%]	Dice[%]	JI[%]	Dice[%]	JI[%]
U-net (2015)	60.1(24.7)*	47.3(23.4)*	70.9(23.6)*	53.1(28.5)*	90.0(7.7)	82.5(11.3)
Curriculum (2019)	67.5(9.9)*	51.8(10.5)*	69.2(14.6)*	50.0(14.5)*	86.6(9.0)*	77.5(13.0)*
Mean Teacher (2017)	52.1(18.7)*	37.3(16.2)*	80.0(14.3)*	68.8(17.5)*	88.8(9.3)*	80.9(13.4)*
UA-MT (2019)	70.7(14.1)*	56.4(15.9)*	80.6(17.8)*	70.7(21.8)*	88.7(10.5)*	81.2(14.7)*
MASSL (2019)	77.4(16.7)*	66.0(18.5)*	86.0(14.9)	77.8(18.5)	90.6(8.8)	84.0(12.6)
Shape-aware (2020)	81.4(14.2)*	70.8(16.9)*	85.0(12.2)*	75.6(15.8)*	91.0(7.8)	84.3(11.5)
MISSS (2021)	81.2(20.9)*	72.4(23.4)*	84.7(15.2)	75.8(18.6)	91.2(5.6)	84.3(8.9)
Proposed	**86.4(11.2)**	**77.6(14.8)**	**87.5(12.9)**	**79.6(16.7)**	**91.4(7.4)**	**84.9(11.1)**

available. Nevertheless, when comparing the performance of the full-supervised U-net with the semi-supervised methods, we find some fluctuations that some semi-supervised methods, e.g., Curriculum, Mean Teacher, perform even worse than U-net. The reason may be that when insufficient labeled knowledge can be accessed, these semi-supervised methods tend to introduce more unintended noise in the unlabeled data learning, thus degrading the performance. In contrast, the methods including UA-MT, MASSL, Shape-aware and MISSS, impose uncertainty estimation, attention, shape constraint and mutual information regularization on the learning of unlabeled data, improving the robustness of corresponding models. Therefore, even at label levels of only 6% and 13%, these semi-supervised methods can surpass U-net with a large gap.

To check whether the improvements by our method against other methods are statistically significant, we conduct the paired t-test on all the comparison results. As can be viewed from Table 3, when few labeled data is available, our proposed method has statistically significant improvements over all the counterparts with $p<0.05$, demonstrating the superiority of the proposed methods. With more labeled data available, all the methods approach more supervision and show an upward trend with a narrowing gap, but

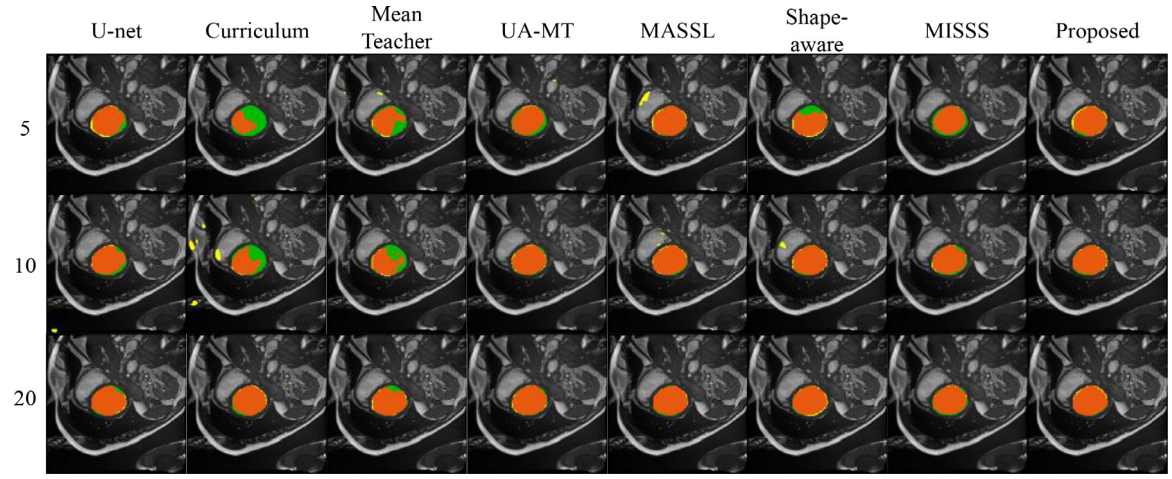

Fig. 5. Visual comparison results on 2017 ACDC dataset with 5, 10, 20 labels. Orange indicates the correct segmented area, green the unidentified and yellow the miss-identified.

