Boundary-Guided Contrastive Learning for Semi-Supervised Medical Image Segmentation

Yang Yang, Jiaxin Zhuang, Guoying Sun, Ruixuan Wang, and Jingyong Su

***Abstract*— Semi-supervised learning methods, compared to fully supervised learning, offer significant potential to alleviate the burden of manual annotations on clinicians. By leveraging unlabeled data, these methods can aid in the development of medical image segmentation systems for improving efficiency. Boundary segmentation is crucial in medical image analysis. However, accurate segmentation of boundary regions is under-explored in existing methods since boundary pixels constitute only a small fraction of the overall image, resulting in sub-optimal segmentation performance for boundary regions. In this paper, we introduce boundary-guided contrastive learning for semi-supervised medical image segmentation (BoCLIS). Specifically, we first propose conservative-to-radical teacher networks with an uncertainty-weighted aggregation strategy to generate higher quality pseudo-labels, enabling more efficient utilization of unlabeled data. To further improve the performance of segmentation in boundary regions, we propose a *boundary-guided patch sampling* strategy to guide the framework in learning discriminative representations for these regions. Lastly, the *patch-based contrastive learning* is proposed to simultaneously compute the (dis)similarities of the discriminative representations across intra- and inter-images. Extensive experiments on three public datasets show that our method consistently outperforms existing methods, especially in the boundary region, with DSC improvements of 20.47%, 16.75%, and 17.18%, respectively. A comprehensive analysis is further performed to demonstrate the effectiveness of our approach. Our code is released publicly at https://github.com/youngyzzZ/BoCLIS.**

***Index Terms*— Semi-supervised learning, image segmentation, contrastive learning, mean teacher network.**

Received 26 November 2024; revised 20 February 2025; accepted 25 March 2025. Date of publication 1 April 2025; date of current version 2 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62376068, Grant U24A20340, and Grant 62071502; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023B1515120065; in part by Guangdong S&T programme under Grant 2023A0505050109; in part by Shenzhen Science and Technology Innovation Program under Grant JCYJ20220818102414031; and in part by Guangdong Excellent Youth Team Program under Grant 2023B1515040025. Recommended by Associate Editor C. Petitjean. *(Corresponding authors: Jingyong Su; Ruixuan Wang.)*

Yang Yang, Guoying Sun, and Jingyong Su are with the School of Computer Science and Technology, Harbin Institute of Technology at Shenzhen, Shenzhen 518055, China (e-mail: yangy@stu.hit.edu.cn; sunguoying@stu.hit.edu.cn; sujingyong@hit.edu.cn).

Jiaxin Zhuang is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: jzhuangad@cse.ust.hk).

Ruixuan Wang is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: wangruix5@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TMI.2025.3556482

I. Introduction

AUTOMATED and precise segmentation of organs, lesions, and tissues is of paramount importance in medical diagnosis and understanding disease progression, as it provides clinicians with valuable and insightful information. In recent years, supervised learning approaches empowered by neural networks have emerged as powerful tools, achieving state-of-the-art performance in various medical image segmentation tasks [1], [2], [3]. This success can largely be attributed to the availability of large-scale annotated datasets. However, the acquisition of pixel-wise annotations on a large scale is labor-intensive, expensive, and demands specialized expertise. Therefore, it is crucial and practically relevant to develop methods that can alleviate these requirements.

Semi-supervised learning methods [4], [5], [6], [7], [8], [9] offer promising directions in addressing this challenge, as they require only a minimal amount of annotations and generate pseudo-labels for a significant portion of unlabeled data. These pseudo-labels are then utilized to train the segmentation network, effectively reducing the reliance on expensive pixel-wise annotations and enabling more efficient and cost-effective training. Previous semi-supervised learning approaches have made significant advancements in improving the accuracy of medical image segmentation, with consistency regularization based [10], [11] and pseudo-label based [12], [13], [14] methods emerging as the two main streams. For consistency regularization based approaches, it relies on the assumption that model predictions should remain consistent under various perturbations [10], [11], such as data augmentation or feature perturbation. To enforce this consistency, common techniques involve minimizing the mean square error or the Kullback-Leibler divergence between the outputs obtained from different perturbations [15]. By minimizing the differences of output labels, these approaches ensure a consistent prediction across different perturbations. Alternatively, pseudo-label based methods [16], [17] involve initializing the model with labeled data to generate initial predictions for unlabeled data. The annotated labels from the labeled data, along with the pseudo-labels generated from the predictions of the previous iteration, are then used as ground-truths to iteratively update the network. The pseudo-label generation is periodically updated as the model is trained, with the expectation that label quality will be progressively improved during the training process. However, pseudo-label based methods still face the challenge that the initial model is heavily based on a limited amount of labeled data for initialization, resulting

