Unified Feature Consistency of Under-Performing Pixels and Valid Regions for Semi-Supervised Medical Image Segmentation

Tao Lei, *Senior Member, IEEE*, Yi Wang, Xingwu Wang, Xuan Wang, Bin Hu, *Fellow, IEEE*, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*—Existing semi-supervised medical image segmentation methods based on the teacher-student model often employ unweighted pixel-level consistency loss, neglecting the varying difficulties of different pixels and resulting in significant deficits in segmenting challenging regions. Additionally, consistency learning often excludes pixels with high uncertainty, which destroys the semantic integrity of a medical image. To address these issues, we propose a novel unified feature consistency (UFC) of under-performing pixels (UPPs) and valid regions for semi-supervised medical image segmentation: 1) high-performing pixels (HPPs) and UPPs are distinguished by confidence differences between the student and teacher models, and then UPPs are mapped into a latent feature space to improve consistency learning effect (UPPFC); 2) in order to obtain richer semantic information from a medical image, vectors of valid regions are selected from both image- and patch-level class feature vectors by using the output probabilities of the teacher model; and 3) these vectors are mapped into the latent feature space for class feature consistency (CFC) learning as a supplement to UPPFC which only focuses on challenging regions for pixel-level consistency learning, thereby enhancing the model's ability to learn structured semantic information from images themselves. Experimental results demonstrate that the proposed UFC achieves sufficient learning for challenging regions and retains the semantic integrity of medical images. Encouragingly, our proposed UFC provides better-segmentation results than the current state-of-the-art methods on three publicly available datasets. Our codes will be released at: https://github.com/SUST-reynole.**

***Index Terms*—Consistency learning, latent feature space, medical image segmentation, semi-supervised learning.**

Received 16 April 2024; revised 23 June 2024 and 9 September 2024; accepted 16 September 2024. Date of publication 23 September 2024; date of current version 5 February 2025. This work was supported in part by the National Natural Science Foundation of China under Program 62271296, Program 62201334, and Program 62301302; and in part by the Scientific Research Program Funded by the Shaanxi Provincial Education Department under Program 23JP014 and Program 23JP022. *(Corresponding author: Xuan Wang.)*

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

Tao Lei and Yi Wang are with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitao@sust.edu.cn; 201806060626@sust.edu.cn).

Xingwu Wang is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 102206, China (e-mail: xingwuwang@bupt.edu.cn).

Xuan Wang is with the Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: wangxuan@nwpu.edu.cn).

Bin Hu is with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: bh@bit.edu.cn).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRPMS.2024.3465561.

Digital Object Identifier 10.1109/TRPMS.2024.3465561

I. Introduction

MEDICAL image segmentation is crucial in computer vision and medical image analysis, aiming to accurately locate and segment various organs or lesions. Fully supervised learning techniques were widely introduced in early medical image segmentation, but a large amount of images with annotations required for training has become a significant barrier to widespread application. For this, semi-supervised learning is proposed to address the scarcity of labeled data by utilizing a small amount of labeled data and a large amount of unlabeled data. Recent advancements in semi-supervised learning have been explored, such as consistency learning [1], [2], [3], [4], [5], [17], [18], [19], [20], [21], [22], [53], [54], [55], [56], [57], [58], [62], adversarial learning [8], [23], [24], [25], [26], [27], self-training [29], [30], [31], and contrastive learning [32], [33], [34], [35], [59], [60], [61]. Among them, consistency learning methods are most commonly used in the field of semi-supervised medical image segmentation.

Although the above methods have achieved great success in medical images, two problems remain to be solved as illustrated in Fig. 1. First, different regions in images contain distinct prior structured information, and those with complex prior structures generally show greater segmentation challenges. However, most existing methods use unweighted pixel-level consistency loss, which neglects the difference in segmentation difficulties of different pixels, resulting in obvious learning insufficiency in pixels within challenging regions. Second, the overall prior structure of a medical image, which includes region-level (image-level and patch-level) semantic information, is crucial as mentioned in [50] and [51]. But most existing methods avoid, including pixels with high uncertainty for consistency learning, which destroys the overall semantic integrity and then harms the model's ability to learn image-level or patch-level semantic information from images themselves.

To address the aforementioned issues, a unified feature consistency (UFC) strategy of under-performing pixels (UPPs) and valid regions for semi-supervised medical image

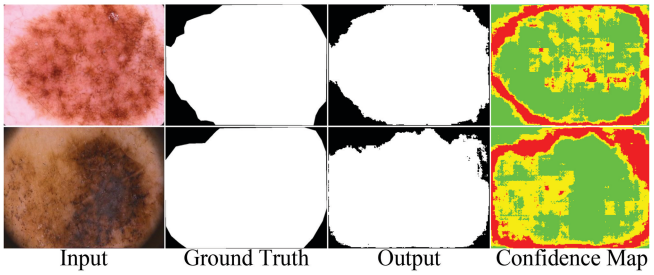


Fig. 1. Segmentation results of skin lesion images using FixMatch [17]. Green, yellow, and red pixels indicate predictions are high-($\varrho \geq 0.95$), middle-($0.8 \leq \varrho < 0.95$), and low-confidence ($\varrho < 0.8$), respectively, where ϱ is the predicted probability for a pixel. From confidence map, pixels within different regions of an image exist distinct segmentation difficulties. Furthermore, it can be clearly observed that a large number of mid-confidence pixels and some low-confidence pixels are predicted correctly, but only high-confidence pixels are used for model training, severely damaging the integrity for structured semantic information in a medical image.

segmentation is proposed: 1) pixels are categorized into high-performing pixels (HPPs) and UPPs based on confidence differences between the student and teacher models; 2) a strong constraint is imposed on UPPs by using consistency learning in a latent feature space, rather than the predicted result of networks as used for HPPs; and 3) finally, a class feature extractor is designed to extract valid region-level class feature vectors, which is helpful for improving the overall feature representation ability of our network. The main contributions of this article are summarized as follows.

1) Unlike existing methods that treat all pixels in a medical image equally, resulting in deficits in learning challenging regions [1], [2], [3], [22], [28], [56], [57], [62], a novel consistency learning strategy focusing on UPPs in medical images is proposed. Based on the difference of confidence from the student and teacher models, UPPs are selected and then are mapped into a latent feature space for better-consistency constraint, which enhances the model's ability in segmenting challenging regions and then boosts robustness of the model.
2) Unlike existing methods that exclude the high-uncertainty pixels [5], [17], [20], [31], [34], [49], [56], [60], damaging the structured semantic information of a medical image, we propose a novel region-level (including image-level and patch-level) class feature consistency (CFC) learning strategy. Region-level class feature vectors are obtained by using the predicted probabilities of the teacher model without damaging the integrity for structured semantic information of medical images, thereby achieving effective learning on region-level semantic information.
3) Experimental results demonstrate the superiority of the proposed UFC over the current state-of-the-art (SOTA) methods on three publicly available datasets, including international skin imaging collaboration (ISIC) [41], multiorgan nuclei segmentation (MoNuSeg) [42], and left atrial (LA) [43]. Furthermore, as a plug-and-play module, UFC can be directly incorporated into any semi-supervised method based on teacher-student models demonstrating its strong universality.

The remainder of this article is structured as follows. Section II reviews related work. Section III presents a detailed description of the proposed framework. Section IV reports the experimental results and discusses each component of our proposed method. Moreover, a further study is presented in Section V. Finally, Section VI provides a summary and draws the conclusion.

II. Related Work

A. Medical Image Segmentation

The current deep learning based medical image segmentation methods [6] can be roughly categorized into two groups: 1) CNN-based and 2) Transformer-based methods. CNN-based methods, e.g., CE-Net [7], 3-D UX-Net [8], LKAU-Net [9], SGU-Net [10], and PHNet [11], utilize hierarchical representations to capture local features of medical images, which often introduce some functional modules, such as pyramid feature fusion, attention mechanism, and depthwise separable convolutions, to enhance the network's ability of feature representation. Transformer-based medical image segmentation methods, e.g., TransUNet [12], UNETR [13], FAT-Net [14], ConvFormer [15], and FCT [16], process global information of images through self-attention mechanisms, allowing networks for better-capturing long-range dependencies between pixels, thereby improving the accuracy and robustness of medical image segmentation. Although the advance of CNN and Transformer has been applied to the various tasks of medical image segmentation, due to the limited labeled data, a good performance has been primarily achieved in fully supervised settings.