Table 4
Quantitative comparison results on PROMISE12 dataset. * means our method is significantly better than the compared method with $p<0.05$ via paired t-test.

n/m	10/25		15/20	
	Dice[%]	JI[%]	Dice[%]	JI[%]
U-net (2015)	56.7(24.0)*	47.3(23.4)*	60.6(20.2)*	46.0(17.5)*
Mean Teacher (2017)	64.6(13.5)*	49.2(13.9)*	67.0(10.0)*	51.2(10.9)*
UA-MT (2019)	62.2(23.5)*	49.0(22.7)*	65.0(18.2)*	50.9(20.5)*
MASSL (2019)	64.2(13.5)*	48.7(13.5)*	67.5(11.3)*	52.1(12.3)*
Shape-aware (2020)	65.4(13.1)*	50.0(13.4)*	68.3(16.9)*	54.0(16.0)*
MISSS (2021)	62.4(12.7)*	46.5(13.3)*	70.4(17.6)	55.3(18.5)
Proposed	**67.4(10.8)**	**51.9(11.6)**	**70.4(11.2)**	**55.5(12.4)**

our method still ranks first. Standard deviations are also provided in the brackets in Table 3, where we can see that the proposed method achieves the smallest standard deviation, indicating the excellence and robustness of our method.

We also give qualitative comparison results on 2017 ACDC dataset in Fig. 5. It can be seen from the figure that the target area is better sketched by our method with more accurate boundaries. By contrast, when a small amount of labeled data is available, the results of other methods always contain more over-segmented or under-segmented regions (marked by green and yellow respectively). Such mis-predictions are particularly severe in methods other than UA-MT, which can be interpreted as that the uncertainty estimation in UA-MT effectively circumvents prediction regions with poor confidence reduced. As for Shape-aware and MISSS which achieve high quantitative results, they also show good visual performance consistently. Nevertheless, compared with all the methods, our method produces the fewest false positive and false negative predictions, thereby generating the most similar segmentation results to the ground truths over all compared methods. This demonstrates that the proposed method can exploit the large amount of unlabeled data more effectively.

Comparison on PROMISE 12: To inspect the generalization ability of our model to the other dataset, we further conduct the comparison experiments on the PROMISE12 dataset. The statistical and visual results are shown in Table 4 and Fig. 6, respectively. Similarly, in Table 4, we can see that our proposed method achieves the best performance in both Dice and JI measures and has statistically significant improvements over most of the methods. Compared with the full-supervised method, our method improves the performance dramatically, rising Dice from 56.7% to 67.4% and JI from 47.3% to 51.9% when n=5, which supports the benefit of semi-supervised learning. Among the semi-supervised methods, Shape-aware takes the second place with 65.4% Dice and 68.3 % Dice in the case of 10 labels and 15 labels, respectively. However, it still lags behind our method by 2.0% and 2.1%. Consistently, our method ranks first with the smallest standard deviation.

Observing the visual results in Fig. 6, it is easy to find that U-net is only able to segment a small portion of target areas, which stems from the supervised training manner with rare labels. In contrast, the semi-supervised methods can assuage this problem to some extent by expanding the training dataset to unlabeled data and providing more knowledge to the learning of the network. Despite so, there are still large fluctuations among the results of different approaches. Consistent with the highest Dice and JI values in Table 4, our method covers the largest orange areas and the smallest yellow and green areas, which means that a more accurate target region can be distinguished by our method.

All in all, whether on 2017 ACDC dataset or PROMISE12 dataset, our method yields the best quantitative results and presents the visual segmentation effects closest to the ground truths, demonstrating the overwhelming advantage of our method.

