Semi-supervised 3D Medical Image Segmentation using Multi-Consistency Learning with Fuzzy Perception-Guided Target Selection

Tao Lei, *Senior Member*, *IEEE*, Wenbiao Song, Weichuan Zhang, *Senior Member*, *IEEE*, Xiaogang Du, Chenxia Li, Lifeng He, *Senior Member*, *IEEE*, and Asoke K. Nandi, *Fellow*, *IEEE*

***Abstract*—Semi-supervised learning methods based on the mean teacher model have achieved great success in the field of 3D medical image segmentation. However, most of the existing methods provide auxiliary supervised signals only for reliable regions, but ignore the effect of fuzzy regions from unlabeled data during the process of consistency learning, which results in the loss of more valuable information. Besides, some of these methods only employ multi-task learning to improve models' performance, but ignore the role of consistency learning between tasks and models, thereby weakening geometric shape constraints. To address the above issues, in this paper, we propose a semi-supervised 3D medical image segmentation framework with multi-consistency learning for fuzzy perception-guided target selection. First, we design a fuzzy perception-guided target selection strategy from multiple perspectives and adopt the fusion method of fuzziness minimization and the fuzzy map momentum update to obtain a fuzzy region. By incorporating the fuzzy region into consistency learning, our model can effectively exploit more useful information from the fuzzy region of unlabeled data. Second, we design a multi-consistency learning strategy that employs intra-task and inter-model mutual consistency learning as well as cross-model cross-task consistency learning to effectively learn the shape representation of fuzzy regions. The strategy can encourage the model to agree on predictions for different tasks in fuzzy regions. Experiments demonstrate that the proposed framework outperforms the current mainstream methods on two popular 3D medical datasets, the left atrium segmentation dataset, and the brain tumor segmentation dataset. The code will be released at: https://github.com/SUST-reynole.**

***Index Terms*—Medical image segmentation, Semi-supervised learning, Fuzzy estimation, Consistency learning**

I. Introduction

MEDICAL image segmentation plays a key role in radiotherapy and computer-aided diagnosis by accurately and automatically segmenting anatomical structures or lesions, which can greatly improve diagnostic efficiency and accuracy [1]. In recent years, many supervised learning-based methods for medical image segmentation have been demonstrated to be effective, such as U-Net [2], TransUnet [3] and Swin-Unet [4]. However, these methods typically require large amounts of high-quality labeled data, but in practice the annotation of medical images is usually very expensive, especially for 3D medical images such as CT and MRI scans. One of the main reasons is that medical image annotation requires a tedious and specialized manual contouring process. Compared to supervised learning, semi-supervised learning is a new learning paradigm to solve the problem of incomplete data supervision [5]. It mainly utilizes unlabeled data to mine effective hidden information, thus improving models' performance. In addition to traditional supervised and semi-supervised methods, new approaches such as one-shot segmentation and template-based joint learning methods have also been reported. One-shot segmentation methods aim to learn effective segmentation models from a single annotated example, which is particularly useful in the medical field where annotated data are scarce [6]. These methods leverage prior knowledge and transfer learning to improve the model's generalization ability from minimal labeled data, thus reducing the dependency on extensive annotations [7]. Template-based joint learning methods, on the other hand, use pre-defined templates or atlases to guide the segmentation process [8]. By aligning new images with these templates, the model can effectively leverage prior structural information, which enhances segmentation accuracy and robustness, especially in complex anatomical regions [9]. Obviously, in the field of medical image segmentation, semi-supervised learning is more suitable for the needs of practical clinical scenarios.

Currently, a large number of semi-supervised medical image segmentation methods have emerged. Most of these methods firstly employ a popular encoder-decoder network architecture as the backbone, perform model pre-training using full supervision on a few labeled data, and then utilize data perturbation, model perturbation, or feature perturbation to achieve the effective utilization of unlabeled data, and finally update the network model. For most semi-supervised learning methods, their core idea is to use consistency learning to mine potential knowledge from unlabeled data and improve the generalization ability of the model. Consistency learning usually employs

This work was supported in part by the National Natural Science Foundation of China (Program No. 62271296, 62201334), in part by Scientific Research Program Funded by Shaanxi Provincial Education Department (23JP014, 23JP022). (Corresponding author: Weichuan Zhang)

This work did not involve human subjects or animals in its research.

T. Lei is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, and the Department of Geriatric Surgery, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China. (E-mail: leitao@sust.edu.cn)

W. Song is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China. (E-mail: 221611055@sust.edu.cn)

W. Zhang, X. Du and L. He are with the College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China. (E-mail: zwc2003@163.com; du423@sina.com; helifeng@ist.aichi-pu.ac.jp)

C. Li is with the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China.(E-mail: saphirli@sina.com)

A. K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, Middlesex, UB8 3PH, U.K., and the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China. (E-mail: asoke.nandi@brunel.ac.uk)

consistency regularization with different perturbations to train a network. One of the most representative methods is the Mean Teacher (MT) [10] and its variant models [11] [12] [13] [14] [15] [16]. This class of methods exploits the perturbation-based consistency loss between the teacher model and the student model, and the supervision loss on labeled data. They face the following two problems when they are applied to 3D medical image segmentation. First, to improve the performance of the model, these methods provide additional supervision in a complex and computationally overhead way [12] [14] and retain only reliable prediction regions when calculating the consistency loss. Reliable regions commonly occupy a large portion of an image but have a small consistency loss, and consistency loss commonly measures the average difference between the predictions of each individual voxel in the image, so the consistency learning on reliable regions potentially reduces the role of fuzzy regions that only occupy a small portion of an image. Second, consistency learning is usually applied to identical models or tasks [15] [17], ignoring the role of the combination of different models and different tasks, which has not fully explored and utilized the geometric shape information and inter-model differentiation information for medical image segmentation.

To solve the above problems, we propose a semi-supervised 3D medical image segmentation framework using multi-consistency learning with fuzzy perception-guided target selection. Since the fuzzy regions provide more useful information for medical image segmentation, the model can learn the most valuable regions from unlabeled data and does not focus too much on voxels that already have good classification. By performing multi-consistency learning on fuzzy regions, the model better encourages consensus in the most challenging regions, allowing network models based on semi-supervised learning to achieve better segmentation performance. The main contributions of this paper are summarized as follows:

- Different from the mainstream semi-supervised medical image segmentation networks that only rely on confidence or uncertainty for consistency learning [12] [14] [18], we propose a new fuzzy perception-guided target selection strategy. The proposed target selection strategy focuses only on the regions with high fuzziness in medical images through a fusion method, which improves the model's learning effect, achieves strong regularization constraints, and thus improves the generalization ability of our network model.
- Different from the mainstream semi-supervised medical image segmentation networks that only focus on intra-task or inter-model consistency learning [15] [17] [18], we propose a novel consistency learning strategy that combines tasks and models. Intra-task and inter-model mutual consistency learning as well as cross-model cross-task consistency learning enable our proposed network to simultaneously take into account geometric shape constraints and differential perturbations between models, which enhances the representation ability of our network on the shape information of targets in 3D medical images.
- Our proposed framework of fuzzy perception-guided target selection can be applied to any semi-supervised medical image segmentation task. We evaluated it experimentally on two public datasets, the 3D left atrium [48] and the 3D brain tumors [49]. It is demonstrated that our proposed framework outperforms state-of-the-art (SOTA) methods.

II. Related Work

A. Semi-supervised Medical Image Segmentation

To overcome the problem of insufficient labeled data in medical image segmentation, researchers have proposed many semi-supervised learning methods. Although semi-supervised methods based on deep learning can provide excellent segmentation results with their powerful feature representation and modeling capabilities [12], there are still challenges in applying these methods to complex 3D medical images. In order to better utilize unlabeled data, more methods focus on improving learning strategies, which are mainly classified into consistency learning [10] [11] [12] [18] [19] [20], pseudo-label learning [21] [22] [23], contrastive learning [24] [25] [26] [27] [28], and adversarial learning [29] [30] [31]. Among them, consistency learning focuses on maintaining the consistency of predictions under different data perturbations to improve models' robustness. However, it is more sensible to choose an appropriate consistency loss function. Pseudo-label learning uses predictions of unlabeled data as pseudo-labels and trains them together with labeled data, but it is easily interfered by noisy samples. Contrastive learning usually requires the design of effective similarity measures to enhance the feature representation of similar samples by learning similarity measures. Adversarial learning introduces adversarial networks to improve the robustness of a model by generating adversarial samples, but the stability and convergence of adversarial networks are the main challenges. Among these methods, consistency learning is the most popular for semi-supervised medical image segmentation in practical applications.