B. Semi-Supervised Learning

Semi-supervised learning overcomes the limitation of scarce labeled data by exploiting effective feature information from a large amount of unlabeled data in medical image segmentation. Consistency learning techniques [1], [2], [3], [4], [5], [17], [18], [19], [20], [21], [22], [53], [54], [55], [56], [57], [58] are widely used in the field of semi-supervised learning. The core idea of these methods is to enforce pixel-level consistency prediction under different perturbations. For example, Yu et al. [20] proposed an uncertainty-aware self-ensembling model (UA-MT), which excludes unreliable predictions by using Monte Carlo Dropout [48]. Due to minor improvements achieved by UA-MT, Xu et al. [54] rethought UA-MT and proposed an ambiguity-consensus mean-teacher model (AC-MT) which flips the selection criteria of UA-MT, focusing on selecting fuzzy but informative voxels from unlabeled data as targets for consistency learning. In order to fully utilize the prior relationship between labeled and unlabeled data, Lei et al. [28] proposed an adversarial consistency learning strategy (ASE-Net) using two discriminators, which focuses on the difference in output quality between labeled and unlabeled data, as well as the difference in output quality of unlabeled data under perturbed and unperturbed conditions. For the same purpose, Gao et al. [55] proposed a correlation aware mutual

learning (CAML), which uses labeled data to construct prototypes and calculates the distance between pixels in unlabeled data and the prototype in the embedding space to transfer prior knowledge of labels to unlabeled data. However, the weight update based on the MT model is unidirectional, which severely limits the learning ability of the teacher and student models, especially in the later stages of training. Motivated by this, He et al. [56] proposed a network using a bilateral exponential moving average strategy for bidirectional supervision (BSNet) to address this issue. To fully utilize disturbance information, Li et al. [2] proposed a transformation-consistent self-ensembling model (TCSM_v2), which exploits consistency from transformation, including rotation, flipping, and scaling operations. To address the issue of insufficient learning caused by supervising only the final layer of the encoder, Lei et al. [62] proposed a knowledge transfer strategy utilizing deep supervision. Additionally, Luo et al. [18], [53] proposed a method that simultaneously predicts pixel-by-pixel segmentation maps and geometrically aware level-set representations of targets as a dual-task-consistency (DTC) regularization strategy. Also, they added multiscale segmentation heads to make predictions at different scales, and ensured that each scale produces similar prediction values through consistency regularization (URPC). Furthermore, Sohn et al. [17] proposed a simplified consistency regularization combining with pseudo-labeling (FixMatch), which compels the model to produce consistency only for high-confidence predictions between images from strong and weak data augmentations Subsequently, Yang et al. [5] noticed that the exploration of perturbation space is still relatively limited, although FixMatch can achieve competitive results. Therefore, they designed a unified perturbation strategy (UniMatch) to utilize broader perturbation space. To further explore the perturbation space, Wang et al. [57] proposed a feature discrepancy loss that enables two branch networks to learn how to infer the input in different ways while arriving at consistent predictions (CCVC), thereby forcing the subnetworks to learn different information. Su et al. [58] compared and evaluated two subnetworks to select more reliable pseudo-labels, thereby preventing the model from being misled by poor predictions. While the aforementioned methods have achieved excellent performance, they often exclude predictions with high uncertainty, leading to a damage on the structured semantic information of medical images. And they do not emphasize learning challenging pixels. In response to these problems, we design a novel UFC module for UPPs and valid regions.

C. Representation Learning

Representation learning [47] refers to the process of learning useful and high-level representations from raw data. Contrastive learning is often regarded as a typical technique in representation learning, which aims to learn representations by pulling closer the distance between similar sample pairs while pushing apart the distance between dissimilar sample pairs. The methods based on contrastive learning, such as MOCO [37], SimCLR [36], and SwAV [38], were initially popular in the field of self-supervised learning. Due to the excellent performance of contrastive learning for learning representation, semi-supervised medical image segmentation methods based on contrastive learning [32], [33], [34], [35], [59], [60], [61] were soon introduced. The core idea of these methods is to assume that pixels with the same semantics are positive samples, while those with different semantics are negative ones. For example, Hu et al. [59] achieved good performance by using a pretraining stage based on self-supervised contrastive learning without any labels, and then combined pixel-level contrastive learning with supervised fine-tuning only on the labeled part of the data, greatly reducing the computational cost of pixel-level contrastive loss. Since pixel-level pseudo labels are not as accurate as we expect, and the difference of features between patches is easier to discriminate correctly, Wu et al. [34] divided the image into different types of patches based on the proportion of pixels with the same category. They thereafter jointly guided the calculation of pixel-level contrastive loss using both patch types and pseudo labels (CDCL). To effectively reduce the noise sampled from the pseudo labels of unlabeled data, Wang et al. [60] proposed an uncertainty-guided contrastive learning method, where the uncertainty is calculated by average ensembling the prediction results obtained from the heterogeneous decoders of CNN and Transformer, thereby guiding pixel-level contrastive learning. Directly using pseudo labels of unlabeled images in cross-entropy loss can lead to erroneous predictions. To address this, Chaitanya et al. [35] used two independent network branches: one for calculating segmentation losses from labeled data and another for contrastive losses from both labeled and pseudo-labeled data. Moreover, Xie et al. [61] proposed a contrastive learning framework based on probabilistic representations, where pixel-level representations are modeled using Gaussian distributions and then adjusted their contributions in the contrastive learning process according to the reliability of the semantic representations, enabling the model to tolerate erroneous semantic representations. Although these methods mentioned above have achieved better results with contrastive learning, they often neglect image-level or patch-level representation learning. Besides contrastive learning, BYOL [39] and SimSiam [40] are also notable representation learning methods, which utilize self-generated predictions for training models. Following these works, we propose a UFC module as the core of our semi-supervised medical image segmentation method.

III. Method

Fig. 2 presents a semi-supervised Siamese network that utilizes our proposed UFC for medical image segmentation. Specifically, UFC incorporates UPP feature consistency learning (UPPFC) and CFC learning. UPPFC focuses on UPPs that are challenging to segment and are not fully mastered by the student model, while CFC fully exploits structured overall semantic information from valid images or patches.

Assume a medical image dataset comprises M labeled images $X_L = \{\chi_l\}_{l=1}^{M}$ with labels $Y_L = \{\gamma_l\}_{l=1}^{M}$, and N unlabeled images $X_U = \{\chi_l\}_{l=M+1}^{M+N}$. According to [17], the network plays the role of a teacher when X_U are subjected to weak augmentation, which does not perform backpropagation to update the network's parameters. Similarly, it takes the

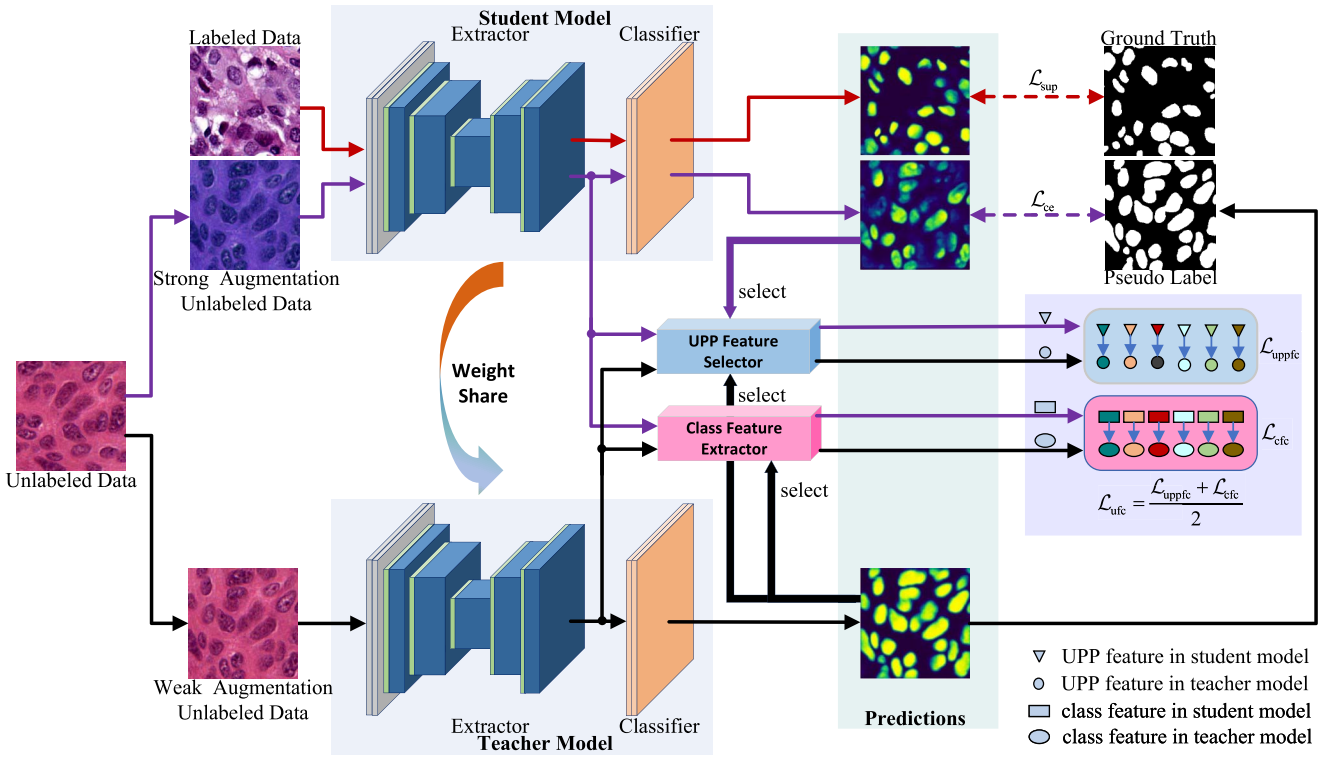


Fig. 2. Overall network architecture. The student model and teacher model share the same architecture and parameters. We update the student model by reducing the weighted sum of $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{un\_ce}}$, and $\mathcal{L}_{\text{ufc}}$. Specifically, when calculating $\mathcal{L}_{\text{ufc}}$, the student model and the teacher model need to first map the two types of features into a latent feature space and then measure their distance by using cosine similarity. The red arrows represent the supervised branch, the purple ones represent the strong augmented branch for unsupervised data, and the black ones represent the weak augmented branch for unsupervised data. ▽ and ○ with same color indicate UPP features at the same position from the teacher model and the student model for a medical image, while ▭ and ⬭ with same color indicate class features of the same class in the same region from the teacher model and the student model for a medical image.