4.4. Ablation study

To investigate the contributions of key modules of our method, we further conduct a series of experiments in different model settings on 2017 ACDC dataset. Similarly, the paired t-test is also performed to inspect the statistical significance of the improvements brought by the components.

Contribution of the two auxiliary tasks: First, to validate the effectiveness of two auxiliary tasks, we compare the models of (1) the segmentation task alone (Seg), (2) the segmentation task and the SDF prediction task (Seg+Sdf), and (3) all the three tasks (Seg+Sdf+Rec). Note that the teacher model is not incorporated in

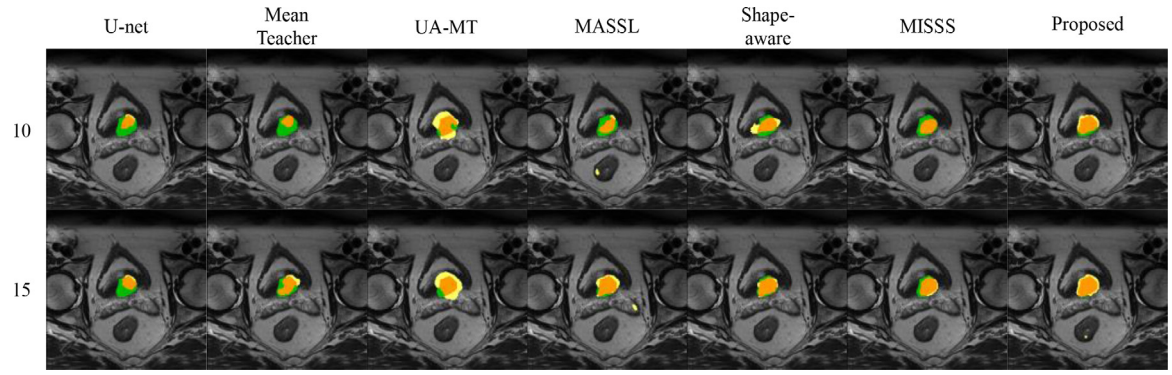

Fig. 6. Visual comparison results on PROMISE12 dataset with 10, 15 labels. Orange indicates the correct segmented area, green the unidentified and yellow the miss-identified.

Table 5
Ablation study on key modules on 2017 ACDC dataset. * means our method is significant better than compared method with $p < 0.05$ via paired t-test.

n/m	5/70		10/65		20/55	
	Dice[%]	JI[%]	Dice[%]	JI[%]	Dice[%]	JI[%]
Seg	60.1(24.7)*	47.3(23.4)*	70.9(23.6)*	53.1(28.5)*	90.0(7.7)	82.5(11.3)
Seg+Sdf	81.0(16.7)*	70.8(19.5)*	83.4(17.2)*	74.5(20.2)*	90.2(8.5)	83.1(12.5)
Seg+Sdf+Rec	81.2(13.8)*	70.9(17.8)*	84.2(15.1)*	75.4(18.6)*	90.7(7.2)	83.8(10.8)
S+T	52.1(18.7)*	37.3(16.2)*	80.0(14.3)*	68.8(17.5)*	88.8(9.3)*	80.9(13.4)*
S+T+UncA	70.8(22.4)*	58.8(22.6)*	82.8(20.5)*	74.7(23.0)*	90.2(8.2)	83.1(12.0)
S+T+UncW	79.3(23.4)*	70.5(25.3)*	83.2(18.7)*	74.7(21.8)*	90.7(7.8)	83.8(11.6)
Proposed w/o. Tri-U and CTL	82.6(13.0)*	72.4(16.6)*	86.0(12.2)	77.1(16.0)*	90.9(8.4)	84.2(12.4)
Proposed w/o. CTL	84.6(13.9)*	75.6(17.6)*	87.1(11.5)	78.8(15.5)	91.3(7.7)	84.9(11.6)
Proposed	**86.4(11.2)**	**77.6(14.8)**	**87.5(12.9)**	**79.6(16.7)**	**91.4(7.4)**	**84.9(11.1)**

the implementation of these three models. The quantitative results are shown in the upper part in Table 5, from which we can see that the Seg model exhibits the worst performance. With the SDF prediction task and the reconstruction task joining, Dice and JI are improved to varying degrees, showing the positive impact of both the two auxiliary tasks on segmentation performance. Especially when n is small, the improvement by Sdf is significant, revealing its considerable contribution to the utilization of unlabeled data.