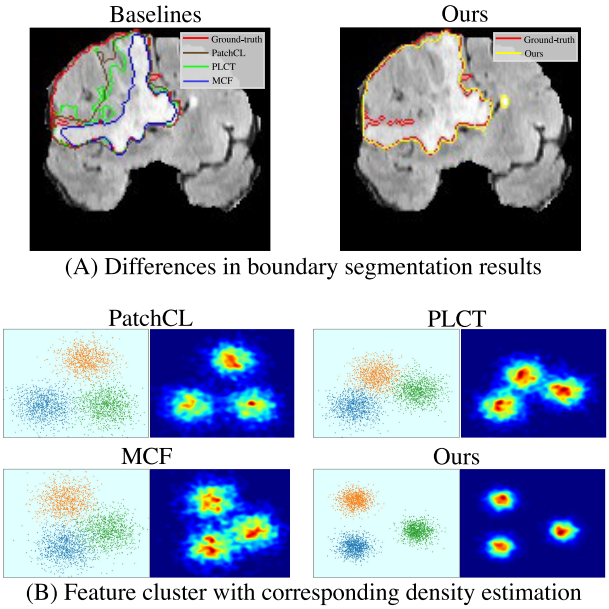

Fig. 1. Motivation of our proposed boundary-guided contrastive learning. (A) Exemplary segmentation results, marked in different colors, are presented to highlight the differences for the challenging boundary region on the BraTS2020 dataset. (B) Illustration of feature clusters with density estimation on ACDC dataset. Features are extracted from the model before the classification layer without accessing true labels. We randomly select 1000 pixel points for each class to perform t-SNE [25] with a perplexity parameter of 32.

in unsatisfactory segmentation performance [18]. This limited segmentation ability can lead to the generation of low-quality pseudo-labels, which subsequently hinders the training process and impedes the improvement of model performance [19]. While some efforts [14], [16], [20] have been made to incorporate uncertainty estimates to mitigate the adverse effects caused by the poor quality of pseudo-labels, there remains significant potential for further improvement in this regard.

In semi-supervised learning, there is an additional challenge that has yet to be resolved. While current methods trained with limited annotations perform well in segmenting the main parts of foreground regions, they are susceptible to generating misclassified predictions in boundary regions, preventing further improvements in segmentation performance [21]. As shown in Fig. 1 (A), current state-of-the-art semi-supervised methods, such as PatchCL [22], PLCT [9] and MCF [23], fail in accurate segmentation of boundaries. This situation is attributed to the severe scarcity of finely labeled data in semi-supervised learning training datasets, where the boundary region pixels represent only a minuscule proportion of total labeled pixels [24]. Fig. 1 (B) reveals that significant density variations persist within each cluster, indicating varying levels of learning difficulty among representations. Current methods face challenges in effectively learning the intricate details of the boundary regions.

Motivated by the above analysis, we propose a novel boundary-guided contrastive learning framework for medical image segmentation (BoCLIS). We have made efforts from two aspects. We propose conservative-to-radical teacher networks to improve the overall segmentation performance of the model and introduce boundary-guided contrastive learning to improve the model's segmentation performance in boundary regions. Specifically, we design conservative-to-radical teacher networks with uncertainty-weighted aggregation to improve the quality of pseudo-labels. For the teacher networks, we assign a gradually increasing momentum for EMA updates, allowing the update paces to shift from conservative to radical and resulting in slightly different predictions. These predictions, along with their corresponding pixel-/voxel-wise uncertainty, are utilized to generate a more reliable aggregation label for the student network by employing an uncertainty-weighted aggregation strategy. To achieve accurate segmentation results in boundary regions, a new sampling strategy, termed boundary-guided patch sampling, is introduced to contrastive learning. This sampling strategy selects patches with the guidance of target's boundary, directing the attention of the network towards boundary regions. The designed contrastive loss function further encourages our framework to learn the (dis)similarity of patch-level representative features in hidden space.

