Exploring Feature Representation Learning for Semi-Supervised Medical Image Segmentation

Huimin Wu, Xiaomeng Li, *Member, IEEE*, and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*— This article presents a simple yet effective two-stage framework for semi-supervised medical image segmentation. Unlike prior state-of-the-art semi-supervised segmentation methods that predominantly rely on pseudo supervision directly on predictions, such as consistency regularization and pseudo labeling, our key insight is to explore the feature representation learning with labeled and unlabeled (i.e., pseudo labeled) images to regularize a more compact and better-separated feature space, which paves the way for low-density decision boundary learning and therefore enhances the segmentation performance. A stage-adaptive contrastive learning method is proposed, containing a boundary-aware contrastive loss that takes advantage of the labeled images in the first stage, as well as a prototype-aware contrastive loss to optimize both labeled and pseudo labeled images in the second stage. To obtain more accurate prototype estimation, which plays a critical role in prototype-aware contrastive learning, we present an aleatoric uncertainty-aware method to generate higher quality pseudo labels. Aleatoric-uncertainty adaptive (AUA) adaptively regularizes prediction consistency by taking advantage of image ambiguity, which, given its significance, is underexplored by existing works. Our method achieves the best results on three public medical image segmentation benchmarks.**

***Index Terms*— Aleatoric uncertainty, consistency regularization, contrastive learning, pseudo labeling, semi-supervised segmentation.**

I. Introduction

MEDICAL image segmentation is a foundational task for computer-aided diagnosis and computer-aided surgery. In recent years, considerable efforts have been devoted to designing neural networks for medical image segmentation, such as U-Net [1], DenseUNet [2], nnUNet [3], and HyperDenseNet [4]. However, training these models requires a large number of labeled images. Unlike natural images, the professional expertise required for pixelwise manual annotation of medical images makes such labeling tasks challenging and time-consuming, resulting in the difficulty of obtaining a large labeled dataset. Hence, semi-supervised learning, which enables training using labeled and unlabeled data, becomes an active research area for medical image segmentation.

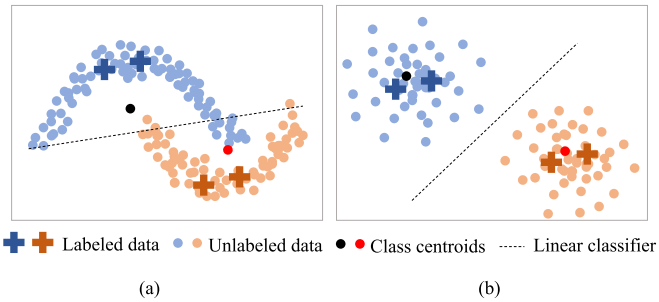

Fig. 1. Two toy examples which (a) visualize the feature space of an indiscriminative semi-supervised model and (b) visualize the feature space of a well-clustered semi-supervised model.

A common assumption of semi-supervised learning is that the decision boundary should not pass through high-density regions. Consistency regularization-based techniques [5], [6], [7] achieve a decision boundary at a low-density area by penalizing prediction variation under different input perturbations. Entropy minimization (Entropy Mini)-based methods aim to achieve high-confidence predictions for unlabeled data either in an explicit manner [8] or an implicit manner [9], [10], [11], [12]. As shown in Fig. 1, an ideal model should pull together data points of the same class and push apart data points from different classes in the feature space. As the training set of semi-supervised learning includes labeled and unlabeled images, it is challenging to directly optimize the unlabeled images in the feature space without explicit guidance. We observe that with unlabeled images, most semi-supervised methods [5], [6], [7] can achieve more accurate segmentation results than the model trained with only labeled data. Therefore, the pseudo segmentation predicted by a semi-supervised model on unlabeled data could possibly be made even more stable and precise.

Motivated by this observation, we present a simple yet effective two-stage framework for semi-supervised medical image segmentation with the key idea to explore representation

Manuscript received 20 June 2022; revised 15 December 2022 and 22 April 2023; accepted 14 July 2023. This work was supported in part by the HKSAR Research Grants Council (RGC) General Research Fund (GRF) under Grant 16203319, in part by the Foshan HKUST Projects under Grant FSUST21-HKUST10E and Grant FSUST21-HKUST11E, and in part by the Shenzhen Municipal Central Government Guides Local Science and Technology Development Special Funded Projects under Grant 2021Szvup139. *(Corresponding author: Xiaomeng Li.)*

Huimin Wu is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: hwubl@connect.ust.hk).

Xiaomeng Li is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China, and also with the Shenzhen Research Institute, The Hong Kong University of Science and Technology, Shenzhen 518057, China (e-mail: eexmli@ust.hk).

Kwang-Ting Cheng is with the Department of Electronic and Computer Engineering and the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: timcheng@ust.hk).

Digital Object Identifier 10.1109/TNNLS.2023.3296652

learning for segmentation from both labeled and unlabeled images. The first stage aims to generate high-quality pseudo labels, and the second stage aims to use pseudo labels to retrain the network to regularize features for both labeled and unlabeled images. Existing uncertainty-based semi-supervised methods [5], [13], [14], [15] have achieved stunning results by considering the reliability of the supervision for the unlabeled images. These methods exploit the epistemic uncertainty, a kind of uncertainty about the model's parameters arising from a lack of data, either in the output space [5], [13], [14] or in the feature space [14], as guidance for identifying trustworthy supervision. Medical images are often noisy, and the boundaries between tissue types may not be well defined, leading to a disagreement among human experts [16], [17], [18]. However, aleatoric uncertainty that represents the ambiguity about the input data and is irreducible by obtaining more data is ignored in these methods.

To obtain high-quality pseudo labels for unlabeled images, we present an aleatoric-uncertainty adaptive (AUA) method for semi-supervised medical image segmentation. Under the framework of the mean teacher (MT) model [19], to obtain reliable target supervision for unlabeled data, instead of estimating the model's epistemic uncertainty [5], [13], [14], we explore the aleatoric uncertainty of the model for noisy input data. AUA first measures the spatially correlated aleatoric uncertainty by modeling a multivariate normal distribution over the logit space. To effectively utilize unlabeled images, AUA encourages the prediction consistency between the teacher model and the student model by adaptively considering the aleatoric uncertainty for each image. Specifically, the consistency regularization automatically emphasizes the input images with lower aleatoric uncertainty, i.e., input images with less ambiguity.

In the second stage, we retrain the network with pseudo labels. To effectively regularize feature representation learning in both stages, we propose stage-adaptive feature regularization, including a boundary-aware contrastive loss (BCL) in the first stage and a prototype-aware contrastive loss (PCL) in the second stage. The main idea of BCL is to fully leverage labeled images for representation learning. A straightforward solution is to pull together the pixels to the same class and push away pixels from different classes using a contrastive loss. However, medical images usually contain a large number of pixels. Simply utilizing contrastive loss would lead to a high computational cost and memory consumption. To this end, we present a BCL, where only randomly sampled pixels from the segmentation boundary are optimized. In the second stage, to effectively utilize both labeled and pseudo-labeled images, i.e., unlabeled images for representation learning, we present a PCL with each pixel's feature pulled closer to its class centroid, i.e., prototype, and pushed further away from the class centroids it does not belong to. The main intuition is that the trained model can generate pseudo labels for unlabeled images in the second stage. Compared with the BCL, the PCL better leverages the pseudo labels, especially those that may not occur at the segmentation boundaries.

In summary, this article makes the following contributions.

1) We introduce stage-adaptive contrastive losses (i.e., BCL and PCL) to regularize a more compact and better-separated feature space, which eases the learning of a segmentation decision boundary.
2) We present AUA, an aleatoric uncertainty adaptive consistency regularization method that paves the way for PCL by improving pseudo label quality and prototype estimation.
3) Our method achieves the state-of-the-art performance on three public datasets. The ablation study validates the effectiveness of our proposed method. Our code is available at GitHub https://github.com/Huiimin5/AUA.