Contribution of the UWI module: Second, since the UWI module is only applied in the segmentation task, here we exclude the influence of other factors and keep only the Seg model in the student model (S) and the teacher model (T). To study the impact of the UWI module, we first combine the student model and the teacher model as a naïve mean teacher model (S+T), then incorporate the uncertainty estimation with averaging strategy (UncA), and the uncertainty estimation with the proposed UWI strategy (UncW) to the S+T model, and compare their results quantitatively. The middle part of Table 5 presents the detailed results. We can find that the S+T model decreases Dice and JI by 8.0% and 10% compared with S only when n=5. This may be explained as that the teacher model is susceptible to noise when labeled data is few, thus degrading the performance. However, by considering the uncertainty estimation, the S+T+UncA model significantly rises Dice and JI remarkably by 18.7% and 21.5% for n=5. Furthermore, our proposed S+T+UncW model improves Dice and JI by 8.5% and 11.7% when n=5, compared with the S+T+UncA model. The higher metric values prove the effectiveness of the proposed UWI strategy, which can be attributed to the fact that it assigns different weights for different MC sampling results.

Contribution of the tripled-uncertainty: Third, to verify the guiding role of the tripled-uncertainty, we firstly remove the contrastive loss (CTL) from the proposed model, and then compare the clipped models with and without the tripled-uncertainty (Tri-U), i.e., proposed w/o. Tri-U and CTL vs. proposed w/o. CTL. The results are displayed in the bottom of Table 5. As observed, the engagement of the tripled-uncertainty gains better performance with Dice values of 84.6%, 87.1%, 91.3% and JI values of 75.6%, 78.8 %, 84.9%, surpassing the proposed w/o. Tri-U and CTL model. It demonstrates that the uncertainty map can effectively guide the student to learn more reliable knowledge from the teacher, thus achieving higher segmentation performance.

Contribution of the contrastive learning: Next, we make a comparison of the complete proposed model and the model without CTL, i.e., proposed vs. proposed w/o. CTL, to evaluate the effectiveness of the contrastive loss. The results can also be found in the bottom of Table. 5. With the assistance of the contrastive loss, our model further boosts the performance by 1.8%, 0.4%, 0.1% on Dice when n=5, 10 and 20 respectively, which proves the positive role of contrastive loss.

Contribution of the model-level consistency losses: Finally, we perform the ablation experiments in an extreme case of n=5 to validate the effectiveness of consistency losses proposed in our model. In this section, we focus on investigating the contribution of the model-level consistency losses. To achieve this, we detach them from the complete model, respectively, to check their impact on the final segmentation result. Concretely, the models without the consistency loss on the SDF prediction task, the reconstruction task, and the segmentation task are separately denoted as Proposed w/o sdf_CSL, Proposed w/o rec_CSL, and Proposed w/o seg_CSL. The results are reported in the upper part in Table 6. As we can see, the removal of different model-level consistency losses could lead to various degrees of performance degradation, especially for the consistency loss of the segmentation task which significantly decreases from 86.4 % to 82.3 % in Dice. These results imply that the model-level consistency loss is really helpful for enhancing the student model.

Contribution of the task-level consistency losses: Besides the model-level consistency losses, we also study the contribution of

Table 6
Ablation study on consistency losses on 2017 ACDC dataset.

Type		Dice	JI
Model-level Consistency	Proposed w/o sdf_CSL	85.5(13.9)	76.1(17.5)
	Proposed w/o rec_CSL	85.8(16.7)	77.0(19.1)
	Proposed w/o seg_CSL	82.3(16.3)	72.0(19.1)
Task-level Consistency	Proposed w/o sdf_seg_CSL	83.0(15.7)	76.1(17.5)
	Proposed w/o rec_seg_CSL	85.7(15.8)	76.7(18.2)
	Proposed	**86.4(11.2)**	**77.6(14.8)**

Table 7
Comparison between the pixel-level contrastive loss and the image-level contrastive loss.