role of a student when X_U suffer from strong augmentation resulting in less reliable outputs. The student model's outputs for X_U are supervised by pseudo-labels from the teacher model with cross-entropy loss $\mathcal{L}_{\text{un\_ce}}$. Meanwhile, according to the predicted probabilities of the proposed network, UPP feature vectors and class feature vectors of X_U are derived from the features extracted by the network's extractor. Subsequently, these vectors are mapped into a latent feature space for UFC

$$\mathcal{L}_{\text{ufc}} = (\mathcal{L}_{\text{uppfc}} + \mathcal{L}_{\text{cfc}})/2 \tag{1}$$

where $\mathcal{L}_{\text{uppfc}}$ is UPPFC loss and $\mathcal{L}_{\text{cfc}}$ is CFC loss. Moreover, the student model's outputs for X_L are supervised by Y_L with cross-entropy loss $\mathcal{L}_{\text{ce}}$ and Dice loss $\mathcal{L}_{\text{dice}}$ as follows:

$$\mathcal{L}_{\text{sup}} = (\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}})/2. \tag{2}$$

Finally, the total loss is defined as

$$\mathcal{L} = \omega_{\text{sup}}\mathcal{L}_{\text{sup}} + \omega_{\text{un\_ce}}\mathcal{L}_{\text{un\_ce}} + \omega_{\text{ufc}}\mathcal{L}_{\text{ufc}} \tag{3}$$

where ω_{sup}, $\omega_{\text{un\_ce}}$, and ω_{ufc} are the corresponding loss weights of $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{un\_ce}}$, and $\mathcal{L}_{\text{ufc}}$, respectively.

A. Under-Performing Pixel Feature Consistency Learning

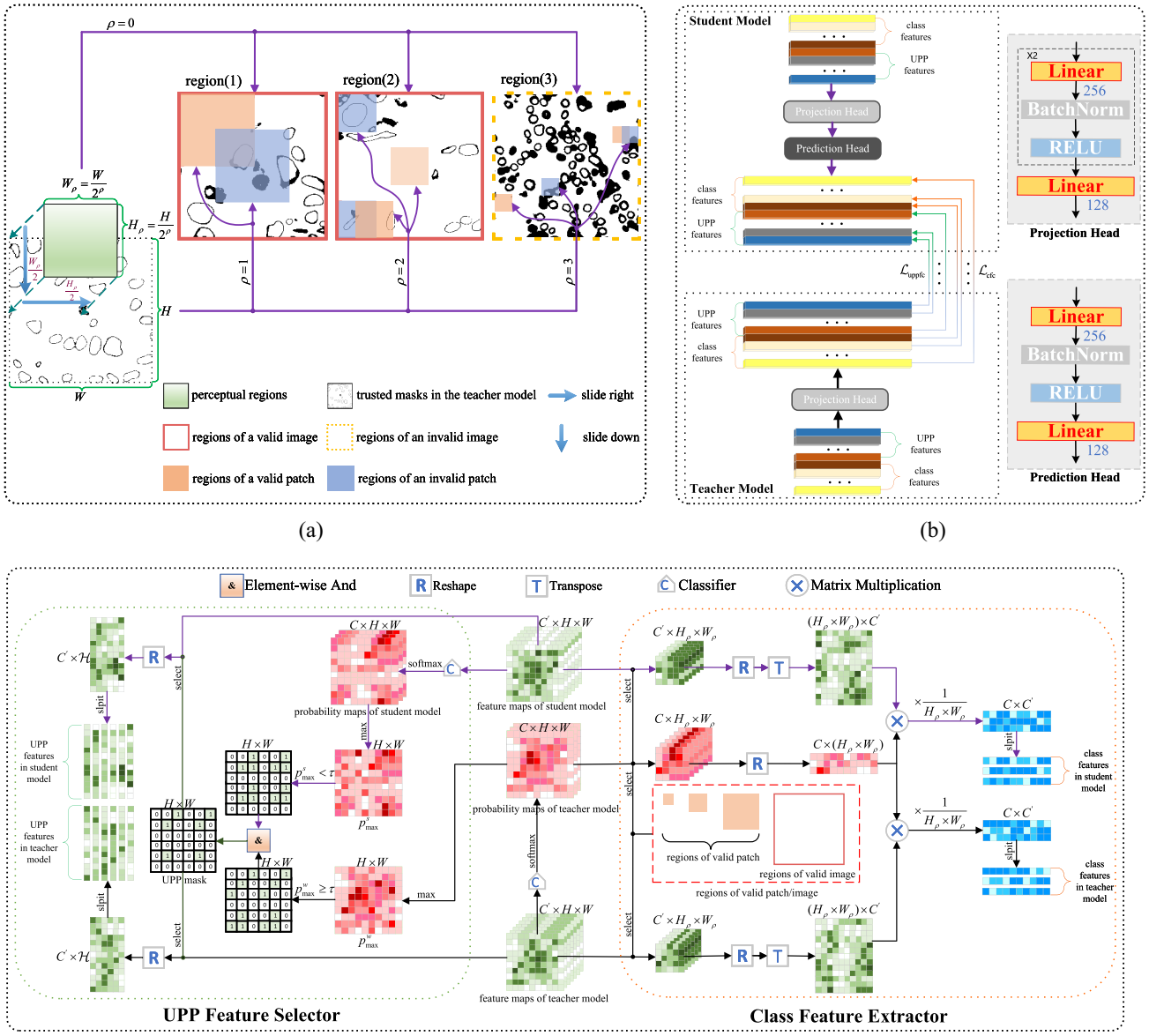
Because the accuracy of challenging region segmentation largely determines the segmentation performance of a model, UPPFC focuses on UPPs in a lower-dimensional latent feature space, thereby enhancing the model's ability of segmenting challenging regions.

x_u^s and x_u^w correspond to the strong and weak augmentation images of unlabeled data $X_u \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$, respectively, where B is the batch size, C_{in} is the number of input channels, and H and W are, respectively, the height and width of a medical image. Then the features z^s and z^w for x_u^s and x_u^w are extracted through the student and teacher models' extractors, respectively, while predictions p^s and $p^w \in \mathbb{R}^{B\times C\times H\times W}$ are obtained through corresponding classifiers, where C is the number of classes. Finally, the maximum predicted probabilities $p^s_{\max}$ and $p^w_{\max} \in \mathbb{R}^{B\times H\times W}$ for x_u^s and x_u^w are obtained, respectively, along with their corresponding pseudo-labels $\hat{y}_s$ and $\hat{y}_w \in \mathbb{R}^{B\times H\times W}$ as follows:

$$\begin{cases} p^s_{\max} = \max(p^s), & \hat{y}_s = \operatorname{argmax}(p^s) \\ p^w_{\max} = \max(p^w), & \hat{y}_w = \operatorname{argmax}(p^w) \end{cases} \tag{4}$$

where max refers to finding the maximum value of a matrix along its first dimension (starting from 0, consistent with coding conventions, that is, channel dimension), while argmax refers to finding the index of the maximum value along the first dimension of the matrix.

For each pixel x_i $(1 \le i \le B \times H \times W)$, the predicted max probabilities from the student model and the teacher model are denoted as $p^{(s,i)}_{\max}$ and $p^{(w,i)}_{\max}$. Compared to the student model, the high-confidence outputs of the teacher model are more



(a) (b) (c)

Fig. 3. Illustrative pipeline of the proposed UFC. By selecting pixels that simultaneously meet high confidence in the teacher model and nonhigh confidence in the student model as UPPs, we perform under-performing feature consistency learning (UPPFC). We utilize probability maps to extract class feature vectors and then perform CFC learning. In (a), for all masks, the black area represents 0 while the white area represents 1. In (b), for all features, the same color represents UPP features at the same position or class features of the same class in the same region from the teacher model and the student model for a medical image. In (c), especially, the purple arrows represent the branch for the student model and the black ones represent the branch for the teacher model.

reliable. In this way, HPPs and UPPs are defined in terms of the high-confidence outputs of the teacher model as follows:

$$\begin{cases} \chi_{\text{hpp}} = \left\{ x_i \mid p_{\max}^{(s,i)} \geq \tau, p_{\max}^{(w,i)} \geq \tau \right\} \\ \chi_{\text{upp}} = \left\{ x_i \mid p_{\max}^{(s,i)} < \tau, p_{\max}^{(w,i)} \geq \tau \right\} \end{cases} \tag{5}$$

where $0 \leq \tau \leq 1$. The acquisition of the UPP mask is illustrated in Fig. 3(c). Most HPPs in the corresponding teacher and student models have similar semantic information, for both of two models are extremely likely to have excellent predictions for HPPs. Thus, according to clustering assumptions [46], their distances in the feature space are extremely close. Therefore, the consistency learning for HPPs in the feature space will lead to redundant computation. In contrast, UPPs that the student model fails to learn well, compared to the teacher model, exhibit the significant difference of semantic information extracted by the teacher model and the student model. Thus, focusing on UPPs for consistency learning is able to effectively improve the learning effect and efficiency of models.