II. Related Work

We briefly discuss related works in semi-supervised medical image segmentation, including pseudo labeling and consistency regularization. We also discuss some techniques related to contrastive learning and uncertainty estimation.

A. Semi-Supervised Medical Image Segmentation

Semi-supervised learning (SSL) refers to training the model with both labeled and unlabeled images. A wide span of tasks has been explored, such as segmentation [7], classification [20], [21], [22], [23], and crowd counting [24]. For medical image segmentation, early work used graph-based methods [25], [26] for semi-supervised segmentation. Recently, semi-supervised medical image segmentation has featured deep learning. The existing methods can be broadly classified into two categories: pseudo labeling-based [9], [12], [27], [28], [29] and consistency regularization-based methods [5], [6], [7], [13], [14], [30], [31], [32], [33], [34], [35], [36], [37].

1) Pseudo Labeling-Based Methods: Pseudo labeling-based methods handle label scarcity by estimating pseudo labels on unlabeled data and using all the labeled and pseudo labeled data to train the model. Self-training is one of the most straightforward solutions [9], [27], [28] and has been extended to the biomedical domain for segmentation [10], [11], [38]. The main idea of self-training is that the model is first trained with labeled data only and then generates pseudo labels for unlabeled data. By retraining the model with both labeled and pseudo labeled images, the model performance can be enhanced. The model can be trained iteratively with these two processes until the model performance becomes stable and satisfactory. To reduce the noise in pseudo labels, different methods have been developed, including identifying trustworthy pseudo labels by uncertainty estimation [10], using a conditional random field (CRF) [39] to refine pseudo labels [38] or using pseudo labels only for fine-tuning [11]. In addition to such an offline pseudo label generation strategy, online self-training methods [12], [40], [41] have been developed recently where pseudo labels are generated after each forward propagation and used as immediate supervision.

Another pseudo labeling-based method is cotraining [42], [43], [44] where multiple learners are trained and their disagreement on unlabeled data is exploited for improving the accuracy of pseudo labels. The basic idea is that each learner

could learn different and complementary information from the other learners. In some self-training methods, more than one learner can be used, such as in [12], and the supervision on unlabeled data is unidirectional. For example, the teacher model [19] generates pseudo labels to supervise the student model, while in a dual-model cotraining method such as [29], supervision is bidirectional. Specifically, each base model's supervision of unlabeled data is based on the fused predictions from the other base models, weighted by the confidence of each model.

However, these methods ignore the class-aware feature regularization, which is a key focus of this study. We will demonstrate the importance of feature representation learning in learning with labeled and pseudo labeled images.

2) Consistency Regularization-Based Methods: The goal of consistency regularization-based semi-supervised methods [19], [45], [46] is to find the model that is not only accurate in predictions but also invariant to input perturbations to enforce the decision boundary traverse the low-density region of the feature space. One line of these methods considers invariance to input domain perturbations. For example, the temporal ensembling model [45] achieves promising results by accumulating soft pseudo labels on randomly perturbed input images. An extension work with soft pseudo label accumulation guided by epistemic uncertainty was proposed in [13]. When the epistemic uncertainty of the prediction is high, it will contribute less to pseudo label accumulation. The MT model [19] achieves invariance to input perturbations by promoting consistency between the predictions of the teacher and the student models where input images fed to the teacher model are added with noises. Extensions have also been made from the perspective of reliability evaluation [5], [14], [15] to provide reliable supervision from the teacher model to the student model or considering structural information of foreground objects [15], [30], [47]. In addition to input domain perturbation, other perturbations that would not change the semantics of the prediction have also been designed and used to promote consistency. For example, consistency among predictions given by differently designed decoders [31], [32], [48], [49] or at different scales [6] or with different modalities [33] where perturbations are beyond the input level is maintained. The distribution-level consistency between predicted segmentation maps on labeled data with those on unlabeled ones [36], [50] has also been proved effective. Aside from perturbations that lead to invariance in output, there is another line of studies [7], [34], [35] that promotes equivariance between the input and the output because some input space transform, especially spatial transform such as rotations, should lead to the same transform in the output space.

Unlike these existing methods that are based on consistency regularization, our method is a two-stage framework, which improves the overall framework by regularizing the feature representation. Moreover, we introduce an aleatoric uncertainty-aware (AUA) method to represent inherent ambiguities in medical images and enhance the segmentation performance by encouraging consistency for images with low ambiguity.

B. Contrastive Learning in Semi-Supervised Image Segmentation

Note that we exclude self-supervised learning methods where unlabeled data are only used for task-agnostic purposes, i.e., pretraining such as in [51], even though performance under semi-supervised setting is also reported. We only consider contrastive learning for task-specific use [52], [53], [54]. Among these works, only [54]'s goal is to promote interclass separation and intraclass compactness. However, in [54], pseudo labels are obtained from the model trained with labeled data only, whose performance is inferior to our first-stage model, where pseudo labels are obtained from a model that takes advantage of consistency regularization on unlabeled data and feature regularization on labeled data. In [53], interclass separation is considered by taking pixels with different pseudo labels as negative pairs, but intraclass compactness is ignored since the positive pair is built on the same pixels from different crops, which essentially is an extension of instance discrimination for the segmentation task. To the best of our knowledge, ours is the first study with pixel-level feature regularization aiming at intraclass compactness and interclass separation for semi-supervised medical image segmentation.

C. Uncertainty Estimation in Semi-Supervised Medical Image Segmentation

Uncertainties generally fall into two categories: epistemic and aleatoric. Epistemic uncertainty is about a model's parameters caused by a lack of data, while aleatoric uncertainty is caused by intrinsic ambiguities or randomness of input data and cannot be reduced by introducing more data. Early methods measure uncertainty using particle filtering and CRFs [55], [56]. More recently, in Bayesian networks, epistemic uncertainty is usually estimated with Monte Carlo dropout [57], which has been extended for the semi-supervised medical image segmentation task [5], [13], [14]. Aleatoric uncertainty is estimated either without considering correlations between pixels [57] or with a limited ability to model spatial correlation since it is captured by uncorrelated latent variables from multivariate normal distribution [16], [17]. Monteiro et al. [18] proposed an aleatoric uncertainty estimation technique where correlations between pixels are considered. Given the ubiquitous existence of noises or ambiguities in medical images, aleatoric uncertainty has been overlooked for semi-supervised medical image segmentation. In this work, we propose an aleatoric uncertainty adaptive consistency regularization technique, where correlations between pixels are considered when measuring aleatoric uncertainty.

III. Method

Fig. 2 visualizes the overview of the proposed two-stage framework. The input image is first fed into the AUA module to get a segmentation model, which generates a high-quality pseudo label. Then, we introduce the stage-adaptive contrastive learning method, consisting of a BCL on labeled data only in the first stage and a PCL on all data in the second stage. By sequential training through the first and second stages, we generate the final segmentation results.