	Dice	JI
Pixel-level contrastive loss	86.54(13.4)	77.8(16.7)
Image-level contrastive loss (proposed)	86.4(11.2)	77.6(14.8)

the task-level consistency losses. In detail, we compare the complete proposed model with the its variant which discards the consistency loss between the SDF prediction task and the segmentation task (denoted as Proposed w/o sdf_seg_CSL) or between the reconstruction task and the segmentation task (denoted as Proposed w/o rec_seg_CSL). From the middle part of Table 6, we can find that the task-level consistency losses can boost the performance by 3.4% Dice and 0.7% Dice in these two cases, demonstrating that constraining the consistency between related task can help excavate more information beneficial to segmentation.

4.5. Image-level contrastive loss vs. pixel-level contrastive loss

Considering that our segmentation task is actually performed by pixel-wise classification, we further investigate whether the pixel-level contrastive loss has an advantage over the image-level contrastive loss. Concretely, we try to replace the image-level contrastive loss with the pixel-level contrastive loss proposed in (Zhong et al., 2021). The results are given in Table 7, where we can see that the performance of pixel-level contrastive loss is on par with that of the image-level contrastive loss. The reason for this result can be explained as follows. Compared with natural images, it is harder for medical images to identify the different-class regions (usually the tumor pixels) due to their small proportion of the whole image and their varied appearance from person to person. To accurately distinguish different-class pixels, high-level expertise is needed for supervision. However, in our paper, we address a more challenging circumstance where most supervision is unavailable. This could bring obstacles for the pixel-level contrastive loss which requires to explicitly distinguish different-class pixels. Therefore, the performance of the pixel-level contrastive loss is compromised. Meanwhile, considering the large amount of computational resource to be consumed for pixel-level comparison loss, we therefore keep the contrastive learning in the form of image-level.

5. Discussion

In this paper, to alleviate the overfitting problem caused by limited manual labels in the medical image segmentation field, we proposed a novel semi-supervised model based on the mean teacher architecture (Tarvainen and Valpola, 2017), which makes use of the adequate unlabeled data to provide more useful information for learning from labeled data. Traditional mean teacher based methods always attempt to constrain the consistency between the outputs from two augmented inputs, i.e., data-level, or from the teacher and student models, i.e., model-level. However, we argue that such constraints are not sufficient to help the network learn some of more task-relevant features. In this regard, we injected the multi-task learning spirit into our network, and additionally introduced the task-level consistency. Specifically, we integrated the foreground and background reconstruction task and the signed distance field (SDF) prediction task into our network to aid the segmentation task to perceive the semantic and shape information of the target. For each task, the output of the teacher model is supposed to regularize the student model training. Nevertheless, when few labeled data is available, the possible overfitting problem may cause the poor quality of the teacher predictions, thus misleading the student to a suboptimal point. To relieve the adverse influence of the misleading knowledge, we introduced the tripled-uncertainty into the consistency losses for the three tasks, guiding the student model to learn more reliable information from the teacher model. When calculating the uncertainty, we found that the Monte Carlo (MC) sampling results varied with each forward pass. To produce a more credible teacher prediction, we designed an uncertainty weighted integration (UWI) module to aggregate the MC samplings for segmentation. Finally, in addition to the consistency loss at the output level, we also incorporated the contrastive loss at the feature level to reinforce the encoder, allowing it to extract distinct features of different samples.

To demonstrate the feasibility and generalizability of our proposed method, we conducted the experiments on both 2017 ACDC and PROMISE12 datasets for cardiac segmentation and prostate segmentation, respectively. In the case of the same amount of labeled data, we can see a significant improvement by our proposed method over the full-supervised counterpart. When comparing with other state-of-the-art semi-supervised segmentation methods, our method is always better in both qualitative and quantitative measures. Even when applying our method to different medical image segmentation tasks, the first place of our method still remains unchanged. These results not only show that our method can effectively leverage the unlabeled data to support the training of labeled data, but also indicate the method is advanced and has a certain universality for different medical image segmentation tasks.