The $z_{\hbar}^{s}$ and $z_{\hbar}^{w}$ are feature vectors of $x_{\hbar} \in \chi_{\text{upp}}$ from the student and teacher models' extractors, respectively. Then $\phi_{\hbar}^{s} = \phi(z_{\hbar}^{s})$ and $\phi_{\hbar}^{w} = \phi(z_{\hbar}^{w})$ are obtained, where ϕ indicates a projection head. Subsequently $\psi_{\hbar}^{s} = \psi(\phi_{\hbar}^{s})$ is acquired, where ψ denotes a prediction head. As depicted in Fig. 3(b), the two heads are the same as [40]. Next, the cosine similarity between

$\psi_{\hbar}^{s}$ and $\phi_{\hbar}^{w}$ can be expressed as

$$S(\psi_{\hbar}^{s}, \phi_{\hbar}^{w}) = \frac{\langle \psi_{\hbar}^{s}, \phi_{\hbar}^{w} \rangle}{\|\psi_{\hbar}^{s}\|_2 \cdot \|\phi_{\hbar}^{w}\|_2} \tag{6}$$

where $\langle \cdot \rangle$ denotes the inner product and $\| \cdot \|_2$ indicates l_2-norm. Then the distance of UPPs between the student model and the teacher model is measured as

$$D(\psi_{\hbar}^{s}, \phi_{\hbar}^{w}) = 1 - S(\psi_{\hbar}^{s}, \phi_{\hbar}^{w}). \tag{7}$$

Finally, the proposed UPPFC loss is defined as

$$\mathcal{L}_{\text{uppfc}} = \frac{1}{\mathcal{H}} \sum_{\hbar=1}^{\mathcal{H}} D(\psi_{\hbar}^{s}, \phi_{\hbar}^{w}) \tag{8}$$

where $\mathcal{H}$ is the number of UPPs in a batch.

B. Class Feature Consistency Learning for Valid Regions

UPPFC focuses on the segmentation quality of UPPs, significantly enhancing the model's ability in segmenting challenging regions. However, UPPs are defined based on high-confidence outputs from the teacher model. Additionally, to address the limitation of pixel-level consistency learning strategies in improving the overall feature representation of our network, we propose CFC, including image-level CFC learning (ICFC) and patch-level CFC learning (PCFC) with perceptual region adjustment, which consider global (image-level) and local (patch-level) regions simultaneously. To be specific, different class features are first obtained by class feature extractor as illustrated in Fig. 3(c). Second, the extracted features with the same class in a valid region from both the student model and the teacher model are mapped into a latent space as shown in Fig. 3(b), and then we measure the distance between them by cosine similarity for consistency learning. ICFC preserves global semantic information integrity, while PCFC enhances the model's perceptual ability by utilizing different-sized perceptual regions due to the spatial variability for target regions of images. Further details will be elaborated in the following two sections.

1) Image-Level Class Feature Consistency Learning: First, for a medical image $\chi_b (1 \le b \le B)$, the corresponding trusted mask is denoted as $\mathcal{M}^b \in \mathbb{R}^{H \times W}$. Then let $\mathcal{M}^b_{(d,q)}$ represent the dth row and qth column element of $\mathcal{M}^b$, i.e.,

$$\mathcal{M}^b_{(d,q)} = \begin{cases} 1, & p^{(w,b,\max)}_{(d,q)} \ge \sigma \\ 0, & p^{(w,b,\max)}_{(d,q)} < \sigma \end{cases} \tag{9}$$

where $p^{(w,b,\max)}_{(d,q)}$ is the maximum predicted probability from the teacher model for the dth row and qth column pixel of χ_b.

Subsequently, the set of valid images is defined as

$$\mathrm{X}_{\text{valid}} = \left\{ \chi_b \middle| \frac{\sum_{q=1}^{W} \sum_{d=1}^{H} \mathcal{M}^b_{(d,q)}}{\mathrm{HW}} \ge \eta, \chi_b \in X_u \right\}. \tag{10}$$

Especially, for a 3-D dataset, 2-D slices obtained from three different dimensions of the 3-D features extracted by the network's extractor, are exploited to select valid images.

A valid medical image $\chi_v \in \mathrm{X}_{\text{valid}}$, undergoes strong and weak augmentations, resulting in corresponding features $z^{(s,v)}$ and $z^{(w,v)}$ from their respective extractors. Subsequently, $z^{(w,v)}$ is predicted by the teacher model's classifier as $p^{(w,v)} \in \mathbb{R}^{C \times H \times W}$. Let $p^{(w,v,c)} \in \mathbb{R}^{H \times W}$ denote the prediction of cth $(0 \le c \le C)$ class according to $p^{(w,v)}$. Using the more reliable predictions from the teacher model compared to the student model, image-level class feature vectors $Z^{(s,v,c)}$ and $Z^{(w,v,c)}$ are extracted from $z^{(s,v)}$ and $z^{(w,v)}$ as follows:

$$\begin{cases} Z^{(s,v,c)} = \frac{1}{\mathrm{HW}} \sum_{q=1}^{W} \sum_{d=1}^{H} z^{(s,v)}_{(d,q)} p^{(w,v,c)}_{(d,q)} \\ Z^{(w,v,c)} = \frac{1}{\mathrm{HW}} \sum_{q=1}^{W} \sum_{d=1}^{H} z^{(w,v)}_{(d,q)} p^{(w,v,c)}_{(d,q)} \end{cases} \tag{11}$$

where $z^{(s,v)}_{(d,q)}$, $z^{(w,v)}_{(d,q)}$, and $p^{(w,v,c)}_{(d,q)}$ are the elements in the dth row and qth column of $z^{(s,v)}$, $z^{(w,v)}$ and $p^{(w,v,c)}$, respectively. As seen from (11) and Fig. 3(c), class feature vectors contains features of all pixels in valid images, which preserves the structured semantic information of medical images. Note that these low-confidence pixels do not significantly affect the final extracted image-level class feature vectors, given their relatively smaller number compared to high-confidence ones, as illustrated in Fig. 3(a).

$\psi^{s}_{(v,c)} = \psi(\phi^{s}_{(v,c)})$ is derived after $\phi^{s}_{(v,c)} = \phi(Z^{(s,v,c)})$ and $\phi^{w}_{(v,c)} = \phi(Z^{(w,v,c)})$ are obtained. Next, the cosine similarity between $\psi^{s}_{(v,c)}$ and $\phi^{w}_{(v,c)}$ is computed as

$$S(\psi^{s}_{(v,c)}, \phi^{w}_{(v,c)}) = \frac{\langle \psi^{s}_{(v,c)}, \phi^{w}_{(v,c)} \rangle}{\|\psi^{s}_{(v,c)}\|_2 \cdot \|\phi^{w}_{(v,c)}\|_2}. \tag{12}$$

Then the distance of class feature vectors between the student model and the teacher model is computed as

$$D(\psi^{s}_{(v,c)}, \phi^{w}_{(v,c)}) = 1 - S(\psi^{s}_{(v,c)}, \phi^{w}_{(v,c)}). \tag{13}$$

Finally, the proposed ICFC loss is defined as

$$\mathcal{L}^{\text{image}}_{\text{cfc}} = \frac{1}{C\mathcal{V}} \sum_{v=1}^{\mathcal{V}} \sum_{c=1}^{C} D\left(\psi^{s}_{(v,c)}, \phi^{w}_{(v,c)}\right) \tag{14}$$

where $\mathcal{V}$ is the number of valid images in a batch.

2) Patch-Level Class Feature Consistency Learning With Perceptual Region Adjustment: Optical medical images, such as those of skin lesions and cell nuclei, often exhibit significant spatial variability. Therefore, we propose PCFC to enhance the model's perceptual ability for segmentation targets, as illustrated in Fig. 3(a).

The feature map $z \in \mathbb{R}^{C' \times H \times W}$ of an unlabeled medical image $x_u \in \mathbb{R}^{C_{\text{in}}} \times H \times W$ is obtained by the extractor. To obtain more diverse feature information, three different-sized perceptual regions are used for various segmentation targets: 1) region(1) $\in \mathbb{R}^{C' \times H_1 \times W_1}$; 2) region(2) $\in \mathbb{R}^{C' \times H_2 \times W_2}$; and 3) region(3) $\in \mathbb{R}^{C' \times H_3 \times W_3}$, where $H_\rho = (H/2^\rho)$ and $W_\rho = (W/2^\rho)$ $(1 \le \rho \le 3)$. Especially, the patch-level region is changed to image-level region as same as Section III-B1 when $\rho = 0$. For each perceptual region region$(\rho) \in \mathbb{R}^{C' \times H_\rho \times W_\rho}$, region-specific information from z is captured as follows:

$$\left\{ R^{(j,k)}_{\rho} \middle| R^{(j,k)}_{\rho} = z \left[0 : C', \frac{H_\rho}{2} j : \frac{H_\rho}{2} j + H_\rho, \frac{W_\rho}{2} k : \frac{W_\rho}{2} k + W_\rho \right], 0 \le j < 2^{\rho+1} - 1, 0 \le k < 2^{\rho+1} - 1 \right\}. \tag{15}$$

As shown in (15), each perceptual region region(ρ) contains $(2^{\rho+1}-1)^2$ patch-level features. Similar to ICFC in Section III-B1, the same class feature extraction operation and loss calculation are used for PCFC to obtain $\mathcal{L}_{\text{cfc}}^{\text{patch}}$. Finally, the proposed CFC loss is as follows:

$$\mathcal{L}_{\text{cfc}} = \mathcal{L}_{\text{cfc}}^{\text{image}} + \mathcal{L}_{\text{cfc}}^{\text{patch}}. \tag{16}$$

By simultaneously using UPPFC and CFC, the proposed UFC is achieved, effectively exploiting both pixel-level and region-level (including image- and patch-level) semantic information, thereby improving the model's segmentation ability.