Consistency learning aims to learn useful features from both labeled and unlabeled images, computing regularly supervised loss on labeled images and unsupervised consistency loss on unlabeled images. Samuli et al. [32] first proposed a temporal ensembling strategy by using exponential moving averages (EMA) to predict unlabeled data as a consistency goal. However, maintaining EMA of the model's predictions during the training process, which increases the computational cost of the training process. To address this problem, Tarvainen et al. [10] used the EMA weights of the teacher model and the student model (MT) to achieve model training, so that the model is able to achieve consistency learning under different data perturbations. Specifically, MT first performs a supervision learning on labeled data, and then utilizes the prediction from the teacher model to provide a pseudo-label for unlabeled data, and leverages different learning strategies to ensure that the predictions of the teacher model and the student model on unlabeled data are consistent, and finally updates the student model with feedback on supervision and consistency losses. Most of the subsequent improvements [11] [12] [14] [15] [13] [28] employ more complex or computationally expensive methods to provide additional supervision signals for consistency regularization strategies, which improves the prediction quality on unlabeled data. For example, Chen et

al. [11] proposed a cross-pseudo-supervision (CPS) method based on model perturbations to encourage high consistency between different predictions from two perturbation models. However, evaluating the consistency between two predictions on unlabeled data can lead to unreliable guidance, which, in turn, affects the accuracy of the final model. For this, Yu et al. [12] proposed an uncertainty-aware framework (UA-MT) based on the MT model, which enables the student model to acquire more reliable targets by estimating uncertainty through multiple forward propagations. To reduce time and memory overhead, Luo et al. [14] [15] proposed to learn multi-scale consistency (URPC) from pyramid predictions at different scales to obtain target volume segmentation and constructed a dual-task-consistency (DTC) regularization method by jointly predicting pixel-by-pixel segmentation maps and geometrically-aware level-set representations of targets. Wu et al. [16] introduced a mutual consistency network (MC-Net) comprising two decoders, which captures model uncertainty information by evaluating the discrepancy between their predictions. The MC-Net can effectively improve pseudo-label quality by adding a regularization term. However, these networks ignore correlations between labeled and unlabeled data and only compute pixel-level consistency. For this reason, Lei et al. [34] adopted two discriminators (ASE-Net) based on consistency learning to obtain the prior relationship between labeled and unlabeled data and computed both the pixel-level and image-level consistency on unlabeled data under different data perturbations in order to improve the quality of predictions. However, the perceptual bias of the model may reduce its segmentation performance. For this, Wang et al. [35] proposed a mutual correction framework (MCF) through a comparative difference review module to find inconsistent prediction regions and dynamically select more reliable pseudo-labels. Although the above improved methods are better than the classic mean teacher model, most of them ignore improving consistency learning from the perspective of voxel target selection. In contrast to UA-MT [12], which is most relevant to voxel target selection and only selects reliable voxel targets with low uncertainty for consistency learning, we choose fuzzy voxel targets with high uncertainty due to higher learning value for model training.

B. Fuzzy Estimation

In the field of medical image segmentation, accurately quantifying fuzziness is crucial to evaluate the confidence of predicted regions [36], as it indicates the image regions in which the model is most likely to be incorrect for target segmentation [37]. However, determining the fuzzy regions in an image remains a challenging task. One of the most critical issues is how to effectively identify the fuzzy regions containing rich information that is more useful for accurate segmentation. In the field of image semantic segmentation, fuzziness is mainly used to guide semi-supervised learning to improve learning efficiency [33]. Fuzziness mainly originates from the estimation of uncertainty, because uncertainty can effectively depict fuzziness.

In semi-supervised learning, the utilization of fuzziness can be roughly divided into two groups. The first group focuses on learning deterministic regions by discarding ambiguous regions. For example, Yu et al. [12] used Monte Carlo sampling to estimate the perceptual uncertainty of each target prediction and only selected samples with low uncertainty for model training, so that the model could obtain more reliable guidance. Although this approach improves the accuracy and confidence of the model for unlabeled data, it ignores potential effectiveness of the regions with high uncertainty. Aiming at the shortcomings of the first group of methods, the second group focuses on learning only fuzzy regions to fully utilize unlabeled data. To identify fuzzy regions in unlabeled data, Czolbe et al. [33] argued that more information about fuzzy regions can be obtained from data with high uncertainty. In addition, Chen et al. [39] revealed the association between model uncertainty and error-prone fuzzy regions in image segmentation, emphasizing the importance of focusing on fuzzy regions. Meanwhile, Zheng et al. [40] obtained larger prediction variance values in regions with fuzzy predictions and pointed out that the estimation of fuzzy regions is related to the variance. The high variance regions mean higher uncertainty. Zheng et al. [40] also observed a considerable overlap between high variance regions and noise in pseudo-labels, which suggests that attention to fuzzy regions can be improved by focusing on high variance regions. To obtain the fuzzy regions in pseudo-labels, Zhang et al. [41], in a self-training based unsupervised domain adaptation study, found that the class prototype is less sensitive to errors in pseudo-labels, which can help to remove noise, and thus more accurately capture the fuzzy regions in pseudo-labels.

III. Methods

A. Overview

In this paper, we propose a semi-supervised 3D medical image segmentation framework for fuzzy perception-guided target selection with multi-consistency learning. As shown in Fig. 1, our framework can be represented as an ensemble of teacher and student models with the same network structure. The framework shares an identical encoder and two slightly different decoders. The segmentation branch uses the original 3D transposed convolution to achieve up-sampling and a softmax activation function at the last layer of the decoder to obtain segmentation probability maps, while the regression branch uses trilinear interpolation and 3D convolution to extend the feature maps and a Tanh activation function at the last layer of the decoder to obtain the signed distance maps. The process of guided consistency learning for the fuzzy maps is as follows. Firstly, three different strategies of information entropy, perceptual uncertainty, and label noise variance identification are used to obtain three fuzzy maps (high uncertainty regions) respectively. Secondly, the three fuzzy maps are fused to form a final fuzzy map through the minimization of the fusion strategy and the way of momentum update on the fuzzy map. Finally, the generated fuzzy maps are incorporated into multi-consistency losses.

In our framework, we improve the consistency learning of the mean teacher model into multi-consistency learning with a new fuzzy perception-guided target selection, which can obtain