Furthermore, a set of ablation experiments were performed on 2017 ACDC datasets to verify the effectiveness of the key modules in our proposed method. Specifically, we first studied the influence of the two auxiliary tasks, i.e., the reconstruction task and the SDF prediction task, on the basis of the only student model. Experimental results show that the performance is remarkably increased by approximately 20% with more auxiliary tasks integrated when labels are limited. Second, we investigated the UWI strategy based on the mean teacher which preserved only the segmentation task by comparing it to the average integration strategy. We found that although the average integration strategy can achieve a nearly 20% rise when labeled data is scarce, our UWI strategy further boosts the performance by approximately 10%, demonstrating its contribution. Then, we removed the triple uncertainty and the contrastive loss to evaluate their impacts. Not surprisingly, dropping either the triple uncertainty or the contrastive loss results in varying degrees of performance degradation, which implies that they do have positive effects on our model regarding the semi-supervised segmentation task. Next, we validated the contributions of the model-level and task-level consistency losses and found that both kind of consistency losses can contribute to the training of the segmentation model. Finally, we compared the pixel-level and image-level contrastive losses in our semi-supervised segmentation task and proved the image-level contrastive loss is more suitable in our setting.

Although the proposed method has achieved promising results, there are also some limitations worth discussing. On the one hand, we only optimized the integration of preliminary sampling results

for the most relevant segmentation task in the teacher network, but took the simple average operation for the preliminary results of the auxiliary tasks. The performance of the two auxiliary tasks, however, is also significant for our method. Therefore, how to treat the differences of the preliminary results of the auxiliary tasks in teacher model and optimize the integration of them is a topic worth exploring. On the other hand, when performing the data-level consistency, we only added some Gaussian noises to the inputs, which are relatively weak for disturbing the data distribution. To strengthen the learning ability and generalization of our model, in the future, we would like to involve more stronger perturbations, such as flip, rotation, crop or color jitter, to make the augmented input as different as possible from the original one, while preserving the high-level semantic information.

6. Conclusion

In this paper, to alleviate the limitation caused by label scarcity, we proposed a tripled-uncertainty guided semi-supervised model for medical image segmentation, which can effectively utilize the unlabeled data to improve the segmentation performance. Based on a mean teacher architecture, our model explores the relationship among the segmentation task, the foreground and background reconstruction task and the SDF prediction task. To eliminate the possible misdirection caused by the noisy unlabeled data, we employed uncertainty estimation on all three tasks in the teacher model. In contrast to the common uncertainty averaging integration strategy, we considered the differences of each sampling and developed a novel uncertainty weighted integration strategy. In addition, we further equipped our model with the contrastive learning constraint to enhance the representative capability of the encoders. The experimental results on two public medical datasets have demonstrated the feasibility and superiority of our method.

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.

CRediT authorship contribution statement

Kaiping Wang: Conceptualization, Methodology, Investigation, Writing – original draft, Validation. **Bo Zhan:** Conceptualization, Methodology, Investigation, Writing – original draft, Validation. **Chen Zu:** Resources, Data curation, Validation, Writing – review & editing. **Xi Wu:** Methodology, Investigation, Software. **Jiliu Zhou:** Resources, Data curation, Validation. **Luping Zhou:** Writing – review & editing, Investigation. **Yan Wang:** Funding acquisition, Conceptualization, Supervision, Writing – review & editing, Validation, Project administration.

Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC 62071314) and Sichuan Science and Technology Program (2021YFG0326).

References

Amyar, A., Modzelewski, R., Li, H., Ruan, S., 2020. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Comput. Biol. Med. 126, 104037.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 253–260.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersson, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A., 2019. Multi-task learning for segmentation of building footprints with deep neural networks. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1480–1484.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 810–818.

Chen, S., Bortsova, G., Juárez, A.G.U., van Tulder, G., de Bruijne, M., 2019. Multi-task attention-based semi-supervised learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 457–465.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp. 1597–1607.

Chen, Z., Zhang, R., Zhang, G., Ma, Z., Lei, T., 2020b. Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation. IEEE Access 8, 41830–41837.

Cheng, F., Chen, C., Wang, Y., Shi, H., Cao, Y., Tu, D., Zhang, C., Xu, Y., 2020. Learning directional feature maps for cardiac mri segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 108–117.