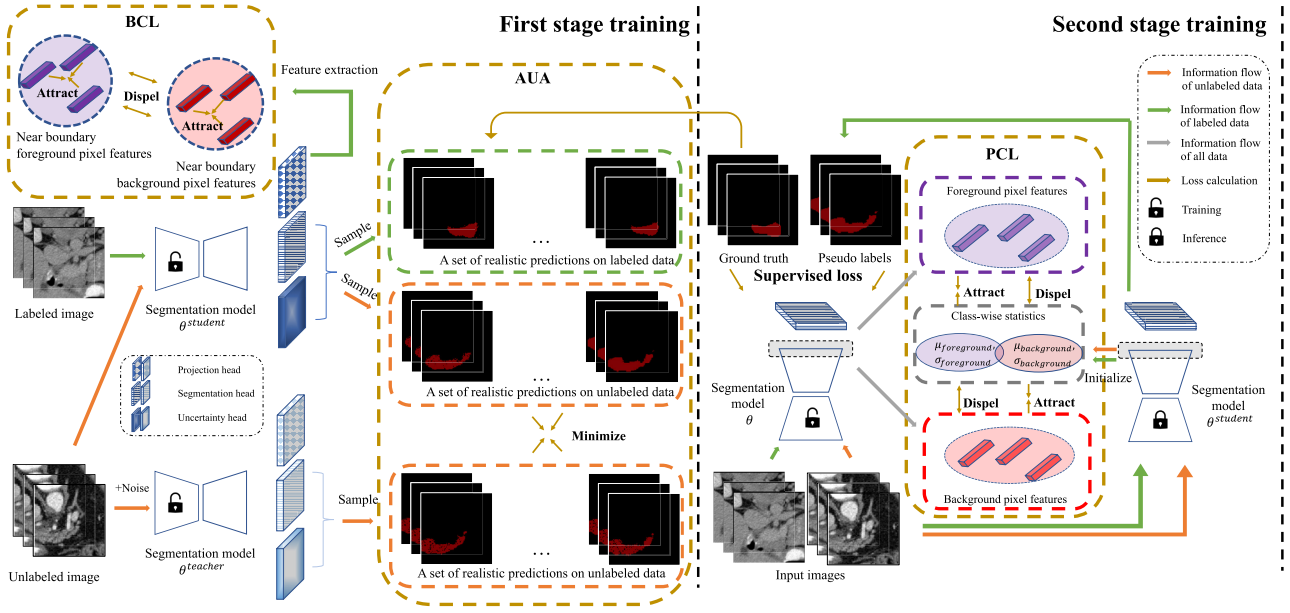

Fig. 2. Overview of our method. The proposed loss functions (AUA, BCL, and PCL) are boxed out in brown. First, we propose AUA under an MT framework, which consists of a student model (parameterized by θ^{student}) and a teacher model (parameterized by θ^{teacher}). AUA is composed of an aleatoric uncertainty estimation module (boxed out in green) and an adaptive consistency regularization module (boxed out in orange). Second, stagewise contrastive learning is proposed, which consists of BCL on boundary pixels of labeled data in the first stage, as well as a PCL, which is applied to all pixels in the second stage.

A. Preliminaries: Deterministic Medical Image Segmentation

Here, we consider a C-class segmentation task on 3-D volumes with size $H \times W \times D \times C$, where H, W, and D denote the height, width, and depth, respectively. Given an image $x \in R^{H\times W\times D\times C}$ and its ground truth y with the same size, the loss function of a general segmentation network is designed to minimize the negative log likelihood, formulated as

$$\mathcal{L}_{\text{sup}} = -\log p(y|x) = -\log \int p(y|g)p(g|x)dg \tag{1}$$

where g denotes logits.

In a deterministic segmentation network, i.e., assuming $p(g|x) = \delta(f(g|x;\theta))$ and independence of each pixel's prediction on the other, where f is a neural network parameterized by θ and δ denotes the Dirac delta function, the loss function in (1) can be rewritten as

$$\mathcal{L}_{\text{sup}_{\text{ce}}} = -\log p(y|g) = -\sum_{i=1}^{V}\sum_{c=1}^{C} y_{ic}\log \text{softmax}(g_i)_c. \tag{2}$$

For simplicity, we use a 1-D scalar i to index each pixel out of a whole set of $V = H * W * D$ pixels in a 3-D volume. Equation (2) is the cross-entropy function commonly used in segmentation models.

B. Preliminaries: Aleatoric Uncertainty Estimation for Segmentation

To model inherent ambiguities of input data, we follow [18] and assume a multivariant Gaussian distribution around logits, i.e., $g|x \sim N(\mu(x), \sigma(x))$, with $\mu(x) \in R^{H\times W\times D\times C}$ and $\sigma(x) \in R^{(H\times W\times D\times C)^2}$. The Monte Carlo integration of S samples is applied to approximate the intractable integral operation, leading us from (1) to

$$\mathcal{L}_{\text{sup}_{\text{au}}} = -\log \frac{1}{S}\sum_{s=1}^{S} p\left(y|g^{(s)}\right) \tag{3}$$

$$= -\text{logsumexp}_{s=1}^{S} \log p\left(y|g^{(s)}\right) + \log(S). \tag{4}$$

The logsumexp (LSE) operation is defined as $\text{LSE}(l_1, \ldots, l_S) = \log(\exp(l_1) + \cdots + \exp(l_S))$ where $l_s = \log p(y|g^{(s)})$. We refer to the calculation of $\log p(y|g^{(s)})$ to (2), where $g^{(s)}$ is a sample out of $g|x \sim N(\mu(x), \sigma(x))$. As pointed out in [18], the full-rank covariance matrix $\sigma(x)$ is computationally infeasible, so we also adopt a low-rank (specifically, r-rank) approximation defined as

$$\sigma(x) = \widetilde{F}\widetilde{F}^T + \widetilde{D} \tag{5}$$

where $\widetilde{F} \in R^{H\times W\times D\times C\times r}$ denotes the factor part of a low-rank form of the covariance matrix and $\widetilde{D} \in R^{H\times W\times D\times C}$ denotes the diagonal part. Compared with a full-rank parameterization, where $(H \times W \times D \times C)^2$ parameters should be estimated, which is beyond what a GPU card can accommodate unless for very small images, its low-rank approximation is more computationally feasible since the complexity reduces from quadratic to linear, i.e., $H \times W \times D \times C \times r$ (for $\widetilde{F}$) $+H \times W \times D \times C$ (for $\widetilde{D}$). This approximation might lead to a compromise of estimated aleatoric uncertainty but still can bring effective guidance for semi-supervised learning, as shown in our experiments.

C. Aleatoric Uncertainty Adaptive Consistency Regularization

It is desirable if the semi-supervised segmentation model can be aware of its chance of making mistakes on unlabeled data. We resort to aleatoric uncertainty that captures input

ambiguities, to guide how much the student model should learn from the teacher model. A consistency regularization technique adaptive to aleatoric uncertainty is proposed.

Given an unlabeled image x^u, the predicted distribution by the student model parameterized by θ^s is denoted as $p_s = p(y^u|x^u;\theta^s)$. Similarly, we can obtain the teacher model's prediction $p_t = p(y^u|x^{u'};\theta^t)$ over the perturbed version of the same input $x^{u'}$ by Gaussian noise injection, where parameters of the teacher model, denoted as θ^t, are updated with an exponential moving average of the parameters of the student model. The consistency between the teacher model's predictions and the student model's predictions on unlabeled data is encouraged by minimizing the generalized energy distance [16], [58], which is defined as

$$\begin{aligned}\mathcal{L}'_{\text{con}} = {} & 2E_{y_s\sim p_s, y_t\sim p_t} d(y_s, y_t) \\ & - E_{y_{s1}\sim p_s, y_{s2}\sim p_s} d(y_{s1}, y_{s2}) \\ & - E_{y_{t1}\sim p_t, y_{t2}\sim p_t} d(y_{t1}, y_{t2}).\end{aligned} \tag{6}$$

To approximate the intractable expectation operation in (6), we take S samples out of p_s and p_t, respectively. The consistency regularization loss function can be reformulated as

$$\begin{aligned}\mathcal{L}_{\text{con}} = {} & 2\sum_{i_s=0}^{S}\sum_{i_t=0}^{S} d\left(y_s^{(i_s)}, y_t^{(i_t)}\right) \\ & - \sum_{i_{s1}=0}^{S}\sum_{i_{s2}=0}^{S} d\left(y_s^{(i_{s1})}, y_s^{(i_{s2})}\right) \\ & - \sum_{i_{t1}=0}^{S}\sum_{i_{t2}=0}^{S} d\left(y_t^{(i_{t1})}, y_t^{(i_{t2})}\right) \\ & y_s^{(i_s)}, \quad y_s^{(i_{s1})}, \ y_s^{(i_{s2})} \sim p_s \\ & y_t^{(i_t)}, \quad y_t^{(i_{t1})}, \ y_t^{(i_{t2})} \sim p_t.\end{aligned} \tag{7}$$

In (7), d is defined as the generalized Dice loss [59]

$$\begin{aligned}& d\left(y^i, y^j\right) \\ & = 1 - \frac{\sum_{k=1}^{H\times W\times D}\sum_{c=1}^{C}\left(y_{kc}^i \cdot y_{kc}^j\right)}{\sum_{k=1}^{H\times W\times D}\sum_{c=1}^{C}\left(y_{kc}^i \cdot y_{kc}^i\right) + \sum_{k=1}^{H\times W\times D}\sum_{c=1}^{C}\left(y_{kc}^j \cdot y_{kc}^j\right)}\end{aligned} \tag{8}$$

where k indexes each pixel out of a whole set of $V = H*W*D$ pixels in a 3-D volume and c indexes each class out of a total of C classes.