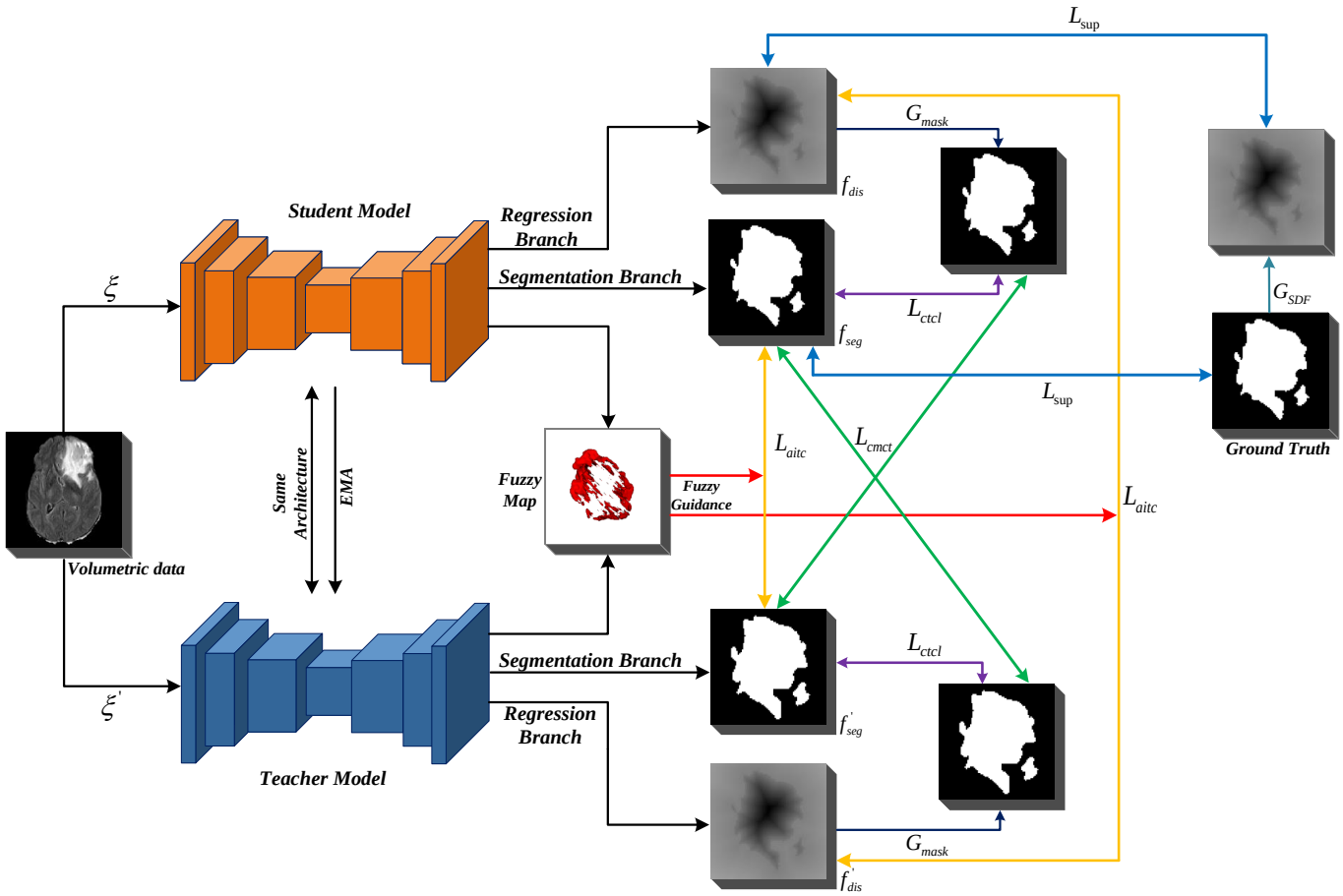

Fig. 1: The overview of a semi-supervised 3D medical image segmentation framework using multi-consistency learning with fuzzy perception-guided target selection. ξ and ξ' are different perturbations applied to the data input to the student model and the teacher model. f_{seg} and f_{dis} are segmentation probability maps and signed distance maps generated by the student model, f'_{seg} and f'_{dis} are segmentation probability maps and signed distance maps generated by the teacher model. The student model is updated via backpropagation and the teacher model is updated via the exponential moving average (EMA) of the student model weights. $\mathcal{L}_{\text{sup}}$ is the supervised loss for labeled data, $\mathcal{L}_{\text{aitc}}$ is the fuzzy mask consistency loss within the task for unlabeled data, $\mathcal{L}_{\text{ctcl}}$ is the cross-task consistency loss for unlabeled data, $\mathcal{L}_{\text{cmct}}$ is the cross-model cross-task consistency loss. The fuzzy map is used to guide the fuzzy mask consistency loss within the task $\mathcal{L}_{\text{aitc}}$ and the cross-task consistency loss $\mathcal{L}_{\text{ctcl}}$. When computing $\mathcal{L}_{\text{cmct}}$ and $\mathcal{L}_{\text{ctcl}}$, the smooth approximation $G_{\text{mask}}(\cdot)$ of the inverse transformation $G_{SDF}(\cdot)$ is applied to convert the level set function into a probability map.

better and more useful target information from fuzzy regions of a large number of unlabeled data. By combining intra-task and inter-model mutual consistency learning as well as cross-model cross-task consistency regularization, we utilize multi-task learning of geometric shape information and differential perturbation information between the teacher and student models to encourage consistent predictions on fuzzy regions. Essentially, fuzzy perception-guided target selection finds the most valuable voxel targets from fuzzy (high uncertainty) regions, and the model learns useful knowledge from these valuable voxel targets.

B. Fuzzy Estimation Guided Target Selection

In medical image analysis, the fuzziness is usually evaluated by information entropy, perceptual uncertainty and label noise variance identification to select valuable target voxels. With the consistency loss of these valuable fuzzy voxels, the network model can selectively focus on learning image information-rich regions from unlabeled data, particularly in challenging regions.

Firstly, the information entropy can be used to measure the model's prediction fuzziness for each voxel. Higher information entropy indicates that the model has an uncertain prediction at that voxel, which also means that the voxel is more challenging to accurately segment. Therefore, the student model generates a predicted probability at each voxel by V-Net [42] and uses information entropy to represent the fuzziness at the voxel:

$$f_v = -\sum_{c \in C} p_s^c \log(p_s^c), \tag{1}$$

where $c \in C$ denotes the category, p_s^c is the predicted probability from the student model at the v-th voxel, f_v is the fuzziness at the v-th voxel, and the high information entropy map $F_{HE} \in \mathbb{R}^{H \times W \times D}$ is obtained by selecting voxels with $f_v \geq H$. Since the number of voxels in the fuzzy

region decreases as the training proceeds, a dynamic threshold $H \in [0.75, 0.60]$ similar to the Gaussian ramp-down paradigm is used to control the degree of fuzziness.

Secondly, when the prediction fuzziness of models is quantified by information entropy, we further consider the high perceptual uncertainty of models in different regions, especially in cases of limited training data or when model convergence is affected, and this high perceptual uncertainty is often associated with high fuzziness. To quantify the perceptual uncertainty of models, Monte Carlo sampling is introduced as an approximation of Bayesian neural networks [43] to describe the probability distribution of models' perceptual uncertainty. Two dropout layers with a dropout rate of 0.5 are used at the last layer of the downsampling stage and the first layer of the upsampling stage of the segmentation network. During the training process, by utilizing these dropout layers to perform T random forward passes for the student model, the perceptual uncertainty for each voxel is estimated. Therefore, the predicted entropy is used to approximate the perceptual uncertainty for each voxel as follows:

$$\tilde{p}^c = \frac{1}{T}\sum_{t=1}^{T} p_s^c, \tag{2}$$

$$f_v = -\sum_{c \in C} \tilde{p}^c \log(\tilde{p}^c), \tag{3}$$

where $\tilde{p}^c$ is the average of T predicted probabilities, and the high perceptual uncertainty map $F_{HAU} \in \mathbb{R}^{H \times W \times D}$ is obtained by selecting voxels with $f_v \geq H$.

Next, we investigate the common issue of fuzziness in semi-supervised learning based on pseudo-labeling. Errors from pseudo-labels usually appear in fuzzy regions, since a model struggles to make accurate predictions with limited labeled data. Moreover, the regions with high variance exhibit clear overlap with noise in pseudo-labels. To capture errors from pseudo-labels without introducing Gaussian noise or additional branches, the noise from pseudo-labels is modeled only by the model's prediction variance. Here, the KL-divergence predicted by the student model and teacher model is used to approximate the variance:

$$Var_v = \sum_{c \in C} p_s^c \log\left(\frac{p_s^c}{p_t^c}\right), \tag{4}$$

where p_t^c is the predicted probability of the teacher model at the v-th voxel, and Var_v is the variance of the teacher and student model's prediction probabilities at the v-th voxel.

If the predictions from the two models are different at a voxel, the variance will be a large value, meaning that the voxel is located in a fuzzy and information-rich region. Label noise variance identification combines pseudo-labeling with consistency regularization, where the student model generates noisy pseudo-labels, while the teacher model recognizes label noise. The erroneous locations in pseudo-labels usually correspond to fuzzy and information-rich voxel regions. Because the location of the target voxel is to be determined in an original image rather than in the randomly perturbed image, no additional perturbation is introduced. We will use the class prototype method in unsupervised domain adaptation to identify errors in the pseudo-labels from the student model, where the features of correctly labeled voxels should be closer to their associated class prototypes. Regarding the generation of class prototypes, the masked average pooling operation [39] will be used to calculate the class prototypes of the foreground and background respectively:

$$q^{\text{obj}} = \frac{\sum_v \mathbb{I}\left[\hat{Y}_v^{'} \in C_{\text{obj}}\right] \cdot F_v^{'} \cdot P_v^{'\text{obj}}}{\sum_v \mathbb{I}\left[\hat{Y}_v^{'} \in C_{\text{obj}}\right] \cdot P_v^{'\text{obj}}}, \tag{5}$$

$$q^{\text{bg}} = \frac{\sum_v \mathbb{I}\left[\hat{Y}_v^{'} \in C_{\text{bg}}\right] \cdot F_v^{'} \cdot P_v^{'\text{bg}}}{\sum_v \mathbb{I}\left[\hat{Y}_v^{'} \in C_{\text{bg}}\right] \cdot P_v^{'\text{bg}}}, \tag{6}$$

where $F' \in \mathbb{R}^{\text{C}\times\text{H}\times\text{W}\times\text{D}}$ is the feature map generated by convolution of the penultimate layer in the teacher model after up-sampling to obtain the feature map, P'^{obj} represents the predicted probability of the teacher model on the foreground, P'^{bg} represents the predicted probability of the teacher model on the background, $\hat{Y}'$ is the label generated by the teacher model used as the mask for prototype generation, and $\mathbb{I}[\cdot]$ is the mask selected as foreground or background based on the label generated by the teacher model. The cosine similarity distance is computed between the v-th feature vector $F_v^{'}$ and the class prototypes q^{obj} and q^{bg}:

$$\cos\left(F_v^{'}, q^{\text{obj}}\right) = \frac{F_v^{'} \cdot q^{\text{obj}}}{\|F_v^{'}\|_2 \cdot \|q^{\text{obj}}\|_2}, \tag{7}$$

$$\cos\left(F_v^{'}, q^{\text{bg}}\right) = \frac{F_v^{'} \cdot q^{\text{bg}}}{\|F_v^{'}\|_2 \cdot \|q^{\text{bg}}\|_2}. \tag{8}$$