The main contributions of this work are summarized below:

- We develop an innovative semi-supervised segmentation scheme by designing a conservative-to-radical teacher learning with the uncertainty-weighted strategy. Unlike previous works with setting varied initialization or architecture to generate model diversity, our framework maintains identical initialization and architecture across teacher networks while implementing conservative-to-radical momentum coefficients in the EMA process. This framework maintains the teacher model diversity across the update iterations and significantly improves the quality of pseudo-labels generated by the teacher networks, allowing for more efficient utilization of both labeled and unlabeled data.
- We propose an efficient boundary-guided patch sampling strategy that leverages the locations and uncertainty estimates of patches to direct the framework in learning discriminative representations for these regions. Additionally, we introduce a patch-level boundary-guided contrastive learning approach that facilitates the learning of representations between boundary and non-boundary regions, as well as local and global information, guided by the target's boundary. To the best of our knowledge, this is the first endeavor specifically aimed at addressing the challenge of semi-supervised boundary segmentation.
- We demonstrate our method is widely applicable to both 2D and 3D datasets for multi-class segmentation of lesions and organs. Compared to existing methods, our approach achieves state-of-the-art performance across various evaluation metrics, particularly in the accurate segmentation of boundary regions.

II. Related Work

Semi-supervised learning has emerged as a prominent and continuously evolving research area in recent decades, with particular relevance to medical image segmentation. In this section, we adhere to discuss and present a comprehensive review of the literatures that are highly relevant to our work.

A. Semi-Supervised Learning

1) Consistency Regularization Based: The fundamental concept behind consistency regularization is that predictions should be robust when subjected to various perturbations of the input samples [10]. The procedure involves introducing data or feature augmentations as well as stochastic perturbations to the input, while maintaining the consistency of the predictions. In the studies conducted by Li et al. [26], shape constraint is explored through the introduction of the signed distance map. Later, Luo et al. [15], [27] and Wu et al. [17] achieve consistency regularization by designing special auxiliary tasks or a mutual consistency network along with perturbations. Similar studies [28], [29] introduce perturbations to generate slightly varied predictions and further promote their consistency through the meticulous design of the network architecture. Gao et al. [30] designs an Omni-Correlation Consistency Module (OCC). The OCC module establishes omni-correlations between the labeled and unlabeled datasets and enforces consistency by regulating the omni-correlation matrix of each sub-model. Chen et al. [31] further presents decoupled consistency and begins to notice the importance of uncertainty in pseudo-labels. Yet, it only uses the uncertainty map as a threshold for selecting pseudo-labels and fails to fully exploit the information embedded within the uncertainty map.

2) Pseudo-Label Based: In the early stages, Lee [32] initialized a model with a limited amount of labeled data and used it to generate pseudo-labels for a large scale of unlabeled data. Laine and Aila [33] further proposes a temporal ensembling mechanism that updates the pseudo-labels by exponential moving average (EMA) to improve their quality. Later, Yu et al. [16] introduces an uncertainty-weighted mean teacher (UAMT) approach that utilizes transformation consistency to improve performance. Wang et al. [34] and Bai et al. [35] further propose the mean teacher framework that incorporates auxiliary tasks to facilitate the learning of distinctive features, resulting in improved predictions. Shi et al. [36] proposed a student network with two decoders, which is different from most existing approaches. Each decoder applies varying levels of penalties to misclassified background regions to enhance model performance. Later, Miao [37] et al. introduce a self-correcting co-training scheme improve target predictions, aligning them more closely with ground-truth labels through collaborative network outputs. Other methods [16], [20], [38], with the aim of improving the quality of pseudo-labels, introduce uncertainty or confidence estimation to generate more reliable pseudo-labels. However, these methods solely depend on the confidence prediction and lack the necessary informativeness to provide reliable guidance. This deficiency is precisely what our proposed work aims to address.

B. Semi-Supervised Learning With Contrastive Learning

With the remarkable performance of contrastive learning [39], [40], [41] in supervised learning, numerous semi-supervised learning approaches [42], [43], [44] have started to leverage its benefits. Alonso et al. [45] employs a teacher-student network and uses both entropy and contrastive loss in pseudo-labels derived from unlabeled images. Zhuo et al. [46] and You et al. [47] also employ contrastive loss using teacher-student networks in a similar manner. Later, Chaitanya et al. [9] and Basak and Yin [22] propose pixel-based and patch-level contrastive learning methods, respectively. However, their proposed contrastive learning approaches face the challenge of effectively learning discriminative features without careful selection of representative regions. Therefore, we design a boundary-guided patch sampling strategy to encourage our framework to learn more comprehensive semantic representations.