The optimum of (7) is 0, which means that the optimum of the first term is the sum of the last two. This consistency regularization metric is adaptive to aleatoric uncertainty in the sense that if the diversity between samples of the student (or the teacher) model is high, i.e., the values of the last two terms of (7) are large, indicating a high aleatoric uncertainty, the pairwise similarity of samples from the student and the teacher models, denoted by the first term of (7), would be less strictly constrained. On the contrary, on input data where a low diversity is estimated, implying the aleatoric uncertainty is low and the model is more likely to generalize well, the student model automatically learns more from the teacher model by optimizing the first term to a smaller value.

To summarize, the AUA loss is defined as follows:

$$\mathcal{L}_{\text{AUA}} = \mathcal{L}_{\text{sup}_{\text{au}}} + \lambda_g \mathcal{L}_{\text{con}} \tag{9}$$

where λ_g is the scaling weight to balance the uncertainty estimation loss and the generalized energy distance loss.

D. Stage-Adaptive Feature Regularization

We introduce a stage-adaptive feature learning method consisting of a BCL and a PCL, to enhance the representation learning with only labeled images and both labeled and pseudo labeled images, respectively. A natural solution is a contrastive loss with features of pixels belonging to the same class (i.e., both foreground pixels or both background pixels) as positive pairs and features belonging to different classes (i.e., one from foreground and the other from background) as negative pairs. This strategy allows us to perform pixelwise regularization but consumes memory quadratically to the number of pixels, so we propose a stage-adaptive contrastive learning method with these concerns properly handled. To reduce the computational cost, at the first stage, we only optimize the feature representation for pixels around the segmentation boundaries, using a BCL. At the second stage, with more accurate pseudo labels on unlabeled data, we introduce a PCL to fully leverage both labeled and pseudo labeled images for representation learning.

1) Boundary-Aware Contrastive Learning: As a balance of benefits of pixelwise feature level regularization and computational costs, we build positive and negative pairs based on a random subset of near-boundary pixels, arriving at the BCL formally defined as

$$\mathcal{L}_{\text{BCL}} = \sum_{i\in NB}\frac{-1}{P(i)}\sum_{pi\in P(i)}\log\frac{\exp\left(f_i^1 \cdot f_{pi}^1/\tau_1\right)}{\sum_{o\in O(i)}\exp\left(f_i^1 \cdot f_o^1/\tau_1\right)} \tag{10}$$

where NB contains indexes of randomly sampled near-boundary pixels from an input image, $O(i)$ contains indexes of the other pixels except pixel i and $P(i)$ contains indexes of pixels in $O(i)$ belonging to the same class as pixel i. The feature vectors f_i^1, f_o^1 and f_{pi}^1 are obtained from a 3-layer convolutional projection head, which is connected after the layer before the last layer. The temperature τ_1 is set to be 0.07. By subsampling, BCL reduces the computational cost from $(H \times W \times D)^2$ (i.e., pixelwise contrastive loss) to NB^2.

2) Prototype-Aware Contrastive Learning: In the second stage, the way to regularize an indiscriminative feature space as in Fig. 1(a) is to encourage each feature to be closer to any other pixels that share the same label and further away from the centroid of opposite class so that forming a feature space in Fig. 1(b) is encouraged, which is defined as

$$\begin{aligned}& \mathcal{L}'_{\text{PCL}} \\ & = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\frac{-1}{|P(i)\backslash\{i\}|} \\ & \quad \times \sum_{pi\in P(i)}\log\frac{\exp\left(f_i^2 \cdot f_{pi}^2/\tau\right)}{\exp\left(f_i^2 \cdot f_{pi}^2/\tau\right) + \frac{1}{|N(i)|}\sum_{ni\in N(i)}\exp\left(f_i \cdot f_{ni}^2/\tau\right)}\end{aligned} \tag{11}$$

where $\mathcal{P}$ contains indexes of all pixels. $P(i)$ and $N(i)$ contains the indexes of positive pixels, i.e., those sharing the same class, and negative pixels, i.e., those with different labels to pixel i, respectively. Features extracted from the second stage model are denoted as f_*^2 where $*$ can be an index of any pixel.

In [60], by assuming a Gaussian distribution for features belonging to each class, the computational cost of (11) can be reduced from quadratic to linear, leading to a regularization formulated as

$$\mathcal{L}_{\text{PCL}} = {f_i^2}^\top \sigma_p f_i^2 - \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\exp\left(\frac{f_i^2\cdot\mu_p}{\tau_2} + \frac{{f_i^2}^\top \sigma_p f_i^2}{2\tau_2^2}\right)}{\exp\left(\frac{f_i^2\cdot\mu_p}{\tau_2} + \frac{{f_i^2}^\top \sigma_p f_i^2}{2\tau_2^2}\right) + \exp\left(\frac{f_i^2\cdot\mu_n}{\tau_2} + \frac{{f_i^2}^\top \sigma_n f_i^2}{2\tau_2^2}\right)} \tag{12}$$

where μ_p and σ_p are the mean and covariance matrix of the positive class to pixel i, respectively, and similarly, μ_n and σ_n are the mean and covariance matrix of the negative class corresponding to pixel i, respectively. These prototype statistics for each class c are estimated from the first-stage model with an moving average update of extracted features with each update at t-step formulated as

$$\mu_t^c = \frac{N_{t-1}^c \mu_{t-1}^c + n_t^c \mu_t^{'c}}{N_{t-1}^c + n_t^c}$$
$$\sigma_t^c = \frac{N_{t-1}^c \sigma_{t-1}^c + n_t^c \sigma_t^{'c}}{N_{t-1}^c + n_t^c} + \frac{N_{t-1}^c n_t^c \left(\mu_{t-1}^c - \mu_t^{'c}\right)\left(\mu_{t-1}^c - \mu_t^{'c}\right)^\top}{\left(N_{t-1}^c + n_t^c\right)^2} \tag{13}$$

where N_{t-1} denotes the total number of pixels belonging to class c seen before time step t and n_t denotes the pixel number of class c in the loaded image at time step t. $\mu_t^{'c}$ and $\sigma_t^{'c}$ denote the mean and covariance, respectively, of features belonging to class c in images at t. The final prototypes are estimated after 3000 iterations and the temperature τ_2 is set to be 100. By utilizing prototypes, BCL reduces the computational cost from $(H \times W \times D)^2$ (i.e., pixelwise contrastive loss) to $H \times W \times D \times C$.

E. Stagewise Training as a Unified Framework

To summarize, in the first stage, the loss function is defined as

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{AUA}} + \lambda_c \mathcal{L}_{\text{BCL}} \tag{14}$$

where λ_c is the scaling weight for BCL loss. To this end, pseudo labels on unlabeled data with higher quality can be obtained thanks to joint prediction regularization (with AUA) and feature regularization (with BCL), which enables retaining a stronger segmentation model at the second stage by regularizing both predictions and features over the whole dataset in a label-aware manner. The loss function in the second stage is given as follows:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{sup}_{\text{ced}}} + \lambda_r \mathcal{L}_{\text{PCL}} \tag{15}$$

where $\mathcal{L}_{\text{sup}_{\text{ced}}}$ is defined as the average of cross-entropy loss and Dice loss as a common practice in segmentation, which serves as pseudo labeling, and λ_r is the weight for PCL loss.