The unlabeled data X_u is input into the student model to generate noisy pseudo-labels $\hat{Y}$. If the v-th voxel $\hat{Y}_v$ generated by the student model is foreground (background) but its cosine similarity distance is closer to the prototype of background (foreground), it will be considered as a mislabeled voxel. Additionally, we incorporate the fuzzy map derived from voxels with a variance $Var_v \geq H$ as a regularization term, which prevents the class prototype method from judging voxels wrongly and reduces discontinuities in the fuzzy map, making the fuzzy map smoother. Therefore, the fuzzy map for label noise variance identification can be defined as:

$$\begin{aligned} F_{LNV} = {} & \mathbb{I}\left[\hat{Y}_v = 0\right] \cdot \mathbb{I}\left[\cos\left(F_v^{'}, q^{obj}\right) \leqslant \cos\left(F_v^{'}, q^{bg}\right)\right] \\ & + \mathbb{I}\left[\hat{Y}_v = 1\right] \cdot \mathbb{I}\left[\cos\left(F_v^{'}, q^{obj}\right) \geqslant \cos\left(F_v^{'}, q^{bg}\right)\right] \\ & + Var_v. \end{aligned} \tag{9}$$

Finally, three different fuzzy maps are obtained by the above three methods, each with slightly different selected targets and characteristics. Given the diversity and differences of the different fuzzy maps, we design and use a minimized fusion strategy and a momentum update way to update the fuzzy maps, which naturally combines multiple different fuzzy maps to generate a comprehensive and representative fuzzy map. This fusion process aims to balance and integrate the information from each fuzzy map to improve the quality

of the final fuzzy map. Specifically, this approach ensures that the fusion of different fuzzy maps produces consistent and reliable results by enhancing the stability of the fuzzy map. The fast convergence property of the momentum method helps to efficiently process the fuzzy regions of large-scale medical image data, reducing the number of iterations and thus reducing the computational cost. Solving the problem of noise in the 3D medical data voxels is another advantage of the proposed method. It effectively maintains the stability of the fuzzy map by considering the historical update direction to cope with the common noise and changes in 3D medical data voxels. Therefore, updating the fuzzy map by the momentum method can ensure stable, accurate and efficient results during the fusion process, which is defined as:

$$F = \min_{v} \left(F_{HE} + F_{HAU} + F_{LNV} \right), \tag{10}$$

$$F_t^{total} = \alpha \cdot F_{t-1}^{total} + (1-\alpha) \cdot F_t^{new}, \tag{11}$$

where t represents the current number of iterations, F_t^{new} is the new fuzzy map generated by the current iteration, F_{t-1}^{total} is the fuzzy map generated by the previous iteration, and α is the weight coefficient that controls the previous fuzzy map.

C. Fuzzy Perception-Guided Multi-Consistency Learning

In many previous methods, medical image segmentation is often regarded as a task of pixel-level classification, where the goal is to generate a segmentation probability map and assign a corresponding class label to each pixel. In addition to employing binary or multi-label masks for pixel classification, other researches focus on methods using signed distance maps. This type of methods converts a binary mask into a gray-level image, where the intensities of a pixel changes depending on the distance from the nearest boundary. The signed distance function (SDF) is a traditional technique [44] [45] used to represent object contours in a high-dimensional space. In medical image segmentation, the SDF is often utilized to describe the geometric features of targets to capture geometric distance information, which improves the segmentation performance of models. Specifically, we apply the transformation of the pixel-level segmentation map of a prediction image to a signed distance map [15] [17]. A regression branch is introduced into the traditional encoder-decoder architecture to generate signed distance maps while working in parallel with the traditional segmentation branch for generating segmentation probability maps. Task-level differences between the two branches lead to model perturbations and encourage the model to learn different representations of segmentation targets from different perspectives. The segmentation branch and regression branch provide supervised information for labeled data. Therefore, the supervised loss can be defined as:

$$\begin{aligned} \mathcal{L}_{sup} = \sum_{x_i, y_i \in D_l} & \mathcal{L}_{\text{dice}} \left(f_{seg}\left(x_i\right), y_i \right) + \mathcal{L}_{bce} \left(f_{seg}\left(x_i\right), y_i \right) \\ & + \mathcal{L}_{dis} \left(f_{dis}\left(x_i\right), G_{SDF}\left(y_i\right) \right), \end{aligned} \tag{12}$$

where $\mathcal{L}_{dice}(\cdot)$ represents the commonly used Dice loss, $\mathcal{L}_{bce}(\cdot)$ represents the Binary Cross Entropy loss, $\mathcal{L}_{dis}(\cdot)$ represents the Mean Squared Error loss, $f_{seg}(\cdot)$ represents the segmentation network model, $f_{dis}(\cdot)$ represents the regression network model, and $G_{SDF}(\cdot)$ represents the signed distance transformation function.

For semi-supervised medical image segmentation, the improvement in model performance comes from generating supervised signals and obtaining unsupervised knowledge from unlabeled data via an unsupervised loss function. The main semi-supervised segmentation methods often use the mean teacher model as its framework, which consists of two models, namely the student model and the teacher model, which have the same network structure but different parameters. During the training process, the network parameters of the teacher model are updated as the exponential moving average (EMA) of the parameters of the student model [10] [12] [32]. Unlike the classical Mean Teacher (MT) [10] that computes all voxels, or its variant Uncertainty Aware Mean Teacher (UA-MT) [12] that only computes reliable regions for consistency learning, our model will compute the consistency loss on the finally generated fuzzy regions because they have higher learning value. Specifically, we redesign the consistency loss as an intra-task fuzzy mask mean squared error loss:

$$\begin{aligned} \mathcal{L}_{aitc}\left(f, f'; F\right) = & \beta \frac{\sum_v \left[F_v \sum_c \left(\left\| f_{\text{seg}} - f'_{\text{seg}} \right\|^2 \right) \right]}{\sum_v F_v} \\ & + (1-\beta) \frac{\sum_v \left[F_v \sum_c \left(\left\| f_{\text{dis}} - f'_{\text{dis}} \right\|^2 \right) \right]}{\sum_v F_v}, \end{aligned} \tag{13}$$

where F_v is the fuzziness at the v-th voxel, $(f_{\text{seg}}, f_{\text{dis}})$ represents the outputs of the segmentation branch and the regression branch of the student model at the v-th voxel for each class $c \in C$, $\left(f'_{\text{seg}}, f'_{\text{dis}}\right)$ represents the outputs of the segmentation branch and the regression branch of the teacher model at the v-th voxel for each class $c \in C$, and β is the weight coefficient that achieves a balance between the segmentation and regression tasks.

For the same input data in different tasks, their predictions should keep consistency when mapped into the same predefined space. In order to efficiently utilize unlabeled data, we perform cross-task consistency learning on fuzzy regions, aiming to ensure that the outputs from the segmentation branch and the regression branch remain consistent. We use the smooth approximation method of the inverse transformation of the SDF to convert the output of the distance map back to the binary segmentation output [15]. The cross-task consistency learning method helps the model to learn the correlations between different tasks comprehensively and establishes consistency between the outputs of two tasks. Therefore, cross-task consistency loss can be defined as:

$$\begin{aligned} \mathcal{L}_{ctcl}\left(f, f'; F\right) = & \beta \frac{\sum_v \left[F_v \sum_c \left(\left\| f_{\text{seg}}(x) - G_{\text{mask}}\left(f_{\text{dis}}(x)\right) \right\|^2 \right) \right]}{\sum_v F_v} \\ & + (1-\beta) \frac{\sum_v \left[F_v \sum_c \left(\left\| f'_{\text{seg}}(x) - G_{\text{mask}}\left(f'_{\text{dis}}(x)\right) \right\|^2 \right) \right]}{\sum_v F_v}, \end{aligned} \tag{14}$$

where $G_{\text{mask}}(\cdot)$ is the smooth approximation to the inverse

transformation of the SDF.

The task differences between the two branches can cause perturbations in the model, which means that the different tasks are able to guide the model in their respective ways, enabling it to learn the segmentation target from multiple perspectives, thus obtaining more diverse and comprehensive target representations. This cross-model cross-task consistency learning can help the model understand and learn the different characteristics and changes of targets. It's loss function is defined as:

$$\begin{aligned}\mathcal{L}_{cmct}\left(f, f'\right) = &\left\| f'_{\text{seg}}\left(x\right) - G_{\text{mask}}\left(f_{\text{dis}}\left(x\right)\right)\right\|^2 \\ &+ \left\| f_{\text{seg}}\left(x\right) - G_{\text{mask}}\left(f'_{\text{dis}}\left(x\right)\right)\right\|^2. \end{aligned} \tag{15}$$

Cross-model cross-task consistency learning further enhances the model's representation abilities, allowing it to deeply understand the relationship between different tasks and different models. It not only helps models improve the segmentation performance, but also can better cope with the diversity and complexity of 3D data voxels in medical image segmentation tasks.