Duan, J., Bello, G., Schlemper, J., Bai, W., Dawes, T.J., Biffi, C., Marvao, A., Doumound, G., Rueckert, D., 2019. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. IEEE Trans. Med. Imaging. 38 (9), 2151–2164.

Duong, L., Cohn, T., Bird, S., Cook, P., 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), pp. 845–850.

Fang, Y., Chen, C., Yuan, Y., Tong, K.Y., 2019. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 302–310.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587.

Girshick, R., 2015. Fast r-cnn. In: 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1440–1448.

Guo, H., Pasunuru, R., Bansal, M., 2018. Dynamic multi-level multi-task learning for sentence simplification. arXiv preprint arXiv:1806.07304.

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.S., Qin, J., 2020. Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 562–571.

Hassani, K., Khasahmadi, A.H., 2020. Contrastive multi-view representation learning on graphs. In: International Conference on Machine Learning, pp. 4116–4126.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2018. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670.

Hu, L., Li, J., Peng, X., Xiao, J., Zhan, B., Zu, C., Wu, X., Zhou, J., Wang, Y., 2021. Semi-supervised NPC segmentation with uncertainty and attention guided consistency. Knowl.-Based Syst., 108021.

Huang, H., Lin, L., Zhang, Y., Xu, Y., Zheng, J., Mao, X., Qian, X., Peng, Z., Zhou, J., Chen, Y.W., Tong, R., 2021. Graph-BAS3Net: Boundary-Aware Semi-Supervised Segmentation Network With Bilateral Graph Convolution. In: 2021 IEEE/CVF International Conference on Computer Vision(CVPR), pp. 7386–7395.

Jia, J., Zhai, Z., Baker, M.E., Hernández-Girón, I., Staring, M., Stoel, B.C., 2021. Multi-task Semi-supervised Learning for Pulmonary Lobe Segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1329–1332.

Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D., 2020. Hard negative mixing for contrastive learning. arXiv preprint arXiv:2010.01028.

Kalluri, T., Varma, G., Chandraker, M., Jawahar, C.V., 2019. Universal semi-supervised semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5259–5270.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision?. arXiv preprint arXiv:1703.04977.

Kervadec, H., Dolz, J., Granger, E., Ayed, I.B., 2019. Curriculum semi-supervised segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 568–576.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Aaron M., Liu C., Krishnan, D., 2020. Supervised contrastive learning. arXiv preprint arXiv:2004.11362.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