IV. Experimental Results

A. Datasets and Preprocessing

1) Pancreas CT Dataset: Pancreas CT dataset [61] is a public dataset containing 80 scans with a resolution of 512 × 512 pixels and slice thickness between 1.5 and 2.5 mm. Each image has a corresponding pixelwise label, which is annotated by an expert and verified by a radiologist.

2) Colon Cancer Segmentation Dataset: Colon cancer dataset is a subset of Medical Segmentation Decathlon (MSD) datasets [62], consisting of 190 colon cancer CT volumes. Pixel-level label annotations are given on 126 CT volumes. Among these volumes, we randomly split 26 CT volumes as a test set and use the rest for training.

3) Left Atrium MR Dataset: The left atrium (LA) dataset contains 100 magnetic resonance (MR) image scans with an isotropic resolution of 0.625 × 0.625 × 0.625 mm. This dataset is fully annotated with pixel-level supervision, among which 80 scans are used for training and the remaining 20 are used for validation.

4) Preprocessing: To fairly compare with other methods, we follow preprocessing in [33] by clipping CT images to a range of [−125, 275] HU values, resampling images to 1 × 1 × 1 mm resolution, center-cropping both raw images and annotations around foreground area with a margin of 25 voxels, and finally normalizing raw images to zero mean and unit variance. On the Pancreas dataset, we apply random crop as an augmentation on the fly, and the Colon dataset is augmented with random rotation, random flip, and random crop. On both CT datasets, 96 × 96 × 96 subvolumes are randomly cropped from raw data and fed to the segmentation model for training. On the LA dataset, we apply center crop as well as normalize the intensities to zero mean and unit variance for preprocessing. During training, we adopt random crop to 112 × 113 × 80 for on-the-fly augmentation.

B. Implementation Details

1) Environment: All experiments in this work are implemented in Pytorch 1.6.0 and conducted under python 3.7.4 running on an NVIDIA TITAN RTX.

2) Backbone: VNet [63] is used as our backbone where the last convolutional layer is replaced by a 3-D 1 × 1 × 1 convolutional layer. On top of that, a projection module and an aleatoric uncertainty module are built for feature regularization and aleatoric uncertainty estimation, respectively. Similar to [54], the projection head constitutes three convolutional layers, each followed by ReLU activations and batch normalization, except for the last layer, which is followed by a unit-normalization layer. The channel size of each convolutional layer is set as 16. The aleatoric uncertainty module is comprised of three one-layer branches predicting means, covariance factors, and covariance diagonals. Output sampling is implemented by calling the sample function of

TABLE I

COMPARISON WITH STATE-OF-THE-ART ON PANCREAS DATASET WITH 20% LABELED DATA. THE UP ARROW IMPLIES THAT THE LARGER THE NUMBER, THE BETTER THE PERFORMANCE. THE DOWN ARROW IMPLIES THAT A LOWER NUMBER INDICATES A BETTER PERFORMANCE

Method	## scans used		Metrics			
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	ASD[voxel]↓	95HD[voxel]↓
V-Net	12	0	70.63	56.72	6.29	22.54
V-Net	62	0	81.78	69.65	1.34	5.13
MT [19]	12	50	75.85	61.98	3.40	12.59
DAN [50]	12	50	76.74	63.29	2.97	11.13
Entropy Mini [40]	12	50	75.31	61.73	3.88	11.72
UA-MT [5]	12	50	77.26	63.82	3.06	11.90
CCT [49]	12	50	76.58	62.76	3.69	12.92
SASSNet [36]	12	50	77.66	64.08	3.05	10.93
DTC [33]	12	50	78.27	64.75	2.25	8.36
URPC [6]	12	48	78.83	66.01	1.96	6.92
MC-Net [48]	12	48	79.27	66.24	1.90	7.07
Ours	12	48	**79.81**	**66.82**	**1.64**	**5.90**

TABLE II

COMPARISON WITH STATE-OF-THE-ART ON PANCREAS DATASET WITH 5% LABELED DATA

Method	## scans used		Metrics			
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	ASD[voxel]↓	95HD[voxel]↓
V-Net	3	0	30.74	18.84	6.97	26.45
V-Net	60	0	81.46	69.18	1.31	5.09
MT [19]	3	57	31.09	18.77	28.14	59.22
DAN [50]	3	57	46.37	30.84	16.87	42.89
Entropy Mini[40]	3	57	50.71	35.13	16.14	42.45
UA-MT[5]	3	57	34.46	21.24	25.73	57.40
CCT [49]	3	57	43.89	28.72	20.81	52.58
DTC [33]	3	57	48.47	32.71	17.03	42.61
SASSNet [36]	3	57	51.96	36.03	16.08	45.36
MC-Net [48]	3	57	50.26	35.41	11.27	30.05
URPC [6]	3	57	55.00	39.23	**6.11**	**22.40**
Ours	3	57	**56.18**	**40.05**	12.47	34.85

torch.distributions.LowRankMultivariateNormal.[1] The teacher model parameters are updated by taking a moving average of the student model parameters. For BCL, the near-boundary pixels are obtained from the difference set of original foreground pixels and resulting foreground pixels after morphology dilation of 1-pixel radius.

3) Training Details: Our model is trained with a stochastic gradient descent (SGD) optimizer with 0.9 momentum and 0.0001 weight decay for 6000 iterations. A step decay learning rate schedule is applied where the initial learning rate is set to be 0.01 and dropped by 0.1 every 2500 iterations. For each iteration, a training batch containing two labeled and two unlabeled subvolumes is fed to the proposed model, with each subvolume randomly cropped with the size of 96 × 96 × 96 for CT volumes and 112 × 112 × 80 for MR imaging (MRI). On the test set, predictions on subvolumes with the same size using a sliding window strategy with a stride of 16 × 16 × 16 (for CT volumes) or 18 × 18 × 4 (for MRI on LA dataset) are fused to obtain the final results.

4) Evaluation Metrics: We use Dice (DI), Jaccard (JA), the average surface distance (ASD), and the 95% Hausdorff distance (95HD) to evaluate the effectiveness of our semi-supervised segmentation method. DI and JA mainly measure the amount of overlap between output segmentation maps and human annotations. The latter two metrics, ASD and 95HD, measure surface distance and are more sensitive to errors over the segmentation boundary.

[1]https://pytorch.org/docs/stable/distributions.html#torch.distributions.lowrank_multivariate_normal.LowRankMultivariateNormal

C. Results on Pancreas Dataset

1) Our Settings: Since the predictions on unlabeled data may be inaccurate in the early stage of training, we follow common practices [5], [33] and use a Gaussian ramping up function $\lambda_g(t) = 0.15 * e^{-5(1-(t/t_{\max}))^2}$ to control the strength of consistency regularization, where t denotes the current time step and $t_{\max}$ denotes the maximal training step, i.e., 6000 as introduced previously. The constant is used to scale BCL, i.e., λ_c is set to be 0.09 given 20% labeled data and 0.01 given 5% labeled data. In the second stage of training, the PCL weight λ_r is always set to be 0.1.