D. Overall Training Process

Our proposed framework can be trained by minimizing the weighted sum of the supervised segmentation loss $\mathcal{L}_{sup}$, the intra-task consistency loss $\mathcal{L}_{aitc}$, the cross-task consistency loss $\mathcal{L}_{ctcl}$ and the cross-model cross-task consistency loss $\mathcal{L}_{cmct}$. The student model utilizes the supervised segmentation loss $\mathcal{L}_{sup}$ to learn from the labeled data. At the same time, the student model and the teacher model learn more challenging information from the unlabeled data under the fuzzy perception-guided target selection. Therefore, the framework can be formulated as minimizing the following function:

$$\begin{aligned}\min_{\theta} &\mathcal{L}_{sup}\left(\theta; \mathcal{D}_L\right) + \lambda\left(\mathcal{L}_{aitc}\left(\theta, \theta'; \mathcal{D}\right)\right. \\ &\left. + \mathcal{L}_{ctcl}\left(\theta, \theta'; \mathcal{D}\right) + \mathcal{L}_{cmct}\left(\theta, \theta'; \mathcal{D}\right)\right), \end{aligned} \tag{16}$$

where λ represents the rising weighting coefficient. According to [46] [47], we use the Gaussian ramp function $\lambda\left(t\right) = e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$ to control the balance between the supervised and semi-supervised losses, mitigating the interference of the consistency loss in the early training stage, where t represents the current step of the iteration and t_{max} represents the maximum training step. The training procedure of our proposed semi-supervised medical image segmentation framework can be described by Algorithm 1.

Algorithm 1 Training procedure of multi-consistency learning with fuzzy perception-guided target selection

Input: A batch of x_l, y_l from labeled dataset D_l and x_u from unlabeled dataset D_u.

Output: Trained network $\mathcal{N}$ with θ

f_{seg} and f_{dis} represent the output predictions of segmentation branch and regression branch to generate segmentation probabilistic maps and signed distance maps, respectively
for minibatch $\{(x_k, y_k)\}_{k=1}^{B} \subset \left(D^l \cup D^u\right)$ **do**
  Generate output segmentation maps f_{seg}, output distance maps f_{dis} and the final fuzzy map F formed by fusion
  Calculate supervised segmentation loss $\mathcal{L}_{sup}$ as Eq.(12)
  Calculate intra-task consistency losses $\mathcal{L}_{aitc}$ as Eq.(13)
  Calculate cross-task consistency losses $\mathcal{L}_{ctcl}$ as Eq.(14)
  Calculate cross-model cross-task consistency losses $\mathcal{L}_{cmct}$ as Eq.(15)
  Update the student model's weights θ with $\mathcal{L} = \mathcal{L}_{sup} + \lambda\left(\mathcal{L}_{aitc} + \mathcal{L}_{ctcl} + \mathcal{L}_{cmct}\right)$
  Update the teacher model's weights with exponential moving average (EMA) of the student model's weights
end for
return θ

IV. Experiments and Analysis

A. Datasets

To evaluate our framework, we conducted a comprehensive evaluation of two different types of medical image datasets, including the 3D left atrium magnetic resonance (MR) image scans [48] and the 3D brain tumor magnetic resonance (MR) image scans [49].

• **3D left atrium segmentation MR dataset:** The left atrium (LA) dataset originates from the 2018 Atrial Segmentation Challenge and includes 100 sets of 3D gadolinium-enhanced MR images, which contain 3D binary masks representing the left atrial cavity. The original isotropic resolution is $0.625 \times 0.625 \times 0.625\text{mm}^3$. Following [12] [15] [50] [51], we split the 100 scans into 80 scans for training and 20 scans for testing. For the 80 training scans, 2.5%/2, 5%/4 and 10%/8 scans are used as labeled data, and the rest of the scans are employed as unlabeled data.

• **3D brain tumor segmentation MR dataset:** The brain tumor (BraTS) dataset comes from the 2019 Multimodal Brain Tumor Segmentation Challenge. It is mainly used to study and evaluate the performance of brain tumor segmentation algorithms. It contains 335 scans usually including T1-weighted, T2-weighted, and contrast-enhanced T1-weighted sequences, with the same resolution of $1 \times 1 \times 1\text{mm}^3$. Following [13] [50], we randomly chose 250 scans for training, 25 scans for validation, and 60 scans for testing. For the 250 training scans, 5%/12 and 10%/25 scans are used as labeled data, and the rest of the scans are employed as unlabeled data.

B. Implementing Details and Evaluation Metrics

Implementing Details: All algorithms in our experiments were implemented on a server with NVIDIA GeForce RTX 3090 24GB, Ubuntu 18.04, and PyTorch 1.7. The batch size is 4, consisting of 2 labeled images and 2 unlabeled images in each mini-batch. We employed V-Net [42] as the backbone for all experiments to achieve a fair comparison. The framework is trained for 6,000 iterations using the SGD optimizer (weight decay = 0.0001, momentum = 0.9). The learning rate is initialized as 0.01 and decayed by 0.1 every 2,500 iterations. We randomly cropped $112 \times 112 \times 80$ on the LA dataset [48] or $96 \times 96 \times 96$ on the BraTS dataset [49] sub-volume as the input. According to [12] [15] [51], data augmentation,

including randomly flip and rotation, is applied to avoid overfitting. For the testing phase, we employed the teacher model. This choice was made due to the teacher model's superior stability and robustness in semi-supervised learning scenarios. By aggregating the results from multiple trainings of the student model, the teacher model effectively reduces noise and uncertainty, leading to higher prediction accuracy compared to the student model.

Evaluation Metrics: According to [50], we adopted four metrics for a comprehensive evaluation, i.e., Dice similarity coefficient (Dice), Jaccard index (Jaccard), Average surface distance (ASD) and 95% Hausdorff distance (95HD), which can be defined as:

$$Dice\left(V_{pred}, V_{gt}\right) = \frac{2|V_{pred} \cap V_{gt}|}{|V_{pred}| + |V_{gt}|}, \tag{17}$$

$$Jaccard\left(V_{pred}, V_{gt}\right) = \frac{|V_{pred} \cap V_{gt}|}{|V_{pred} \cup V_{gt}|}, \tag{18}$$

$$ASD(A,B) = \frac{1}{2}\left(\frac{\sum_{a\in A} \min_{b\in B} d(a,b)}{\sum_{a\in A} 1} + \frac{\sum_{b\in B} \min_{a\in A} d(a,b)}{\sum_{b\in B} 1}\right), \tag{19}$$

$$HD(A,B) = max[\sup_{a\in A} \inf_{b\in B} d(a,b), \sup_{b\in B} \inf_{a\in A} d(a,b)], \tag{20}$$

where V_{pred} is the set of voxels in the predicted pixel-level probability map from the segmentation network, and V_{gt} is the set of voxels in a pixel-level probability map from the ground truth. A and B represent two sets of contour points, and $d(a,b)$ denotes the Euclidean distance between the two points a and b.

C. Comparisons with SOTA Methods

In this section, we conducted a large number of comparative experiments on the LA dataset [48] and BraTS dataset [49] to verify the superiority and effectiveness of the proposed framework under three general semi-supervised experimental settings. To fairly evaluate the various methods, we used the same V-Net [42] as the backbone network, as well as the same experimental platform and hyperparameter settings in all the comparison experiments. In addition, we used 2.5%, 5%, and 10% labeled data as training datasets on the LA dataset [44] to demonstrate the segmentation performance obtained by the V-Net [42] network under different settings, as shown in Table I, respectively. On the BraTS dataset [49], using 5% and 10% labeled data as a training dataset, the segmentation performance obtained by the V-Net [42] network under different settings is demonstrated, as shown in Table II, respectively.

Comparison on LA dataset: In order to demonstrate the effectiveness of our proposed framework, a comprehensive comparison with existing methods is performed on the LA dataset. We evaluate our framework by comparing it with several recent state-of-the-art semi-supervised segmentation methods, including Mean Teacher (MT) [10], Uncertainty-aware Mean Teacher (UA-MT) [12], Shape-aware Semi-Supervised Network (SASSNet) [51], Dual-Task Consistency (DTC) [15], Uncertainty Rectified Pyramid Consistency (URPC) [14], Mutual Consistency Learning (MC-Net) [16], Adversarial Consistency Learning (ASE-Net) [34], Dual-stage Semi-supervised Pre-training Approach (DSTP) [52], Dual-contrastive Dual-consistency Dual-transformer (3D-ViT) [53], Bidirectional Copy-Paste (BCP) [54] and Mutual Correction Framework (MCF) [35].