C. Cross-Entropy in Consistency Learning

By applying UFC after the extractor, strong feature representations are effectively learned from unlabeled data. However, the UFC cannot directly optimize the parameters of the classifier. Also, as shown in (5), HPPs are high-confidence in both the student and teacher models. Compared to the teacher model, high-confidence predictions from the student model are more likely to be 'overconfident'. Moreover, applying feature consistency learning to all HPPs, as done for UPPs, would significantly increase unnecessary computation, as illustrated in Section III-A. Therefore, inspired by FixMatch [17], a simplified consistency regularization for HPPs is introduced, which not only directly optimizes the classifier's parameters but is also essential for learning HPPs themselves.

The prediction of the student model for $x_e \in \chi_{\text{hpp}}$ is denoted as p_{hpp}^s, and its corresponding pseudo-label from the teacher model is $\hat{y}_e^w$. By applying the cross-entropy loss function φ, the consistency loss is defined as follows:

$$\mathcal{L}_{\text{un\_ce}} = \frac{1}{\mathcal{E}} \sum_{e=1}^{\mathcal{E}} \varphi\left(p_{\text{hpp}}^s, \hat{y}_e^w\right) \tag{17}$$

where $\mathcal{E}$ refers to the number of HPPs in a batch.

IV. Experiments

A. Datasets

Skin Lesion Segmentation Dermoscopy Dataset: This dataset is from the 2018 ISIC skin lesion segmentation challenge [41]. It contains 2594 training images and 100 validation images, featuring various skin lesion types and resolutions. Following [28], all images are resized to 256 × 192, then 10% (259 images) and 20% (519 images) are still used as labeled data, and the rest are used as unlabeled data.

MoNuSeg Dataset: This dataset is from the 2018 MoNuSeg challenge [42]. It consists of 30 training images and 14 validation images with H&E stained tissue from various organs, featuring low contrast between targets and background tissues. The original MoNuSeg images, sized at 1000 × 1000, are uniformly cropped into nonoverlapping 250 × 250 subimages, which are then resized to 256 × 256. 1 image (16 subimages) and 2 images (32 subimages) are randomly selected as labeled data and the rest as unlabeled data.

3-D Left Atrium Segmentation MR Dataset: This dataset is from the 2018 LA Segmentation Challenge [43]. It includes 100 3-D gadolinium-enhanced MR images with a resolution of 0.625 × 0.625 × 0.625 mm^3. Following [3], [18], [20], [28], 80 scans are used for training while 20 scans are used for validation. A standardized data preprocessing scheme that randomly crops the LA data to the size of 112 × 112 × 80 is adopted. In our experiment, 10% (8 scans) and 20% (16 scans) are still used as labeled data while the rest are used as unlabeled data.

B. Implementation Details

Network Architecture: We use UNet++ [44] on both ISIC and MoNuSeg datasets while VNet [45] is used on LA dataset. The extractor in Fig. 2 refers to all components in UNet++ [44]/VNet [45] except for the classifier, with an output channel of 64/16. For the ISIC and LA datasets, to reduce information redundancy and computational cost during the training process, the features obtained by the extractor are downsampled from the original size of 64 × 256 × 192 and 16 × 112 × 112 × 80 to 64 × 128 × 96 and 16 × 56 × 56 × 40, respectively. For MoNuSeg dataset, which involves a small target segmentation task where precise segmentation of every pixel is crucial, the original-sized features are used.

Hyperparameter Settings: ω_{sup}, ω_{ufc}, σ and η are set to 1, 0.1, 0.8 and 90%, respectively. The first two hyperparameters are empirical values while hyperparameters σ and η will be discussed in Section V-A. Following [2], $\omega_{\text{un\_ce}}$ is set to $0.1e^{(-5(1-t/t_{\max})^2)}$, where t denotes the current training epoch, and $t_{\max}$ represents the total training epoch. Following [17], τ is set to 0.95 for defining high-confidence pixels. All of the experiments are conducted on a server equipped with NVIDIA GeForce RTX 3090 24 GB. To fully utilize the GPU memory, batch size is set to 32, 28, and 8 for the ISIC, MoNuSeg, and LA datasets, respectively. The Adam optimizer is used with a learning rate of 0.001 for both the ISIC and MoNuSeg datasets, and 0.0005 for the LA dataset.

Data Augmentation: Weak augmentation for all images is used, including random horizontal flipping for the ISIC dataset, random horizontal flipping, and random rotations (90°, 180°, and 270°) for the MoNuSeg dataset, and random cropping for the LA dataset. Additionally, for all three datasets, three types of strong augmentation are used, including 1) CutMix [19]; 2) Gaussian blur; and 3) color jitter.

C. Comparison With State-of-the-Art Methods

Our method is compared with several SOTA methods, including MT [1], UA-MT [20], TCSM_v2 [2], FixMatch* [17], CPS [22], DTC [18], MC-Net [3], CDCL [34], AC-MT [54], ASE-Net [28], ASE-Net† [28], UniMatch [5], and MCF [52]. In particular, * refers to the extra use of CutMix [19], and † indicates that DyBAC [28] is additionally used.

Our method is evaluated using Dice coefficient (DC) and Jaccard coefficient (JC) as the main indicators. For the ISIC and MoNuSeg datasets, accuracy (AC), specificity (SP), and sensitivity (SE) are also used. For the LA dataset, 95%

TABLE I
STATISTICAL COMPARISON WITH SOTA METHODS ON THE ISIC AND MONUSEG DATASETS. THE BEST RESULTS ARE IN BOLD

	ISIC						MoNuSeg					
Method	Label	DC (%)	JC (%)	SE (%)	AC (%)	SP (%)	Label	DC (%)	JC (%)	SE (%)	AC (%)	SP (%)
SupOnly		82.57	73.55	88.31	91.01	93.76		68.79	53.18	85.92	83.99	83.18
MT [1] [NeurIPS'17]		84.58	76.54	87.25	92.02	95.69		69.28	53.68	83.23	85.01	85.16
UA-MT [20] [MICCAI'19]		84.80	78.02	88.63	91.94	95.82		72.92	57.99	84.20	87.52	88.12
TCSM_V2 [2] [TNNLS'20]		84.71	75.55	90.22	91.92	95.77		71.07	55.89	85.17	85.78	85.63
FixMatch* [17] [NeurIPS'20]		85.49	77.81	89.88	92.29	96.04		73.36	58.32	75.04	89.65	92.99
CPS [22] [CVPR'21]		84.72	76.81	86.87	91.87	95.42		72.43	57.29	78.73	88.16	90.16
DTC [18] [AAAI'21]	259	84.56	76.33	87.19	91.79	95.54	1	68.99	53.49	**88.61**	84.18	82.88
MC-Net [3] [MICCAI'21]	/	84.81	76.64	87.41	91.91	95.97	/	72.72	57.69	81.05	87.91	89.33
CDCL [34] [CVPR'22]	2594	84.92	78.06	88.43	91.54	94.64	30	72.98	57.73	85.12	88.34	91.65
ASE-Net [28] [TMI'23]		84.70	77.94	90.18	91.72	95.80		71.01	55.74	85.51	85.92	85.74
ASE-Net† [28] [TMI'23]		85.19	78.80	90.38	92.40	96.15		72.06	56.70	84.43	87.90	88.83
UniMatch [5] [CVPR'23]		85.84	78.15	91.32	92.07	**96.42**		73.25	57.96	75.33	88.76	90.24
MCF [52] [CVPR'23]		84.71	77.19	89.16	91.20	95.30		73.18	58.08	73.77	**89.80**	**93.50**
Ours		**86.70**	**79.06**	**93.13**	**92.58**	96.17		**74.17**	**59.32**	79.84	89.37	91.55
SupOnly		84.36	75.64	88.83	92.15	94.95		69.25	54.43	68.86	89.36	94.02
MT [1] [NeurIPS'17]		85.83	77.48	89.97	92.57	94.46		76.56	62.44	78.49	91.02	93.93
UA-MT [20] [MICCAI'19]		86.19	78.06	90.94	92.71	94.49		76.84	62.79	77.48	91.27	94.45
TCSM_V2 [2] [TNNLS'20]		86.16	77.98	91.07	92.56	94.26		77.27	63.26	78.37	91.26	94.30
FixMatch* [17] [NeurIPS'20]		85.56	78.41	89.40	92.71	95.93		77.07	63.00	77.27	91.49	94.57
CPS [22] [CVPR'21]		86.34	78.17	90.57	92.72	94.78		77.24	63.31	77.48	91.39	94.59
DTC [18] [AAAI'21]	519	85.91	77.63	90.24	92.79	94.40	2	76.50	62.37	78.45	91.01	93.92
MC-Net [3] [MICCAI'21]	/	86.37	78.11	90.85	92.61	94.64	/	77.16	63.15	78.49	91.23	94.16
CDCL [34] [CVPR'22]	2594	86.15	78.04	90.17	92.45	95.18	30	76.45	62.38	78.76	91.35	93.89
ASE-Net [28] [TMI'23]		86.67	78.59	90.94	92.51	95.85		76.31	62.10	81.18	90.48	92.56
ASE-Net† [28] [TMI'23]		87.21	79.25	91.15	**93.09**	94.52		76.98	62.91	**81.76**	91.00	93.47
UniMatch [5] [CVPR'23]		86.35	78.55	90.72	92.44	95.37		76.78	62.59	77.85	91.32	94.64
MCF [52] [CVPR'23]		86.33	78.32	90.19	92.60	**96.66**		76.32	62.17	75.18	91.34	**95.04**
Ours		**87.33**	**79.84**	**91.84**	92.82	96.29		**78.34**	**64.64**	80.75	**91.53**	93.99
FullSup	100%	87.67	80.06	90.65	93.29	96.78	100%	79.58	66.44	79.58	92.38	95.36