2) Compared Methods: Table I shows the results on the Pancreas dataset. We compare with recent algorithms, including MT [19], deep adversarial network (DAN) [50], Entropy Mini [40], uncertainty-aware mean teacher (UAMT) [5], cross-consistency training (CCT) method [49], shapeaware adversarial network (SASSNet) [36], dual-task consistency (DTC) [33], uncertainty rectified pyramid consistency (URPC) [6], and mutual consistency network (MC-Net) [48]. Previous methods are mainly benchmarked on the first version of the Pancreas dataset with 12 labeled volumes and 50 unlabeled volumes, where, however, two duplicates of scan #2 are found. In case some of these three samples are in the training set and the rest

TABLE III
COMPARISON WITH STATE-OF-THE-ART ON COLON TUMOR DATASET WITH 5% LABELED DATA

Method	# scans used		Metrics			
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	ASD[voxel]↓	95HD[voxel]↓
V-Net	5	0	34.07	23.09	10.12	26.52
V-Net	100	0	62.31	49.47	2.14	13.49
MT [19]	5	95	38.64	26.38	14.41	33.08
DAN [50]	5	95	34.02	24.14	14.21	32.42
Entropy Mini [40]	5	95	38.62	26.78	18.02	39.01
UA-MT[5]	5	95	40.61	28.01	15.31	34.92
CCT [49]	5	95	43.74	30.64	11.23	26.21
SASSNet [36]	5	95	41.64	30.07	11.93	28.96
DTC [33]	5	95	43.29	29.84	10.62	26.22
MC-Net [48]	5	95	38.71	26.90	12.19	28.52
URPC [6]	5	95	46.43	33.01	9.31	24.57
Ours	5	95	**49.00**	**35.15**	**9.04**	**22.32**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART ON LA DATASET WITH 20% LABELED DATA

Method	# scans used		Metrics			
	Labeled	Unlabeled	Dice[%]↑	Jaccard[%]↑	ASD[voxel]↓	95HD[voxel]↓
V-Net	16	0	86.03	76.06	3.51	14.26
V-Net	80	0	91.14	83.82	1.52	5.75
MT[19]	16	64	88.42	79.45	2.73	13.07
DAN [50]	16	64	87.52	78.29	2.42	9.01
Entropy Mini[40]	16	64	88.45	79.51	3.72	14.14
UA-MT [5]	16	64	88.88	80.21	2.26	7.32
ICT [64]	16	64	89.02	80.34	1.97	10.38
SASSNet [36]	16	64	89.27	80.82	3.13	8.83
DTC [33]	16	64	89.42	80.98	2.10	7.32
Chaitanya et al. [51]	16	64	89.94	81.82	2.66	7.23
SimCVD [47]	16	64	90.85	**83.80**	1.86	6.03
MC-Net [48]	16	64	91.07	83.67	**1.67**	5.84
Ours	16	64	**91.08**	83.67	1.80	**5.60**

are in the test set after a random split, we use version 2 where two duplicates are removed, leaving us the same number of labeled data but 2 less, i.e., 48 unlabeled data. Even under a more strict scenario, our proposed model achieves the best performance among existing works.

3) Results Analysis: The first row, i.e., a fully supervised baseline on the partial dataset, shows the lower bound of the semi-supervised segmentation methods, whereas the second row, i.e., a fully supervised model on a fully labeled dataset, shows the upper bound performance. We can observe that our method achieves 79.81% on Dice, surpassing the current state-of-the-art by 0.54%. Notably, our method is very close to the fully supervised model that employs all volumes supervised by human annotations, showing the effectiveness of the proposed semi-supervised method.

4) More Challenging Setting With 5% Labeled Data: To further validate our method under a more challenging scenario, we reduce the number of labeled data to only 5% and use the rest 95% as unlabeled. As shown in Table II, in such a small-data regime, a performance drop of every semi-supervised learning method is observed compared to its counterpart in a big-data regime in Table I where 20% labeled are available, which confirms common sense. It is observed that our method consistently outperforms other methods. Specifically, our method surpasses all the other semi-supervised methods and outperforms the current state-of-the-art by 1.18% on Dice, which demonstrates that the effectiveness of our method is more obvious in a more challenging setting.

5) Comparison of Computational Cost: Due to a two-stage pipeline, our method takes around 2× hours to finish training compared with existing single-stage works. However, during inference, the proposed method does not introduce heavy computational overhead. Existing works use V-Net as the backbone for inference and their computational time costs are very similar. As mentioned in Section IV-B, we only append one more layer after V-Net and the time cost is very close: 4.70 (ours) versus 4.67 (V-Net), measured by seconds per sample. It means that in practical use, our method is as efficient as existing works.

D. Results on Colon Dataset

Table III shows the results on the Colon dataset. To get a result, we set $\lambda_g(t)$ to be $0.15 * e^{-5(1-(t/t_{\max}))^2}$, and the scaling weight of BCL, i.e., λ_c, is set to be 0.03. In the second stage of training, the weight PCL is set to be 0.1. We compare our method with several state-of-the-art methods using 5% data as labeled and the rest as unlabeled. Again, we tune hyperparameters for previous methods so that these methods can reach the best performance on this dataset. In Table III, by comparing the second row with Table II, we notice that under a fully supervised setting using a full dataset, the performance on the Colon dataset is lower than the Pancreas dataset, indicating

TABLE V

ABLATION STUDY ON THE PANCREAS DATASET. BCL REFERS TO BOUNDARY-AWARE CONTRASTIVE LEARNING AND PCL REFERS TO PROTOTYPE-AWARE CONTRASTIVE LEARNING. PSEUDO LABELING REFERS TO DIRECTLY RETRAINING THE NETWORK WITH PSEUDO LABELS WITHOUT PCL

Method	Metrics			
	Dice[%] ↑	Jaccard[%] ↑	ASD[voxel]↓	95HD[voxel]↓
Supervised baseline	70.63	56.72	6.29	22.54
AUA	76.13	62.19	2.25	9.35
AUA + BCL (First stage)	77.15	63.34	2.04	7.00
AUA + BCL + Pseudo labeling	79.08	65.91	1.91	6.69
AUA + BCL + Pseudo labeling + PCL (Full)	**79.81**	**66.82**	**1.64**	**5.90**

TABLE VI

ABLATION STUDY ON THE COLON DATASET. BCL REFERS TO BOUNDARY-AWARE CONTRASTIVE LEARNING AND PCL REFERS TO PROTOTYPE-AWARE CONTRASTIVE LEARNING. PSEUDO LABELING REFERS TO DIRECTLY RETRAINING THE NETWORK WITH PSEUDO LABELS WITHOUT PCL

Method	Metrics			
	Dice[%] ↑	Jaccard[%] ↑	ASD[voxel]↓	95HD[voxel]↓
Supervised baseline	34.07	23.09	10.12	26.52
AUA	42.74	30.20	15.00	35.43
AUA + BCL (First stage)	43.70	30.92	14.74	33.34
AUA + BCL + Pseudo labeling	46.75	33.62	12.39	28.49
AUA + BCL + Pseudo labeling + PCL (Full)	**49.00**	**35.15**	**9.04**	**22.32**

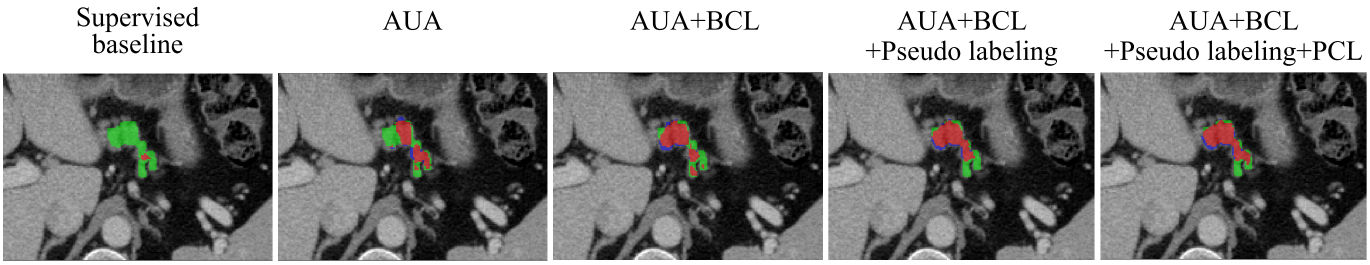

Fig. 3. Visualized ablation study. Regions highlighted in red are true positive areas, i.e., pixels correctly predicted. Green and blue regions represent false negatives, i.e., foreground pixels incorrectly predicted as background, and false positives, i.e., background pixels incorrectly predicted as foreground.

that the Colon dataset is more challenging. By comparing semi-supervised segmentation methods with a fully supervised setting using the partial dataset, i.e., the result in the first row of Table III, we observe stronger performance, showing that leveraging unlabeled data can improve the segmentation performance, which confirms common sense. Our method achieves superior performance compared with all previous works by a large margin (3.43%), which indicates that our method can make better use of unlabeled data.