TABLE I: Quantitative comparisons of different methods on the LA dataset by utilizing 2.5%, 5%, and 10% labeled data of training set. The best values are in bold. The number in red indicates the improvement of our method compared with the best of the other methods.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
V-Net	4	0	52.55	39.60	9.87	47.05
V-Net	8	0	79.99	68.12	5.48	21.11
V-Net	80	0	91.14	83.82	1.52	5.75
MT[NeurIPS'17]	2	78	72.78	58.88	3.65	34.16
UA-MT[MICCAI'19]	2	78	74.63	60.79	3.39	35.98
SASSNet[MICCAI'20]	2	78	73.35	59.12	3.47	36.52
DTC[AAAI'21]	2	78	73.24	58.54	3.52	35.64
URPC[MIA'22]	2	78	74.87	60.82	3.54	34.63
MC-Net[MIA'22]	2	78	75.73	60.94	3.22	32.18
ASE-Net[TMI'22]	2	78	76.94	63.58	3.03	30.04
DSTP[TAI'23]	2	78	75.76	63.45	3.16	31.59
3D-ViT[ICCV'23]	2	78	76.85	64.09	3.01	32.13
BCP[CVPR'23]	2	78	77.19	65.17	3.18	29.86
MCF[CVPR'23]	2	78	76.88	63.48	3.90	30.11
Ours	2	78	**78.98**↑1.79	**66.65**	**2.88**	**28.14**
MT[NeurIPS'17]	4	76	80.67	68.85	4.03	15.24
UA-MT[MICCAI'19]	4	76	82.26	70.98	3.82	13.71
SASSNet[MICCAI'20]	4	76	81.60	69.63	3.58	16.16
DTC[AAAI'21]	4	76	81.25	69.33	3.99	14.90
URPC[MIA'22]	4	76	82.48	71.35	3.65	14.65
MC-Net[MIA'22]	4	76	83.59	72.36	2.70	14.07
ASE-Net[TMI'22]	4	76	83.33	71.79	4.33	15.70
DSTP[TAI'23]	4	76	82.15	70.76	4.10	16.74
3D-ViT[ICCV'23]	4	76	82.47	71.38	3.86	15.12
BCP[CVPR'23]	4	76	84.50	72.71	2.56	**12.96**
MCF[CVPR'23]	4	76	84.39	73.17	3.31	14.85
Ours	4	76	**85.85**↑1.35	**75.47**	**2.32**	14.76
MT[NeurIPS'17]	8	72	84.24	73.26	2.71	19.40
UA-MT[MICCAI'19]	8	72	84.25	73.48	3.36	13.48
SASSNet[MICCAI'20]	8	72	86.81	76.92	3.94	12.54
DTC[AAAI'21]	8	72	86.57	76.55	3.74	14.47
URPC[MIA'22]	8	72	85.02	75.98	2.96	15.21
MC-Net[MIA'22]	8	72	87.71	78.31	2.18	9.36
ASE-Net[TMI'22]	8	72	87.83	78.45	2.17	9.86
DSTP[TAI'23]	8	72	86.74	77.19	2.27	8.67
3D-ViT[ICCV'23]	8	72	87.62	78.12	2.66	8.92
BCP[CVPR'23]	8	72	87.91	78.58	2.10	8.99
MCF[CVPR'23]	8	72	86.63	77.01	2.95	8.97
Ours	8	72	**89.11**↑1.2	**80.48**	**2.01**	**8.54**

We used the same V-Net [42] backbone in all these methods for a fair comparison. Table I shows the results of the comparison of different methods on the left atrial test set in the case of utilizing 2.5%, 5% and 10% labeled data. It can be seen that UA-MT [12] improves Dice by 4.26%, compared to V-Net [42] when utilizing the same percentage of labeled data, suggesting that UA-MT effectively uses unlabeled data for better segmentation results by using consistency learning for reliable labels only. By leveraging the consistency of the fuzzy

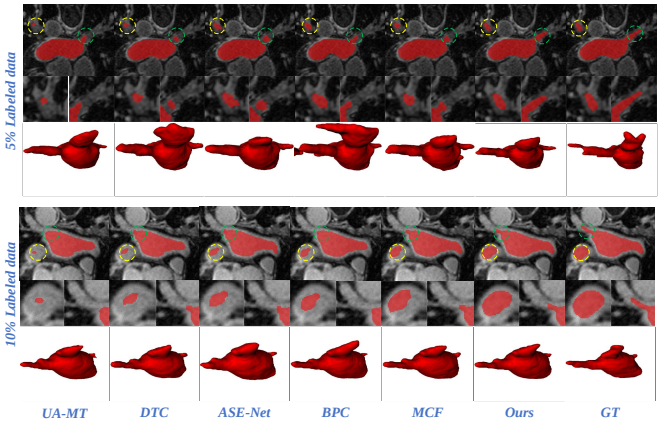

Fig. 2: Visualization results of different methods on the LA dataset by utilizing 5% and 10% of the labeled data in the training set, respectively.

regions of a large amount of unlabeled data during training, the proposed framework obtains significant performance improvements (Dice from 52.55% to 85.85%, Jaccard from 39.60% to 75.47%, ASD from 9.87 to 2.32, and 95HD from 47.05 to 14.76 under 5% labeled data). Compared to the most recent semi-supervised method Bidirectional Copy-Paste (BCP) [54], it reduces the distribution gap between labeled and unlabeled data by enforcing the invariance of predictions under different distributions, but it does not fully exploit the shape constraints and boundary-level distance information of most fuzzy regions from the unlabeled data and does not set a threshold to focus on region-level consistency learning, whereas our framework focuses on local region learning in the unlabeled data and puts more emphasis on fuzzy regions of consistency. Therefore, compared with it, the value of Dice increases by 1.35% under 5% labeled data and by 1.2% under 10% labeled data. In addition, Fig. 2 shows the segmentation results provided by our framework, it is clear that our framework provides better segmentation results than other methods used for comparison.

Comparison of BraTS dataset: To further validate our proposed framework, thirteen state-of-the-art methods are compared on the BraTS dataset [49], including Mean Teacher (MT) [10], Uncertainty-aware Mean Teacher (UA-MT) [12],Shape-aware Semi-Supervised Network (SASSNet) [51], Dual-Task Consistency (DTC) [15], Uncertainty- Rectified Pyramid Consistency (URPC) [14], Smoothness and Class Separation Consistency Learning (SS-Net) [55], Mutual Consistency Learning (MC-Net) [16], Adversarial Consistency Learning (ASE-Net) [34], Dual-stage Semi-supervised Pre-training Approach (DSTP) [52], Dual-contrastive Dual-consistency Dual-transformer (3D-ViT) [53], Bidirectional Copy-Paste (BCP) [54] and Mutual Correction Framework (MCF) [35]. Table II shows the comparative results of different methods on the brain tumor test set using 5% and 10% labeled data. The segmentation results on the BraTS dataset [49] are shown in Fig. 3. Our results demonstrate a closer alignment with the ground truth. Our predicted segmentation results have smoother transitions at the boundaries and reduce misclassification cases due to boundary fuzziness, which allows for better capturing of fuzzy boundaries and provides more accurate segmentation results.

The results presented in Figures 2 and 3 show that the performance of the semi-supervised approach is still insufficient in some critical regions (e.g., edges and small lesions). For example, in the edge region, the segmentation results of the model may appear blurred or inaccurate, which may lead to misjudgments by clinicians when determining the tumor boundary. In addition, for the identification of small lesions, the sensitivity of the model is low, which may lead to some small lesions being overlooked. In practical clinical applications, these limitations may have a significant impact on diagnosis and treatment. For example, for tumor resection surgery, accurate tumor boundary determination is crucial. If the model is not accurate enough in edge detection, it may lead to incomplete tumor resection or mistakenly cut normal tissues, affecting the surgical outcome and patient prognosis. In addition, in early cancer screening, identifying small lesions is crucial for early diagnosis and treatment. If the model misses these small lesions, it may lead to delayed treatment and affect the survival rate of patients. Consequently, in actual clinical practice, doctors usually combine multiple images (e.g., CT, MRI, etc.) to make a comprehensive judgment. If the model only relies on a single modality image for training and prediction, it may not be able to fully utilize other imaging information, thus affecting the accuracy of diagnosis.