C. Mean Teacher Network

EMA is widely used as an updated approach to ensemble model weights, resulting in smoother and more stable model weights [13], [16], [33], [34]. For instance, Tarvainen and Valpola [13] proposes the use of EMA to integrate the model weights obtained from different prior networks into a singular ensemble model. While the EMA method can achieve a more stable teacher model, it will cause inevitable useful information loss during the aggregation process. Laine and Aila [33] employs EMA to average predictions instead of model weights, disregarding the fact that the labels obtained in the initial stage of training may still be subpar and the reliability of these labels. In practice, a student may access multiple teachers, and the collective guidance from multiple teacher networks can greatly benefit the training of the student network [48], [49], [50]. Studies in this field can be divided into three categories. The co-training framework [51], [52], [53] incorporates a complementary student model, enabling mutual supervision between both models. However, these methods not only introduce additional computational overhead and lose EMA stability but also fail to generate sufficient model diversity through mere differences in initialization and network structures. The multi-teacher alternating framework [54] introduces two non-trainable teacher networks that are momentum-updated periodically and randomly in an alternate fashion. However, these methods produce model diversity by designing alternating update strategies for teacher networks, while also preventing each update of the teacher network from fully benefiting from all the data. The multi-teacher ensembling framework [29], [55], [56] encourages the student network to iteratively update different teacher networks, each initialized with different parameters or structures. A notable limitation of these methods is that their teacher network update strategies have not been sufficiently designed to generate diverse supervision. The diversity among multiple teacher networks is progressively reduced as they receive continuous updates with identical momentum coefficients. Furthermore, existing ensembling methods typically employ averaging strategies and train the model exclusively from their consistent predictions, while the potential benefits of learning from the confidence remain largely unexplored. In contrast to previous methods that rely on varied initialization or architecture, we propose a multi-teacher framework update strategy that maintains sufficient model diversity by setting conservative-to-radical momentum coefficients while maintaining the stability of EMA. The diversity of our multi-teacher

networks persists across the update iterations. We design an uncertainty-weighted strategy to further improve the quality of pseudo-labels.

D. Boundary in Segmentation

Boundary problem is a fundamental aspect of medical imaging with a well-established historical foundation [57], [58], [59]. Some previous efforts [60], [61], [62] propose various mechanisms to refine the segmentation maps from coarse to fine. For example, Yuan et al. [63] introduced a model-agnostic post-processing scheme that estimates an offset map based on the original prediction to improve boundary quality in segmentation results. Other studies [64], [65], [66] also directly exploit the boundary information to improve the segmentation. For instance, Peng et al. [67] proposed circular convolution for efficient feature learning on boundary regions, though it is prone to mis-segmentation caused by grayscale non-uniformity and noise sensitivity. Wang et al. [68] introduced a boundary-aware context neural network that captures detailed boundary information at each stage but suffers from the added complexity of feature fusion. Furthermore, some boundary-guided methods [69], [70], [71], [72] aim to design specialized boundary-guided modules for effectively extracting boundary information. However, these methods are limited to focusing solely on the information within boundary regions while ignoring the connections between boundary and non-boundary regions. Unlike previous approaches that learn boundary information from the features of the entire image, we propose a boundary-guided contrastive learning approach, where the boundary-guided sampling strategy enables our framework to directly learn the (dis)similarities between the representations of boundary and non-boundary regions, thereby improving the model's segmentation capability in boundary regions.

III. Methodology

Given a limited labeled dataset $D_l = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l}$, and massive unlabeled images $D_u = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$, the goal of semi-supervised medical image segmentation is to mine effective supervision from unlabeled images with the help of limited annotations, ultimately achieving comparable segmentation performance to its fully supervised counterpart.

Observing that current semi-supervised learning methods [9], [22], [23] perform unsatisfactorily in boundary region segmentation, we made efforts to improve segmentation performance from both an overall and a boundary-specific perspective. The overview of the proposed BoCLIS framework is illustrated in Fig. 2. Built upon conservative-to-radical teacher networks, as described in Section III-A, our framework incorporates two main novel components. The first is conservative-to-radical teacher learning, as described in Section III-B. The second is boundary-guided contrastive learning, as detailed in Section III-C. These components work together through a two-step process. First, the conservative-to-radical teacher networks employ gradually increasing momentums for the EMA updates to generate slightly different predictions with corresponding uncertainty maps. Subsequently, the student network can learn from more reliable aggregation labels derived from the predictions and uncertainty maps to improve overall segmentation performance. Second, the student network is equipped with a contrastive module, which applies a boundary-guided patch sampling strategy to extract representative features, particularly from boundary regions. It then drives features of the same class toward the approximated cluster centers while pushing features of different classes farther apart, dynamically shrinking cluster volume and enhancing intra-cluster compactness. This process ultimately results in improved performance in boundary region segmentation.