E. Results on LA Dataset

To demonstrate that our method is generalizable to different medical image modalities, we also conduct comparative experiments on the LA dataset, as shown in Table IV. To get a result, we set $\lambda_g(t)$ to be $0.15 * e^{-5(1-(t/t_{\max}))^2}$, and the scaling weight of BCL, i.e., λ_c, is set to be 0.09. In the second stage of training, the weight PCL is set to 0.1. We compare with state-of-the-arts benchmarked in [47] using 20% data as labeled and the rest as unlabeled. As a sanity check, all semi-supervised methods can outperform a fully supervised baseline on the partial dataset (i.e., 16 labeled MR images), demonstrating a meaningful utilization of unlabeled data. In Table IV, compared with previous works, the proposed method achieves the best results, closing the performance gap with a fully supervised upper bound (i.e., trained with a full dataset).

F. Ablation Studies

Here, we ablate each component of our proposed framework on the Pancreas dataset with 20% as labeled (Table V) and on the Colon dataset with 5% as labeled (Table VI). We gradually add our proposed component and showcase their performances in terms of four metrics.

First, we validate the effectiveness of the adaptive supervision fitting scheme: AUA. On both datasets, as shown in the second row of Tables V and VI, applying AUA achieves superior performance over the fully supervised model, i.e., V-Net. This performance gain mainly comes from its ability to identify and adaptively learn from trustworthy supervision. Specifically, AUA automatically estimates the aleatoric uncertainty of input data and downweights supervision from low-quality images so that the student model learns from more accurate supervision of the teacher model. Second, we demonstrate the effectiveness of our stage-aware contrastive learning. BCL can boost the performance on top of AUA further by a margin of 1.02% and 0.96% on Pancreas and Colon datasets, respectively, and PCL improves its baseline (as shown in the fourth row) by 0.73% and

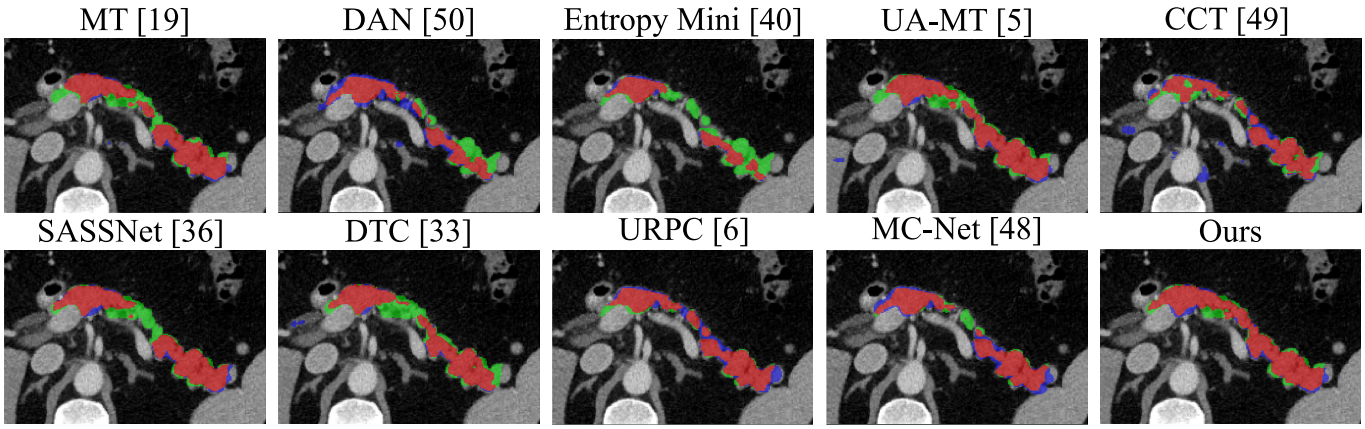

Fig. 4. Comparison between visualized segmentation maps obtained by the state-of-the-art and our method. Regions highlighted in red are true positive areas, i.e., pixels correctly predicted. Regions highlighted in green and blue are false negatives, i.e., foreground pixels incorrectly predicted as background, and false positives, i.e., background pixels incorrectly predicted as foreground.

TABLE VII

ABLATION STUDY ON THE EFFECT OF HYPERPARAMETER λ_c ON THE PANCREAS DATASET

Method	Metrics			
	Dice[%] ↑	Jaccard[%] ↑	ASD[voxel]↓	95HD[voxel]↓
0.03	79.26	66.03	1.84	6.45
0.05	79.61	66.49	2.03	7.53
0.07	79.45	66.31	2.04	6.44
0.09	**79.81**	**66.82**	**1.64**	**5.90**

TABLE VIII

ABLATION STUDY ON AN INCREASING NUMBER OF UNLABELED DATA

#Unlabeled	Metrics			
	Dice[%]↑	Jaccard[%]↑	ASD[voxel]↓	95HD[voxel]↓
12 (1/4)	76.77	62.95	1.76	8.47
24 (1/2)	78.63	65.37	1.81	6.79
48 (Full)	**79.81**	**66.82**	**1.64**	**5.90**

2.25%, respectively. Both techniques are designed to pull features of pixels belonging to the same class closer and push features belonging to opposite classes further, which shapes a more compact (inside each class) and better-separated (across different classes) feature space, leading to a more robust and effective semi-supervised method.

To get a qualitative sense of the effectiveness of each component of our pipeline, in Fig. 3, we plot segmentation results by gradually adding the proposed technique. We illustrate a failure case of the baseline model where the foreground pixels are mislabeled as background. Adding our technique one by one is able to recall more foreground pixels and output segmentation maps with gradually larger overlap with the ground truth.

In addition, we ablate the effect of λ_c, the weight balancing AUA loss, and the BCL loss in the first stage of training. Its values are chosen from 0.03, 0.05, 0.07, and 0.09. Table VII shows that the final results are robust to various choices of λ_c.

Finally, we ablate the robustness of the proposed method to the amount of unlabeled data. In Table VIII, we demonstrate the performance of our method by increasing the unlabeled data number from a small split (i.e., 1/4) to full. We can observe a growing trend in performance, and it is safe to conjecture that with more in-domain unlabeled data, our method can obtain extra performance gain.

G. Qualitative Comparison With the State-of-the-Arts

We visualize the segmentation predictions obtained from other state-of-the-art methods and ours in Fig. 4. We highlight true positive, false negative, and false positive pixels in red, green, and blue, respectively. We can observe that for the other state-of-the-art works, they either achieve a lower recall, such as MT [19], DAN [50], Entropy Mini [40], UA-MT [5], SASSNet [36], DTC [33], and MC-Net [48], or suffer from more false positives, such as DAN [50] and URPC [6]. However, the prediction of our method has a greater overlap with the ground truth.

V. DISCUSSION

In this article, to alleviate heavy reliance on pixelwise labels, which requires considerable human efforts, we propose a novel semi-supervised learning method by taking advantage of aleatoric uncertainty estimation and exploring feature representation learning. First, on top of the MT framework, we present AUA that estimates each image's aleatoric uncertainty and automatically downweights supervision on ambiguous images so that the trained model is able to generate more reliable pseudo labels. However, image ambiguity is an underexplored aspect in semi-supervised learning. Second, we explore representation learning and propose a state-adaptive contrastive learning method. In the first stage, a BCL is designed to regularize labeled image features, and in the second stage, we use a PCL to regularize both labeled and unlabeled features. Superior performance across Pancreas-CT, Colon cancer, and LA datasets validates the superior performance of our method as well its generality to different data modalities.

This study has comprehensive applicability. First, the proposed method can be used to automate downstream diagnosis. It is found in recent research [65], [66] that combining segmentation results benefits disease diagnosis classification tasks. The proposed method allows for training a segmentation model that achieves satisfactory results without relying on

large-scale human labels and thus can be applied to downstream disease diagnosis tasks. Second, in clinical practice, our method can serve as another expert for doctors' reference. For example, a doctor may take into consideration colon cancer segmentation results of the proposed method and make a better diagnosis of cancer staging.