TABLE II: Quantitative comparisons of different methods on the BRATS dataset by utilizing 5% and 10% labeled data of training set. The best values are in bold. The number in red indicates the improvement of our method compared with the best of the other methods.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
V-Net	12	0	70.28	60.42	2.82	38.44
V-Net	25	0	74.43	61.86	2.79	37.11
V-Net	250	0	86.95	78.03	1.75	6.56
MT[NeurIPS'17]	12	238	80.31	70.37	2.83	11.69
UA-MT[MICCAI'19]	12	238	77.25	63.56	3.80	17.56
SASSNet[MICCAI'20]	12	238	76.17	66.43	3.32	13.09
DTC[AAAI'21]	12	238	74.21	64.89	3.16	13.54
URPC[MIA'22]	12	238	78.74	68.20	4.51	14.43
SS-Net[MICCAI'22]	12	238	78.03	68.11	2.76	13.70
MC-Net[MIA'22]	12	238	78.69	68.38	4.49	13.44
ASE-Net[TMI'22]	12	238	78.53	68.03	3.57	15.99
DSTP[TAI'23]	12	238	77.26	66.82	3.06	14.90
3D-ViT[ICCV'23]	12	238	77.66	67.08	3.05	14.93
BCP[CVPR'23]	12	238	79.27	68.69	**2.25**	**12.25**
MCF[CVPR'23]	12	238	78.67	67.94	2.89	12.59
Ours	12	238	**80.09**↑0.82	**70.07**	3.33	13.78
MT[NeurIPS'17]	25	225	81.21	70.83	2.45	14.72
UA-MT[MICCAI'19]	25	225	80.85	70.32	2.57	14.61
SASSNet[MICCAI'20]	25	225	79.19	68.80	6.67	16.36
DTC[AAAI'21]	25	225	81.75	71.63	2.56	15.73
URPC[MIA'22]	25	225	82.59	72.11	3.72	13.88
SS-Net[MICCAI'22]	25	225	82.00	71.82	1.98	10.68
MC-Net[MIA'22]	25	225	79.63	70.10	2.45	12.28
ASE-Net[TMI'22]	25	225	83.24	73.43	2.15	10.32
DSTP[TAI'23]	25	225	83.13	72.77	2.02	12.45
3D-ViT[ICCV'23]	25	225	82.56	72.62	2.33	13.25
BCP[CVPR'23]	25	225	83.31	73.63	2.23	10.86
MCF[CVPR'23]	25	225	83.28	73.99	2.93	11.29
Ours	25	225	**84.17**↑0.81	**74.28**	**1.92**	**9.60**

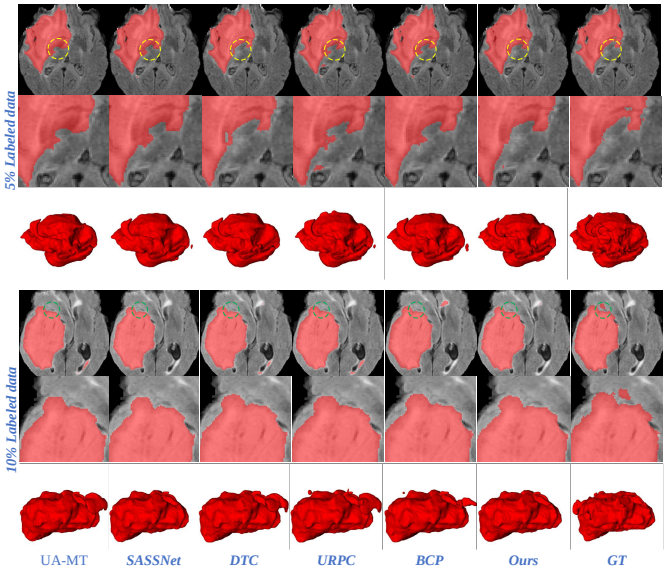

Fig. 3: Visualization results of different methods on the BraTS dataset by utilizing 5% and 10% of the labeled data in the training set, respectively.

D. Ablation Experiments

We conducted ablation experiments on the LA dataset [48] using 8 labeled and 72 unlabeled images to check the effectiveness of each component, as shown in Table III. The results show that when only a small amount of labeled data is available for training, the performance can be improved by mining meaningful latent information from the 72 unlabeled images. Specifically, the model achieves the best performance when fuzzy map, supervised loss, and consistency learning guided by fuzzy perception are introduced. In particular, the intra-task consistency loss and the cross-task consistency loss guided by fuzzy perception, as well as the cross-model cross-task consistency loss, play an important role in performance improvement because they help to improve the model's utilization of unlabeled data and enhance the model's ability to perceive fuzziness in the data. Experimental results (3), (4), (5), (6), (7), and (8) show that consistency learning on fuzzy regions can more fully exploit the valuable information from unlabeled data and significantly improve the segmentation accuracy. Moreover, experimental results (9), (10), and ours show that a combination of intra-task consistency learning and cross-task consistency learning on fuzzy regions helps the model learn the features of different tasks from unlabeled data and emphasize the importance of fuzzy regions, resulting in significant performance improvement. In addition, experimental results (8) and ours demonstrate the effectiveness of the proposed cross-model cross-task consistency learning, which enables the model to learn segmentation targets from multiple perspectives and obtain more diverse and comprehensive target representations. All the above strategies emphasize small branches or edges by fully mining the most valuable voxel targets of fuzzy regions from unlabeled data, which is meaningful guidance for challenging regions.

According to Table IV, the model performs better in various metrics when the value of α is 0.9. One major cause is that it better balances the effects between the old and new fuzzy maps and takes into account most of the information of the previous fuzzy maps, which helps to ensure the consistency of the fuzzy maps during the iterative stage. This consistency is crucial for producing consistent and reliable final results when fusing multiple fuzzy maps, and more historical fuzzy map information is retained, which has a positive impact on combating noise and data changes.

TABLE III: Comparison of ablation experiments on the LA dataset by utilizing 10% labeled data of training set. The best values are in bold.

Method	Fuzzy Map	Supervised Loss		Consistency Loss			Dice(%)
		$\mathcal{L}_{seg}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{aitc}$	$\mathcal{L}_{ctcl}$	$\mathcal{L}_{cmct}$	
Scheme.1	-	✓	-	-	-	-	83.08
Scheme.2	-	✓	✓	-	-	-	84.29
Scheme.3	-	✓	✓	✓	-	-	85.96
Scheme.4	✓	✓	✓	✓	-	-	87.42
Scheme.5	-	✓	✓	-	✓	-	86.48
Scheme.6	✓	✓	✓	-	✓	-	87.65
Scheme.7	-	✓	✓	✓	✓	-	87.38
Scheme.8	✓	✓	✓	✓	✓	-	88.17
Scheme.9	✓	✓	✓	✓	-	✓	88.25
Scheme.10	✓	✓	✓	-	✓	✓	88.37
Ours	✓	✓	✓	✓	✓	✓	**89.11**

TABLE IV: Comparison of different balance weight α used for the old fuzzy map and new fuzzy map on the BRATS dataset by utilizing 5% labeled data of training set. The best values are in bold.

α	Metrics			
	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
0	78.62	68.36	3.48	14.19
0.5	79.12	69.20	3.37	13.89
0.7	80.04	69.79	**2.98**	13.81
0.9	**80.09**	**70.07**	3.33	**13.78**

Since the classical consistency learning process is based on segmentation prediction, we used balanced weights β to control the consistency learning between the segmentation and regression tasks. We performed experiments to assess the selection of β within our consistency learning framework, with the outcomes presented in Table V. When the value of β sets 1 or 0, the model's performance diminishes as it relies only on the segmentation or regression branch while ignoring the other branch. It can be found that the framework achieves the best performance when β = 0.75. Therefore, we set β = 0.75 for our model in the experiments.

V. Discussion

Threshold Selection (Comparison of Fuzzy and Reliable)

In the medical image segmentation task, it is crucial for the region selection of targets, so we adopted a dynamic threshold [0.60, 0.75] with a Gaussian ramp-down paradigm to divide the fuzzy value, thereby selecting the target region for consistency learning. Specifically, when the fuzzy value of a voxel is less and equal to the given threshold, it is

TABLE V: Comparison of different balance weight β used for segmentation task and regression task on the BRATS dataset by utilizing 10% labeled data of training set. The best values are in bold.

β	Metrics			
	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
0	83.47	74.05	2.16	11.83
0.25	83.98	74.14	2.15	11.94
0.5	84.01	73.42	2.13	11.37
0.75	**84.17**	**74.28**	**1.92**	**11.29**
1	83.68	73.71	2.10	11.88

classified as a reliable region, and when the fuzzy value is greater and equal to the given threshold, it is labeled as a fuzzy region. In our study, we compared the effects of using common traditional fixed thresholds (e.g., 0.5, 0.6 and 0.7) and dynamic thresholds, as shown in Table VI. However, experiments have proven that the application of a dynamic threshold achieves better results for medical image segmentation. The dynamic threshold takes into account the fact that the fuzzy region will reduce with the process of training, which makes the selection of the threshold more flexible and adaptable. It can better reflect the distribution of fuzziness in different periods, and can also be adjusted according to the characteristics of data and the learning ability of the model, so that the fuzzy boundaries can be obtained, which helps to distinguish fuzzy regions and reliable regions more precisely.

TABLE VI: Comparison of fixed thresholds and dynamic thresholds on the LA dataset by utilizing 5% labeled data of training set. The best values are in bold.