TABLE II
STATISTICAL COMPARISON WITH SOTA METHODS ON THE LA DATASET. THE BEST RESULTS ARE IN BOLD

	LA								
Method	Label	DC (%)	JC (%)	95HD (mm)	ASD (mm)	Label	DC (%)	JC (%)	95HD (mm)	ASD (mm)
SupOnly		79.99	68.12	21.11	5.48		86.03	76.06	14.26	3.51
MT [1] [NeurIPS'17]		84.24	73.26	19.41	2.71		88.42	79.45	13.07	2.73
UA-MT [20] [MICCAI'19]		84.25	73.48	13.84	3.36		88.88	80.21	7.32	2.26
TCSM_V2 [2] [TNNLS'20]		84.21	73.19	19.56	3.07		86.26	76.56	9.67	2.35
FixMatch* [17] [NeurIPS'20]		87.79	78.33	9.42	2.44		90.33	82.43	6.36	1.64
CPS [22] [CVPR'21]		84.09	73.17	22.55	2.41		87.87	78.61	12.87	2.16
DTC [18] [AAAI'21]	8	86.57	76.55	14.47	3.74	16	89.42	80.98	7.32	2.10
MC-Net [3] [MICCAI'21]	/	87.71	78.31	9.36	2.18	/	90.34	82.48	6.00	1.77
AC-MT [54] [MedIA'23]	80	87.64	78.10	16.59	4.06	80	89.15	80.54	13.86	3.61
ASE-Net [28] [TMI'23]		87.10	77.36	9.93	2.37		89.43	81.00	9.81	2.12
ASE-Net† [28] [TMI'23]		87.83	78.45	9.86	2.17		90.29	82.76	7.18	1.64
UniMatch [5] [CVPR'23]		87.58	78.17	10.24	1.82		89.62	81.38	8.84	2.27
MCF [52] [CVPR'23]		86.63	77.01	**8.37**	2.95		89.07	80.76	8.19	2.49
Ours		**88.75**	**79.86**	8.58	**1.87**		**90.64**	**83.02**	**5.85**	**1.47**
FullSup	100%	91.14	83.82	5.75	1.52	100%	91.14	83.82	5.75	1.52

Hausdorff distance (95HD) and average symmetric surface distance (ASD) are also used. As shown in Tables I and II, our method always outperforms other competitive methods in different number of labeled data. On the ISIC dataset, particularly with 259/2594 labeled data, our method surpasses FixMatch* [17] by 1.21%. On the MoNuSeg dataset, especially with 2/30 labeled data, our method surpasses TCSM_V2 [2] by 1.07% in DC. On the LA dataset, especially with 8/80 labeled data, our method surpasses ASE-Net† [28] by 0.92% in DC.

The comparison of segmentation visualization is shown in Figs. 4 and 5. These visualizations further demonstrate the superiority of our proposed method.

D. Ablation Studies

As shown in Table III, to evaluate each component of the proposed method, an ablation study on the ISIC dataset is conducted, where only 259/2594 data is labeled.

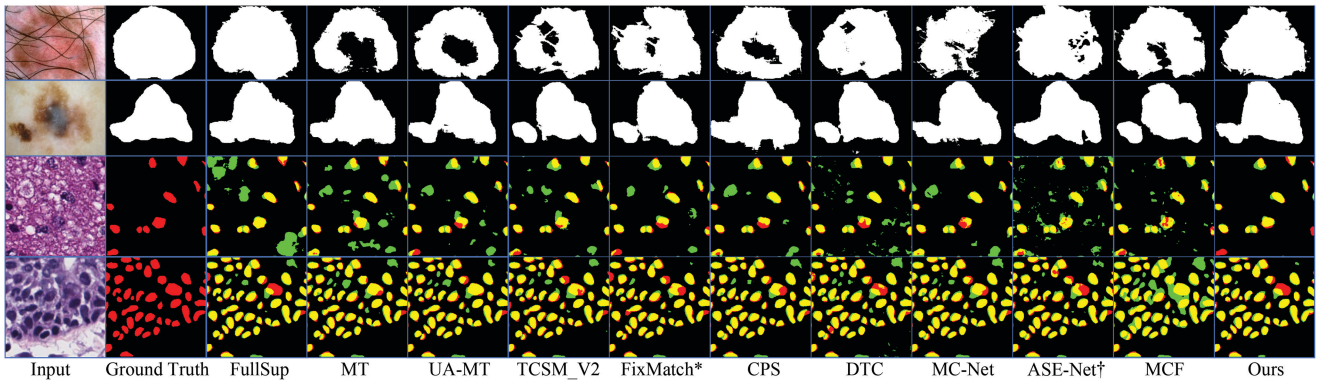


Fig. 4. Visualization result with different SOTA methods in skin and nuclei image segmentation. FullSup is trained with 100% labeled data while other methods are trained in a semi-supervised manner with 259/2594 labeled data for skin image, 2/30 labeled data for nuclei image, and rest for unlabeled data. Especially, for better presenting the segmentation results of nuclear cell, we employ different colors for visualization. Green and red pixels indicate the predictions and ground truth, respectively. Yellow pixels represent the overlapping regions between the prediction and ground truth.

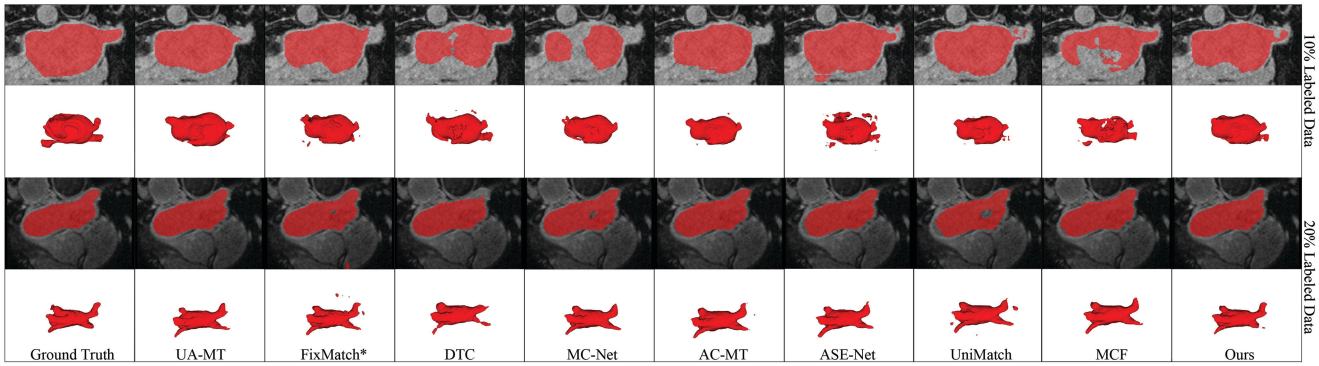


Fig. 5. Visualization result with different methods on the left atrium validation set by utilizing 8/80 and 16/80 labeled data, respectively.

TABLE III
STATISTICAL COMPARISON OF OUR ABLATION STUDIES IN DC METRIC ON THE ISIC DATASET WITH 259/2594 LABELED DATA

Method	$\mathcal{M}_{\text{un\_ce}}$	$\mathcal{L}_{\text{uppfc}}$	$\mathcal{L}_{\text{cfc}}$	$\mathcal{L}_{\text{un\_ce}}$	DC (%)
SupOnly					82.57
Scheme.1	NULL	✓	✓	✓	86.09
Scheme.2		✓			85.85
Scheme.3			✓		85.81
Scheme.4	UPPs			✓	85.44
Scheme.5		✓	✓		86.38
Scheme.6	UPPs	✓		✓	86.29
Scheme.7	UPPs		✓	✓	86.25
Ours	UPPs	✓	✓	✓	**86.70**

Without any semi-supervised techniques, the base segmentation network, UNet++ [44], is trained using only labeled data (SupOnly). By selecting UPPs and optimizing $\mathcal{L}_{\text{uppfc}}$, we achieve UPPFC (Scheme 2). To enhance the network's overall feature representation ability, we achieve CFC (Scheme 3) by optimizing $\mathcal{L}_{\text{cfc}}$ after selecting valid regions and obtaining class feature vectors using the class feature extractor. By simultaneously optimizing $\mathcal{L}_{\text{uppfc}}$ and $\mathcal{L}_{\text{cfc}}$ (Scheme 5), we improve the segmentation quality both locally and globally. Finally, we utilize $\mathcal{L}_{\text{un\_ce}}$ as an auxiliary loss to impose consistency regularization on the output predictions (Scheme 1, Scheme 4, Scheme 6, Scheme 7, Ours), better optimizing the classifier parameters and further improving the segmentation quality.