A. Update Strategy for Conservative-to-Radical Teacher Networks

In the teacher-student network paradigm of semi-supervised learning, the teacher network effectively exploits semantic information from unlabeled data to generate pseudo-labels, guiding the training of the student network. However, most existing methods [16], [34] employ a single teacher network to generate pseudo-labels, which may not yield high-quality pseudo-labels with extremely limited labeled data, leading to the unsatisfactory performance of student network. Considering fully exploiting the massive semantic information within unlabeled data, the conservative-to-radical teacher networks are introduced. The student network f_θ^S, parameterized by $\boldsymbol{\theta}$, is optimized by gradient descent, while the m-th teacher network $f_{\zeta_m}^T$ is updated using the EMA of the student as follows:

$$\boldsymbol{\zeta}_m^t = (1-\alpha_m)\boldsymbol{\zeta}_m^{t-1} + \alpha_m \boldsymbol{\theta}^t, \alpha_m \in [\kappa_1, \kappa_2], \tag{1}$$

where t tracks the step number, and α_m is the momentum coefficient [73] to control the pace of update. Here, as α_m increases progressively with m, the update paces of teacher networks shift from conservative to radical, allowing their diversity preserved across the update iterations. For each unlabeled image, the predictions of the teacher networks are aggregated to generate a confident and robust supervision label. This aggregation label, denoted by $\bar{\mathbf{p}}^u$, guides the student network training.

B. Conservative-to-Radical Teacher Learning

1) Learning with Labeled Data: For labeled data D_l, we use Cross-Entropy (CE) loss to train segmentation network for N_l labeled images $\{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l} \in D_l$, which can be formed as,

$$\mathcal{L}_s = -\frac{1}{N_l}\frac{1}{HW}\sum_{i=1}^{N_l}\sum_{j=1}^{HW}\ell_{CE}(\mathbf{p}_{i,j}^l, \mathbf{y}_{i,j}^l), \tag{2}$$

where $\mathbf{p}_{i,j}^l$ denotes the probability generated by the student network on the j-th pixel in the i-th labeled image and $\mathbf{y}_{i,j}^l$ denotes the piexl-wise corresponding annotations. H and W represent the height and width of the input image, respectively.