Threshold	Metrics			
	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
0.5	83.08	71.31	3.76	19.15
0.6	83.77	72.34	3.31	23.49
0.7	83.97	72.65	2.98	20.92
Ours	**85.85**	**75.47**	**2.32**	**14.76**

In addition, we further studied the effect of selecting fuzzy regions or selecting reliable regions for model training. The results show that selecting fuzzy regions for model training can lead to more accurate segmentation results, as shown in Table VII. This suggests that the learning of models that focus too much on fuzzy regions is crucial for the success of the medical image segmentation task, while focusing on reliable regions may lead to the neglect of fuzziness and reduce the segmentation performance, as shown in Fig. 4. Error-prone fuzzy regions contain richer information and more valuable clues in unlabeled data. Therefore, individually selecting fuzzy regions for model training under dynamic thresholds proves to be a more effective strategy, which helps the model to learn and process the most challenging and attention-demanding regions in the medical images, thus enhancing the performance and reliability of the segmentation model.

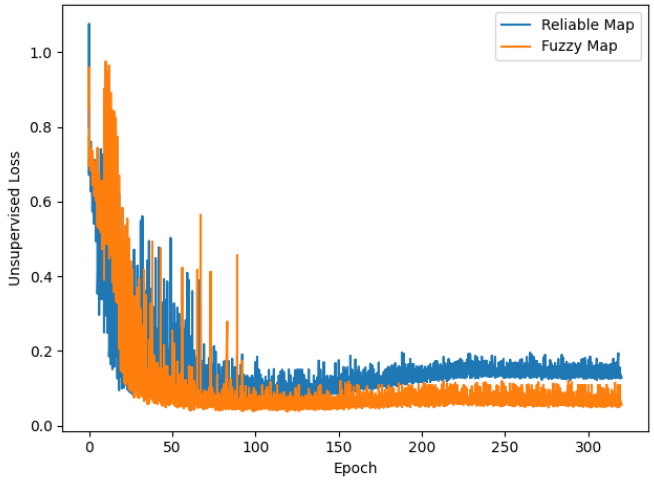

Fig. 4: The unsupervised loss curve of comparison between fuzzy map and reliable map on the LA dataset by utilizing 10% of the labeled data in the training set.

TABLE VII: Comparison of reliable map and fuzzy map on the LA dataset by utilizing 10% labeled data of training set. The best values are in bold.

Method	Metrics			
	Dice(%)↑	Jaccard(%)↑	ASD(voxel)↓	95HD(voxel)↓
Reliable Map	87.38	77.81	2.59	15.08
Fuzzy Map	**89.11**	**80.48**	**2.01**	**8.54**

VI. Conclusion

Although existing deep learning-based medical image segmentation methods have achieved great success, they are limited by the requirement for large amounts of labeled data. Semi-supervised medical image segmentation, which encourages segmentation models to utilize more easily collected unlabeled data, demonstrates potential in overcoming this limitation. In this study, we have proposed a new semi-supervised medical image segmentation with multi-consistency learning for fuzzy perception-guided target selection. First, the framework introduces fuzzy perception-guided target selection to identify the most challenging targets in fuzzy regions of unlabeled data, which allows the model to enhance the learning of the representation for these valuable regions and thus obtain a fuzzy map. Then, the fuzzy map is incorporated into intra-task and inter-model mutual consistency learning as well as cross-model cross-task consistency regularization to further improve segmentation accuracy. Extensive experiments on two challenging public datasets demonstrate that the proposed framework provides a general and effective solution for achieving high-quality 3D medical image segmentation compared with other methods using small amounts of labeled data.

Acknowledgment

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this article.

REFERENCES

[1] P. -H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 7, no. 6, pp. 545-569, 2023.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[3] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[4] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision*. 2022, pp. 205–218.

[5] A. Peláez-Vegas, P. Mesejo, and J. Luengo, "A survey on semi-supervised semantic segmentation," arXiv preprint arXiv:2302.09899, 2023.

[6] W. Lei et al., "One-shot weakly-supervised segmentation in 3D medical images," *IEEE Transactions on Medical Imaging*, 2023.

[7] Y. Feng, Y. Wang, H. Li, M. Qu, and J. Yang, "Learning what and where to segment: A new perspective on medical image few-shot segmentation," *Medical Image Analysis*, vol. 87, p. 102834, 2023.

[8] S. Y. Seo et al., "Unified deep learning-based mouse brain MR segmentation: template-based individual brain positron emission tomography volumes-of-interest generation without spatial normalization in mouse Alzheimer model," *Frontiers in Aging Neuroscience*, vol. 14, p. 807903, 2022.

[9] G. Wang et al., "DeepIGeoS: a deep interactive geodesic framework for medical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1559–1572, 2018.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of International Conference on Neural Information Processing Systems*. 2017, pp. 1195–1204.

[11] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2613–2622.

[12] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 605–613.

[13] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12674–12684.

[14] X. Luo et al., "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.

[15] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021, pp. 8801–8809.

[16] Y. Wu et al., "Mutual consistency learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 81, p. 102530, 2022.

[17] Y. Zhang and J. Zhang, "Dual-task mutual learning for semi-supervised medical image segmentation," in *The 4th Chinese Conference on Pattern Recognition and Computer Vision*. 2021, pp. 548–559.

[18] Y. Zhang, R. Jiao, Q. Liao, D. Li, and J. Zhang, "Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation," *Artificial Intelligence in Medicine*, vol. 138, p. 102476, 2023.

[19] T. Lei et al., "Semi-supervised 3D medical image segmentation using shape-guided dual consistency learning," in *IEEE International Conference on Multimedia and Expo*. 2022, pp. 01–06.

[20] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proceedings of International Conference on Neural Information Processing Systems*. 2020, vol. 33, pp. 596–608.

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

[22] H. Zeng, K. Zou, Z. Chen, R. Zheng, and H. Fu, "Reliable Source Approximation: Source-Free Unsupervised Domain Adaptation for Vestibular Schwannoma MRI Segmentation," arXiv preprint arXiv:2405.16102, 2024. (Early accepted by MICCAI 2024)

[23] L. Qiu, J. Cheng, H. Gao, W. Xiong, and H. Ren, "Federated semi-supervised learning for medical image segmentation via pseudo-label denoising," *IEEE Journal of Biomedical and Health Informatics*, 2023.

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

[25] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 83, p. 102656, 2023.

[26] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 87, p. 102792, 2023.

[27] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 23924–23935.

[28] A. Lou, K. Tawfik, X. Yao, Z. Liu, and J. Noble, "Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[29] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "CAT: Constrained adversarial training for anatomically-plausible semi-supervised segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[30] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "PEFAT: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15671–15680.

[31] C. Xu et al., "BMAnet: Boundary mining with adversarial learning for semi-supervised 2D myocardial infarction segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 87–96, 2022.

[32] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," arXiv preprint arXiv:1610.02242 , 2016.

[33] T. Lei, H. Liu, Y. Wan, C. Li, Y. Xia, and A. K. Nandi, "Shape-guided dual consistency semi-supervised learning framework for 3D medical image segmentation," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 7, no. 7, pp. 719-731, 2023.

[34] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1265-1277, 2023.

[35] Y. Wang, B. Xiao, X. Bi, W. Li, and X. Gao, "MCF: Mutual correction framework for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15651–15660.

[36] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019 pp. 48–56.

[37] Z. Xu et al., "Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 88, p. 102880, 2023.

[38] S. Czolbe, K. Arnavaz, O. Krause, and A. Feragen, "Is segmentation uncertainty useful?," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 715–726.

[39] C. Chen, Q. Liu, Y. Jin, Q. Dou, and P.-A. Heng, "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 225–235.

[40] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1106–1120, 2021.

[41] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12414–12424.

[42] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *The 4th International Conference on 3D Vision*. 2016, pp. 565–571.

[43] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Proceedings of International Conference on Neural Information Processing Systems*. 2017, vol. 30.

[44] Y. Wang et al., "Deep distance transform for tubular structure segmentation in ct scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3833–3842.

[45] J. Ma et al., "How distance transform maps boost segmentation CNNs: an empirical study," *Medical Imaging with Deep Learning*, pp. 479–492, 2020.

[46] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semi-supervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[47] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 431–443, 2020.

[48] Z. Xiong et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[49] S.S.Bakas, "Brats miccai brain tumor dataset," 2020. Available: https://dx.doi.org/10.21227/hdtd-5j88

[50] Z. Xu et al., "All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3174–3184, 2022.

[51] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 552–561.

[52] R. C. Aralikatti, S. Pawan, and J. Rajan, "A dual-stage semi-supervised pre-training approach for medical image segmentation," *IEEE Transactions on Artificial Intelligence*, 2023.

[53] Z. Wang and C. Ma, "Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 870–879.

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

[55] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 34–43.