Ablation Studies for $\mathcal{L}_{\text{uppfc}}$ and $\mathcal{L}_{\text{cfc}}$: In Scheme 2 and Scheme 3, it is notable that UPPFC and CFC still outperform the SOTA methods when used independently. Furthermore, in Scheme 5, where $\mathcal{L}_{\text{uppfc}}$ and $\mathcal{L}_{\text{cfc}}$ are simultaneously optimized, an increase in DC from 85.85% to 86.38% is observed compared to Scheme 2. This indicates that CFC can complement UPPFC, addressing the challenge of enhancing the network's overall feature representation capacity, which is difficult for UPPFC that focuses on local information.

Ablation Studies for $\mathcal{M}_{\text{un\_ce}}$ and $\mathcal{L}_{\text{un\_ce}}$: In Scheme 1, following [17], a simplified version of consistency regularization is applied using all high-confidence pixels without masking any of them from the teacher model. However, when compared to Scheme 5, this approach does not result in the expected improvement in optimizing the classifier parameters. Instead, the DC decreases from 86.38% to 86.09%. In Ours, we employ the simplified consistency regularization only on HPPs, excluding UPPs. As a result, DC increases from 86.38% to 86.70%. This demonstrates that applying simplified consistency regularization to UPPs optimizes not only the classifier's parameters but also those before the classifier, leading to conflicts with the optimization introduced by UPPFC.

TABLE IV
IMPROVEMENT OF OTHER SEMI-SUPERVISED METHODS WITH UFC IN DC METRIC ON THE MONUSEG DATASET WITH 1/30 LABELED DATA

Method	DC (%)
MT [1] [NeurIPS'17]	69.28→71.82 (**+2.54**)
UA-MT [20] [MICCAI'19]	72.92→73.20 (**+0.28**)
TCSM_V2 [2] [TNNLS'20]	71.07→73.43 (**+2.36**)
ASE-Net [28] [TMI'23]	71.01→72.21 (**+1.20**)
FixMatch* [17] [CVPR'20]	73.36→74.17 (**+0.81**)

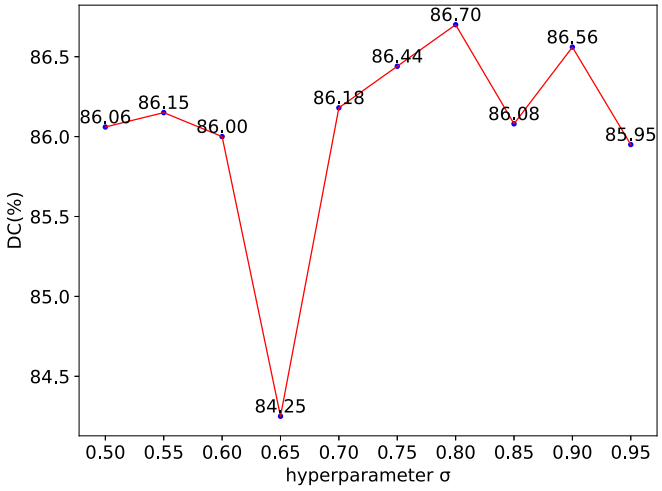


Fig. 6. Quantitative evaluation of CFC with different threshold σ on the ISIC dataset with 259/2594 labeled data.

E. UFC Used in Other Semi-Supervised Methods

In our proposed UFC, strong and weak augmentations [17] are employed as data perturbations. More importantly, UFC is designed as a versatile plug-and-play module, seamlessly integrable into any semi-supervised method with teacher-student models. Therefore, UFC is embedded into other semi-supervised methods, such as MT [1], UA-MT [20], TCSM_v2 [2], and ASE-Net [28], and experiments are conducted on the MoNuSeg dataset with only 1/30 labeled data. Table IV shows the improvement in DC achieved by applying UFC to the mentioned above semi-supervised methods.

V. DISCUSSION

A. Hyperparameter σ and η

Different values of thresholds σ and η in defining valid image/patch are evaluated using 259/2594 labeled data from the ISIC dataset. Figs. 6 and 7 illustrate the effects of varying thresholds on DC. It is evident that DC and thresholds do not exhibit a linear relationship but rather result from a tradeoff between quantity and quality.

As shown in Fig. 6, we initially keep $\eta = 0.9$ and study the impact of different σ values on DC. Specifically, when $\sigma = 0.5$, it indicates the use of CFC without any filtering on images and patches, yielding a DC of 86.06%. The highest-DC observed is at $\sigma = 0.8$, reaching 86.70%, signifying a 0.64% enhancement in DC by selecting valid images and patches. Subsequently, as shown in Fig. 7, with σ fixed at 0.8, we examine the impact of different η values on DC, where the highest DC is achieved at $\eta = 0.9$, reaching 86.70%.

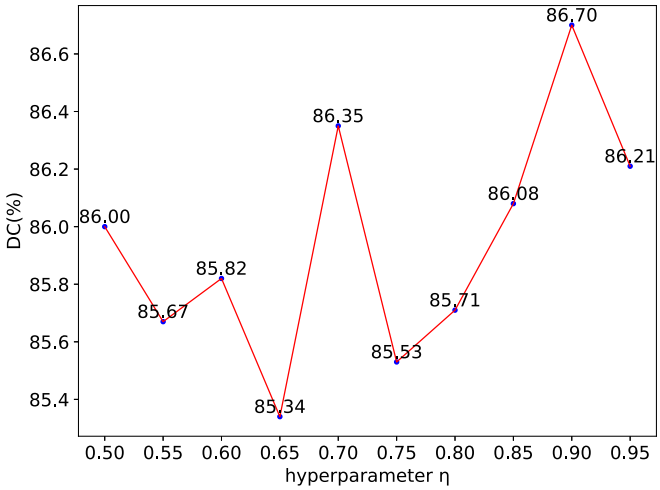


Fig. 7. Quantitative evaluation of CFC with different threshold η on the ISIC dataset with 259/2594 labeled data.

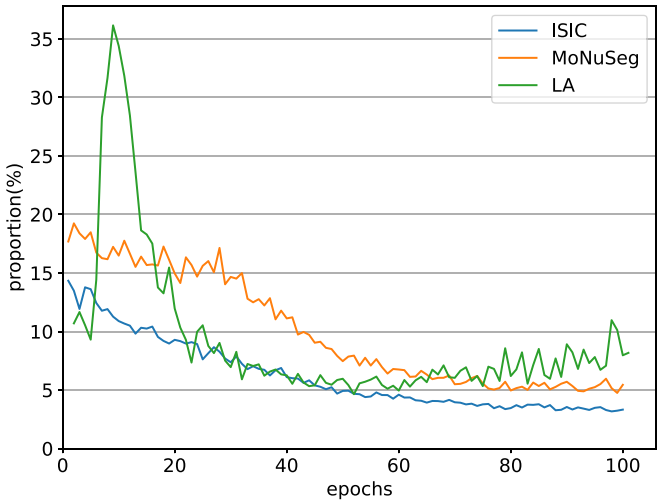


Fig. 8. Proportion curves of UPPs/voxels (UPVs) on the training sets of ISIC, MoNuSeg, and LA datasets with 259/2594, 1/30, and 8/80 labeled data, respectively.

B. Proportion of Under-Performing Pixels/Voxels and Valid Images/Slices

As shown in Fig. 8, the proportion of UPPs and voxels (UPVs) generally decreases with training epochs, stabilizing around 5%. Additionally, even in the early stages of training, the proportion of UPPs/UPVs is not substantial. For the ISIC and LA datasets, we perform downsampling on features to reduce redundancy, utilizing only 1/4 and 1/8 of UPPs/UPVs, respectively. From Section IV-D, it is evident that solely employing UPPFC yields quite satisfactory performance. This indicates that the proposed UPPFC is an effective learning strategy.

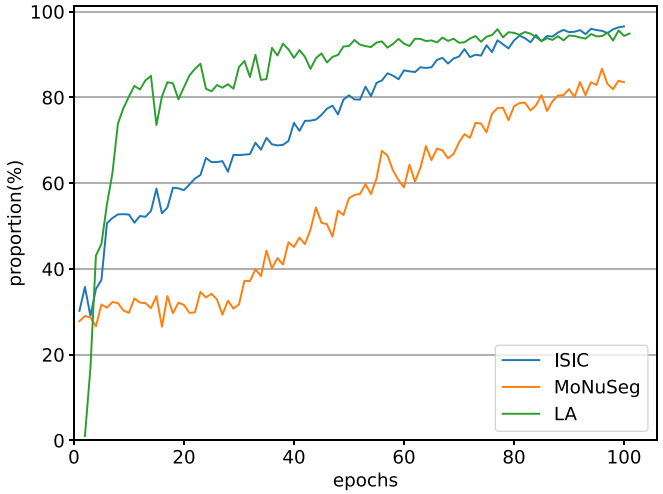


Fig. 9. Proportion curves of valid images/ 2-D slices on the ISIC, MoNuSeg, and LA datasets with 259/2594, 1/30, and 8/80 labeled data, respectively.