Li, H., Peng, X., Zeng, J., Xiao, J., Nie, D., Zu, C., Wu, X., Zhou, J., Wang, Y., 2022. Explainable attention guided adversarial deep network for 3D radiotherapy dose distribution prediction. Knowl.-Based Syst., 108324.
Li, S., Zhang, C., He, X., 2020. Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 552–561.
Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikoz, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhushi, A., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. 18 (2), 359–373.
Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Trans. Med. Imaging 39 (3), 718–728.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.
Luo, X., Chen, J., Song, T., Wang, G., 2020. Semi-supervised medical image segmentation through dual-task consistency. arXiv preprint arXiv:2009.04448.
Luo, Y., Zhou, L., Zhan, B., Fei, Y., Zhou, J., Wang, Y., Shen, D., 2021. Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis. Med. Image Anal., 102335.
Milletari, F., avab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp. 565–571.
Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M., 2019. Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 7223–7226.
Nie, D., Gao, Y., Wang, L., Shen, D., 2018. Asdnet: Attention based semi-suervised deep networks for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 370–378.
Pandey, P., Pai, A., Bhatt, N., Das, P., Makharia, G., AP, P., 2021. Contrastive Semi-Supervised Learning for 2D Medical Image Segmentation. arXiv preprint arXiv:2106.06801.
Park, S., Hwang, W., Jung, K.H., 2018. Integrating reinforcement learning to self training for pulmonary nodule segmentation in chest x-rays. arXiv preprint arXiv:1811.08840.
Park, T., Efros, A.A., Zhang, R., Zhu, J.Y., 2020. Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. Springer, pp. 319–345.
Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C., 2020. Deep co-training for semi–supervised image segmentation. Pattern Recognit. 107, 107269.
Peng, J., Pedersoli, M., Desrosiers, C., 2020. Mutual information deep regularization for semi-supervised segmentation. Med. Imaging Deep Learn. 601–613.
Peng, J., Pedesoli, M., Desrosiers, C., 2021. Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization. arXiv preprint arXiv:2103.04813.
Rebuffi, S.A., Ehrhardt, S., Han, K., Vedaldi, A., Zisserman, A., 2020. Semi-supervised learning with scarce annotations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 762–763.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H., 2021. Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7660–7669.
Ruder, S., 2017. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.
Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., Garnavi, R., 2019. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 282–290.
Shi, Y., Zu, C., Hong, M., Zhou, L., Wang, L., Wu, X., Zhou, J., Zhang, D., Wang, Y., 2022. ASMFS: Adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease. Pattern Recognit. 126, 108566.
Song, L., Lin, J., Wang, Z.J., Wang, H., 2020. An end-to-end multi-task deep learning framework for skin lesion analysis. J. Biomed. Health Inform. 24 (10), 2912–2921.
Sun, T., Shao, Y., Li, X., Liu, P., Yan, H., Qiu, X., Huang, X., 2020. Learning sparse sharing architectures for multiple tasks. Proc. AAAI Conf. Artif. Intell. 34 (05), 8936–8943.
Sun, Y., Yang, H., Zhou, J., Wang, Y., 2022. ISSMF: Integrated semantic and spatial information of multi-level features for automatic segmentation in prenatal ultrasound images. Artif. Intell. Med., 102254.
Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., Wang, Y., 2022. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl.-Based Syst., 108215.
Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780.
Wang, Y., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., Shen, D., Zhou, L., 2018a. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage 174, 550–562.
Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2021. Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 450–460.
Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 542–551.
Wang, Y., Zhou, L., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., Shen, D., 2018b. 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. IEEE Trans. Med. Imaging. 38 (6), 1328–1339.
Wu, Z., iong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 3733–3742.
Xi, N., 2019. Semi-supervised Attentive Mutual-info Generative Adversarial Network for Brain Tumor Segmentation. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–7.
Xiao, T., Wang, X., Efros, A.A., Darrell, T., 2020. What should not be contrastive in contrastive learning. arXiv preprint arXiv:2008.05659.
Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W., Huang, X., 2020. Shape-aware organ segmentation by predicting signed distance maps. Proc. AAAI Conf. Artif. Intell. 34 (07), 12565–12572.
Yang, Y., Hospedales, T.M., 2016. Trace norm regularised deep multi-task learning. arXiv preprint arXiv:1606.04038.
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: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.
Zhan, B., Xiao, J., Cao, C., Peng, X., Zu, C., Zhou, J., Wang, Y., 2021. Multi-constraint generative adversarial network for dose prediction in radiotherapy. Med. Image Anal., 102339.
Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L., 2019. Et-net: A generic edge-attention guidance network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 442–450.
Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 408–416.
Zhang, Y., Li, H., Du, J., Qin, J., Wang, T., Chen, Y., Liu, B., Gao, W., Ma, G., Lei, B., 2021a. 3D Multi-Attention Guided Multi-Task Learning Network for Automatic Gastric Tumor Segmentation and Lymph Node Classification. IEEE Trans. Med. Imaging. 40 (6), 1618–1631.
Zhang, Y., Li, Y., Kong, Y., Wu, J., Yang, J., Shu, H., Coatrieux, G., 2021b. GSCFN: A graph self-construction and fusion network for semi-supervised brain tissue segmentation in MRI. Neurocomputing 455, 23–37.
Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng..
Zheng, H., Perrine, S.M.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020a. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 802–812.
Zheng, Z., Wang, X., Zhang, X., Zhong, Y., Yao, X., Zhang, Y., Wang, Y., 2020b. Semi–supervised Segmentation with Self-training Based on Quality Estimation and Refinement. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 30–39.
Zhong, Y., Yuan, B., Wu, H., Yuan, Z., Peng, J., Wang, Y., 2021. Pixel Contrastive–Consistent Semi-Supervised Semantic Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7273–7282.