The main limitation of this study is lacking an automatic mechanism to differentiate incorrect pseudo labels from correct ones in the second stage. In this work, we put more effort into generating more accurate pseudo labels prior to their use. However, given pseudo labels, how to make better use is also a nontrivial question. Online confidence thresholding [67], [68] can be a potential solution to identify noisy pseudo labels out of clean ones. In addition, developing a more noisy label-tolerant loss function on our design could also get an extra performance gain.

VI. Conclusion

This article presents a simple yet effective two-stage framework for semi-supervised medical image segmentation, with the key idea of exploring the feature representation from labeled and unlabeled images. We propose a stage-adaptive contrastive learning method, including a BCL and a PCL. In the first stage, the BCL loss regularizes the features by pulling features sharing the same labels closer and pushing features with opposite labels further, arriving at a more compact and well-separated feature space. This loss function, together with AUA, which adaptively encourage consistency by considering the ambiguity of medical images, enhances pseudo label quality after the first stage of training. Improved pseudo labels not only provide higher quality supervision for the segmentation head but also generate more accurate prototypes, which allows PCL to regularize a well-separated feature space further. Specifically, the feature of each pixel is pulled closer to its class centroid and pushed away from its opposite class centroid, which translates to a more accurate segmentation model. Our method achieves the best results on three public medical image segmentation benchmarks, and the ablation study validates the effectiveness of our proposed method. Our future works include extending this work to different types of medical data, such as X-ray images, fundus images, and surgical videos.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[2] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[3] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2020.

[4] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[5] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*. Cham, Switzerland: Springer, 2019, pp. 605–613.

[6] X. Luo et al., "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Proc. MICCAI*, 2021, pp. 318–329.

[7] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[8] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[9] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn. (ICML)*, vol. 3, 2013, pp. 1–6.

[10] S. Sedai et al., "Uncertainty guided semi-supervised segmentation of retinal layers in OCT images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 282–290.

[11] D.-P. Fan et al., "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[12] S. Reiß, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9527–9537.

[13] 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 Trans. Med. Imag.*, vol. 40, no. 1, pp. 431–443, Jan. 2021.

[14] Y. Wang et al., "Double-uncertainty weighted method for semi-supervised learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent* Cham, Switzerland: Springer, 2020, pp. 542–551.

[15] K. Wang et al., "Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102447.

[16] S. Kohl et al., "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6965–6975.

[17] C. F. Baumgartner et al., "PHiSeg: Capturing uncertainty in medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent*. Cham, Switzerland: Springer, 2019, pp. 119–127.

[18] M. Monteiro et al., "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12756–12767.

[19] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, pp. 1195–1204, 2017.

[20] Y. Wang, S. Chen, and Z.-H. Zhou, "New semi-supervised classification method based on modified cluster assumption," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 689–702, May 2012.

[21] Y. Duan et al., "MutexMatch: Semi-supervised learning with mutex-based consistency regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 19, 2022, doi: 10.1109/TNNLS.2022.3228380.

[22] Y. Yang et al., "Semi-supervised multiscale dynamic graph convolution network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 21, 2022, doi: 10.1109/TNNLS.2022.3212985.

[23] M. Gong, H. Zhou, A. K. Qin, W. Liu, and Z. Zhao, "Self-paced co-training of graph neural networks for semi-supervised node classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 21, 2022, doi: 10.1109/TNNLS.2022.3157688.

[24] P. Zhu, J. Li, B. Cao, and Q. Hu, "Multi-task credible pseudo-label learning for semi-supervised crowd counting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 8, 2023, doi: 10.1109/TNNLS.2023.3241211.

[25] H. Su, Z. Yin, S. Huh, T. Kanade, and J. Zhu, "Interactive cell segmentation based on active and semi-supervised learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 762–777, Mar. 2016.

[26] M. Borga, T. Andersson, and O. D. Leinhard, "Semi-supervised learning of anatomical manifolds for atlas-based segmentation of medical images," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3146–3149.

[27] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10684–10695.

[28] R. Ke, A. I. Aviles-Rivero, S. Pandey, S. Reddy, and C.-B. Schönlieb, "A three-stage self-training framework for semi-supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 1805–1815, 2022.

[29] Y. Xia et al., "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101766.

[30] W. Hang et al., "Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 562–571.

[31] K. Fang and W.-J. Li, "DMNet: Difference minimization network for semi-supervised segmentation in medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 532–541.

[32] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," 2021, *arXiv:2103.02911*.

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

[34] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," in *Proc. BMVC*, 2018.

[35] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 810–818.

[36] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 552–561.

[37] H. Yang, C. Shan, and A. F. Kolen, "Deep Q-network-driven catheter segmentation in 3D US by hybrid constrained semi-supervised learning and dual-UNet," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 646–655.

[38] W. Bai et al., "Semi-supervised learning for network-based cardiac mr image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2017, pp. 253–260.

[39] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 109–117.

[40] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[41] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*. Cham, Switzerland: Springer, 2020, pp. 614–623.

[42] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Jul. 1998, pp. 92–100.

[43] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. Eur. Conf. Comput. Vis. (eccv)*, pp. 135–152, 2018.

[44] D.-D. Chen, W. Wang, W. Gao, and Z.-H. Zhou, "Tri-Net for semi-supervised deep learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2014–2020.

[45] L. Samuli and A. Timo, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, vol. 4, 2017, p. 6.

[46] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[47] 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.

[48] Y. Wu et al., "Mutual consistency learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102530.

[49] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12671–12681.

[50] 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. Intervent*. Cham, Switzerland: Springer, 2017, pp. 408–416.

[51] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[52] J. Iwasawa, Y. Hirano, and Y. Sugawara, "Label-efficient multi-task segmentation using contrastive learning," in *Proc. Int. MICCAI Brainlesion Workshop*. Lima, Peru, Oct. 2020, pp. 101–110.

[53] X. Lai et al., "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1205–1214.

[54] X. Zhao et al., "Contrastive learning for label-efficient semantic segmentation," 2020, *arXiv:2012.06985*.

[55] A. Blake, R. Curwen, and A. Zisserman, "A framework for spatiotemporal control in the tracking of visual contours," *Int. J. Comput. Vis.*, vol. 11, no. 2, pp. 127–145, Oct. 1993.

[56] X. He, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 2.

[57] 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, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5574–5584.

[58] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *J. Stat. Planning Inference*, vol. 143, no. 8, pp. 1249–1272, Aug. 2013.

[59] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.* Cham, Switzerland: Springer, 2017, pp. 240–248.

[60] S. Li et al., "Semantic distribution-aware contrastive adaptation for semantic segmentation," 2021, *arXiv:2105.05013*.

[61] H. R. Roth et al., "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 556–564.

[62] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*.

[63] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[64] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3635–3641.

[65] J. Wu et al., "SeATrans: Learning segmentation-assisted diagnosis model via transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2022, pp. 677–687.

[66] Y. Zhou et al., "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2074–2083.

[67] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 8, 2021, pp. 6912–6920.

[68] L.-Z. Guo and Y.-F. Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 8082–8094.

Huimin Wu received the bachelor's and master's degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Hong Kong University of Science and Technology (HKUST), Hong Kong, with a focus on medical image segmentation and label-efficient learning, including semi-supervised and self-supervised learning.

She is broadly interested in computer vision, medical imaging, machine learning, and artificial intelligence.

Xiaomeng Li (Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2019.

She is currently an Assistant Professor with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong. Prior to joining HKUST, she worked as a Post-Doctoral Researcher at Stanford University, Stanford, CA, USA. Her research focuses on the intersection of artificial intelligence and medical image analysis, with the overarching goal of leveraging machine intelligence to advance healthcare.

Kwang-Ting Cheng (Fellow, IEEE) is currently the Vice-President for Research and Development at The Hong Kong University of Science and Technology (HKUST), Hong Kong, where he is also a Chair Professor with the Department of Electronic and Computer Engineering and the Department of Computer Science and Engineering.