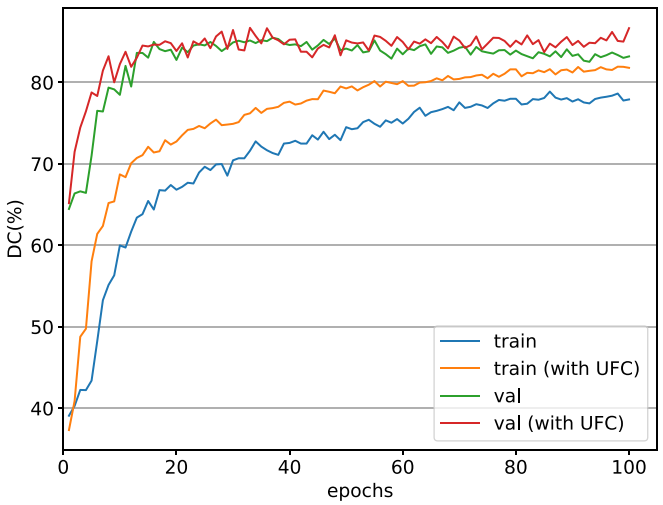


Fig. 10. DC curves of training and validation sets on the ISIC dataset with 259/2594 labeled data.

As shown in Fig. 9, with the increase of training epochs, the proportion of valid images/slices is generally on the rise. Specifically, for the LA dataset, the proportion of valid slices quickly reaches a high ratio, indicating that the vast majority of slices are deemed valid. Conversely, for medical images obtained through optical imaging methods, such as skin lesion and cell nuclei images, due to issues like blurred object boundaries, object occlusions, and artifacts, low-uncertainty segmented pixels are relatively fewer compared to images like CT and MRI. Consequently, the proportion of valid images in the ISIC and MoNuSeg datasets is relatively lower, particularly with images that are more challenging to segment, such as nuclei images. This is also another reason why we use the PCFC with perceptual region adjustment on the ISIC and MoNuSeg datasets.

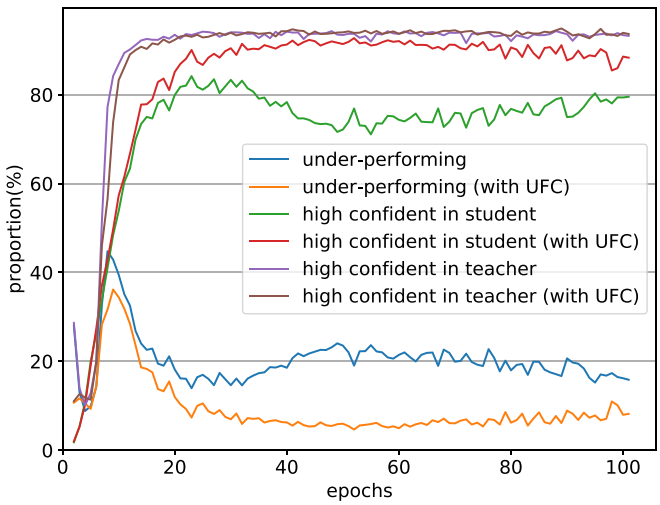


Fig. 11. Proportion curves of under-performing and high-confidence voxels on the LA dataset with 8/80 labeled data.

C. Analysis of UFC

The proposed method is built upon the semi-supervised paradigm of FixMatch. Therefore, we compare the disparities between scenarios with and without the utilization of UFC.

As shown in Fig. 10, we visualize the DC curves of training and validation sets on the ISIC dataset with 259/2594 labeled data. It is worth noting that, the DC is noticeably higher than the original FixMatch after using UFC during the training process. Additionally, while the original FixMatch showed signs of overfitting in the latter stage of training, this tendency disappeared when employing UFC. This indicates that the use of our proposed UFC module enhances the model's generalization ability.

Additionally, as shown in Fig. 11, we visualize the proportion curves of under-performing and high-confidence voxels on the LA dataset with 8/80 labeled data. It is worth noting that after using UFC, there is little impact on the proportion of high-confidence voxels in the teacher model, while there is a significant increase in the proportion of high-confidence voxels in the student model. This suggests that the network is more decisive in segmenting strong augmented images after using UFC, reducing the model's uncertainty.

VI. Conclusion

In this study, we have presented a novel semi-supervised approach for medical image segmentation using UFC. Our approach not only focuses on the local segmentation quality of UPPs, but also considers the global image-level and local patch-level segmentation quality of valid regions by preserving the structured semantic information of medical images, significantly enhancing the ability of feature extraction for the segmentation network. We apply consistency regularization to the HPPs, directly optimizing classifier parameters to achieve higher-quality predictions. This provides guidance for UFC, further improving medical image segmentation effect.

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] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[2] 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 Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[3] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 297–306.

[4] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11514–11524.

[5] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7236–7246.

[6] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, 2022.

[7] Z. Gu et al., "CE-Net: Context encoder network for 2-D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[8] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3-D UX-Net: A large kernel volumetric ConvNet modernizing hierarchical transformer for medical image segmentation," 2022, *arXiv:2209.15076*.

[9] H. Li, Y. Nan, J. Del Ser, and G. Yang, "Large-kernel attention for 3-D medical image segmentation," *Cogn. Comput.*, vol. 16, pp. 2063–2077, Jul. 2024.

[10] T. Lei, R. Sun, X. Du, H. Fu, C. Zhang, and A. K. Nandi, "SGU-Net: Shape-Guided ultralight network for abdominal image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 3, pp. 1431–1442, Mar. 2023.

[11] Y. Lin, X. Fang, D. Zhang, K. T. Cheng, and H. Chen, "A permutable hybrid network for volumetric medical image segmentation," 2023, *arXiv:2303.13111*.

[12] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[13] A. Hatamizadeh et al., "UNETR: Transformers for 3-D medical image segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[14] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327.

[15] X. Lin, Z. Yan, X. Deng, C. Zheng, and L. Yu, "ConvFormer: Plug-and-play CNN-style transformers for improving medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2023, pp. 642–651.

[16] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, "The fully convolutional transformer for medical image segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3660–3669.

[17] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.

[18] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 8801–8809.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[20] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3-D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 605–613.

[21] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.

[22] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2613–2622.

[23] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2017, pp. 408–416.

[24] G. Chen et al., "MTANS: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, Dec. 2021, Art. no. 118568.

[25] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8300–8311.

[26] G. Jin, C. Liu, and X. Chen, "Adversarial network integrating dual attention and sparse representation for semi-supervised semantic segmentation," *Inf. Process. Manag.*, vol. 58, no. 5, 2021, Art. no. 102680.

[27] D. Xu and Z. Wang, "Semi-supervised semantic segmentation using an improved generative adversarial network," *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 9709–9719, 2021.

[28] 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 Trans. Med. Imag.*, vol. 42, no. 5, pp. 1265–1277, May 2023.

[29] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4268–4277.

[30] E. W. Teh, T. DeVries, B. Duke, R. Jiang, P. Aarabi, and G. W. Taylor, "The GIST and RIST of iterative self-training for semi-supervised segmentation," in *Proc. 19th Conf. Robots Vis.*, 2022, pp. 58–66.

[31] Y. Shi et al., "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 608–620, Mar. 2022.

[32] 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 Trans. Med. Imag.*, vol. 42, no. 10, pp. 2832–2841, Oct. 2023.

[33] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022.

[34] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11666–11675.

[35] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102792.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[37] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[38] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[39] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[40] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15750–15758.

[41] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A Challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[42] N. Kumar et al., "A multi-organ nucleus segmentation challenge," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1380–1391, May 2020.

[43] Z. Xiong et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101832.
[44] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.
[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3-D Vision (3DV)*, 2016, pp. 565–571.
[46] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. New York, NY, USA: Springer, 2006, p. 138.
[47] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[48] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[49] Y. Zhao, K. Lu, J. Xue, S. Wang, and J. Lu, "Semi-supervised medical image segmentation with voxel stability and reliability constraints," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 8, pp. 3912–3923, Aug. 2023.
[50] J. Liu et al., "Clip-driven universal model for organ segmentation and tumor detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2023, pp. 21152–21164.
[51] Z. Ding et al., "Exploring structured semantic prior for multi label recognition with incomplete labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3398–3407.
[52] Y. Wang, B. Xiao, X. Bi, W. Li, and X. Gao, "MCF: Mutual correction framework for semi-supervised medical image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15651–15660.
[53] X. Luo et al., "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102517.
[54] Z. Xu et al., "Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102880.
[55] S. Gao, Z. Zhang, J. Ma, Z. Li, and S. Zhang, "Correlation-aware mutual learning for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2023, pp. 98–108.
[56] A. He, T. Li, J. Yan, K. Wang, and H. Fu, "Bilateral supervision network for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 5, pp. 1715–1726, May 2024.
[57] Z. Wang, Z. Zhao, X. Xing, D. Xu, X. Kong, and L. Zhou, "Conflict-based cross-view consistency for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19585–19595.
[58] J. Su, Z. Luo, S. Lian, D. Lin, and S. Li, "Mutual learning with reliable pseudo label for semi-supervised medical image segmentation," *Med. Image Analy.*, vol. 94, May 2024, Art. no. 103111.
[59] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 481–490.
[60] T. Wang, J. Lu, Z. Lai, J. Wen, and H. Kong, "Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation," in *Proc. 31st Int. Joint Conf. Artif. Intell. IJCAI*, 2022, pp. 1444–1450.
[61] H. Xie et al., "PRCL: Probabilistic representation contrastive learning for semi-supervised semantic segmentation," *Int. J. Comput. Vis.*, vol. 132, pp. 4343–4361, May 2024.
[62] T. Lei, H. Liu, Y. Wan, C. Li, Y. Xia, and A. K. Nandi, "Shape-guided dual consistency semi-supervised learning framework for 3-D medical image segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 7, pp. 719–731, Sep. 2023.