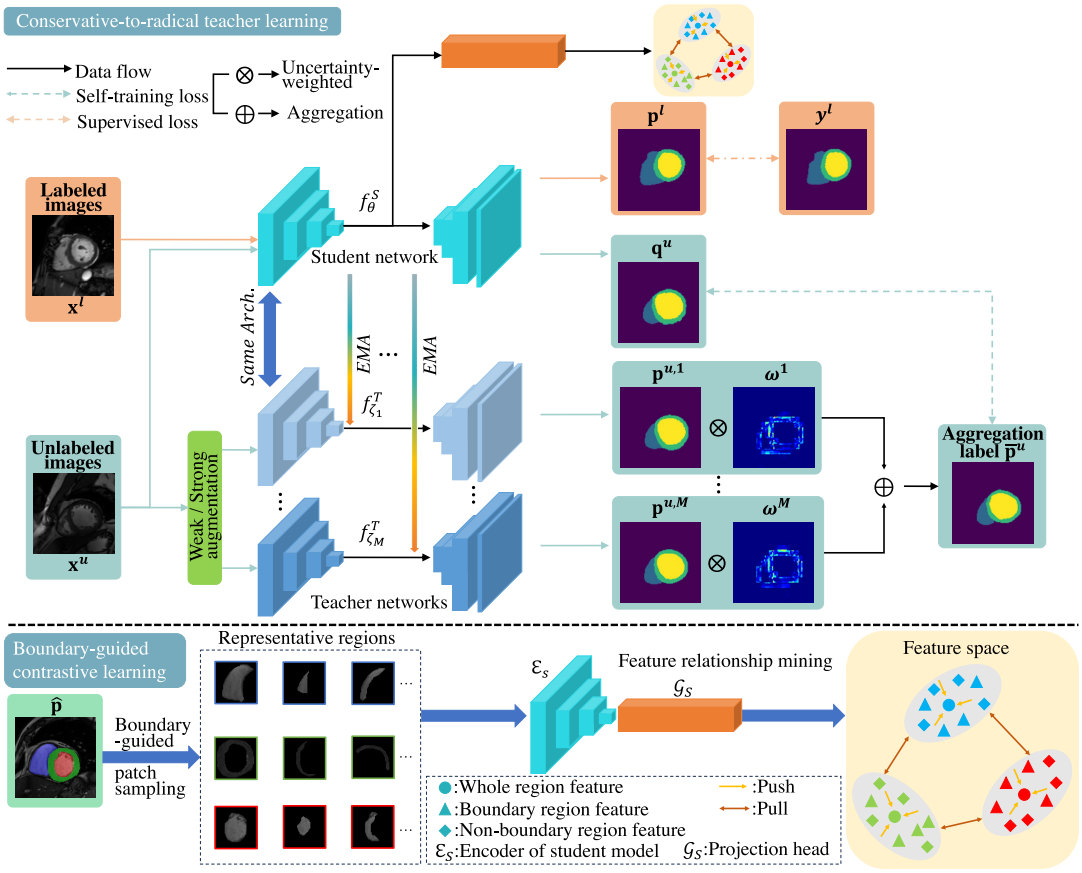

Fig. 2. Overview of the boundary-guided contrastive learning for semi-supervised medical image segmentation framework (BoCLIS). In the conservative-to-radical teacher learning, each labeled image $\mathbf{x}^l$ is exclusively fed into the student network for fully supervised learning. Each unlabeled image $\mathbf{x}^u$ is processed by the conservative-to-radical teacher networks to generate the uncertainty-weighted aggregation label. In the boundary-guided contrastive learning, the aggregation labels are employed to select representative regions.

2) Learning with Unlabeled Data: For each unlabeled image $\mathbf{x}^u$, we first employ the conservative-to-radical teacher networks to obtain M predictions $\{\mathbf{p}^{u,1}, \mathbf{p}^{u,2} \ldots, \mathbf{p}^{u,m}, \ldots, \mathbf{p}^{u,M}\}$, where $\mathbf{p}^{u,m} = f^T_{\zeta_m}(\mathbf{x}^u)$. The dimension of $\mathbf{p}^{u,m}$ is $H \times W \times C$, with C representing the total number of classes for segmentation. Considering that the teacher networks generate pseudo-labels for unlabeled data without the help of corresponding supervisory signals, the results $\mathbf{p}^{u,m}$ obtained by the teacher network may be unreliable. Therefore, we introduce an uncertainty-weighted aggregation strategy to refine the generated pseudo-labels. The uncertainty for the m-th preliminary prediction can be calculated as,

$$\mathcal{H}(\mathbf{p}^{u,m}) = -\sum_{c=1}^{C} p_c^{u,m} \log p_c^{u,m}, \tag{3}$$

where $p_c^{u,m}$ is the pixel-wise probability of class c. The entropy of the prediction for class c only reflects the confidence in belonging to this class. $\mathcal{H}(\mathbf{p}_{i,j}^{u,m})$ captures both the confidence in class c and the uncertainty in the remaining $C-1$ classes. Since entropy reflects the uncertainty degree of information, pixels with low uncertainty, *i.e.*, high confidence, should have a greater impact on the aggregation prediction. This impact can be represented by a weight value $\omega^m = e^{-\mathcal{H}(\mathbf{p}^{u,m})} / \sum_{k=1}^{M} e^{-\mathcal{H}(\mathbf{p}^{u,k})}$. Then, the uncertainty-weighted aggregation prediction $\bar{\mathbf{p}}^u$ can be defined as $\bar{\mathbf{p}}^u = \sum_{m=1}^{M} \omega^m \cdot \mathbf{p}^{u,m}$. For the student network, consistency cost for N_u unlabeled images can be defined as the CE loss:

$$\mathcal{L}_u = -\frac{1}{N_u}\frac{1}{HW}\sum_{i=1}^{N_u}\sum_{j=1}^{HW} \ell_{CE}(\mathbf{q}_{i,j}^u, \bar{\mathbf{p}}_{i,j}^u), \tag{4}$$

where $\mathbf{q}_{i,j}^u$ denotes the probability corresponding to the j-th pixel of the unlabeled image $\mathbf{x}_i^u$ from student network, and $\bar{\mathbf{p}}_{i,j}^u$ is the aggregated prediction of teacher networks. In general, the conservative-to-radical teacher networks generate slightly different predictions and also estimate the corresponding uncertainty maps for any unlabeled input. The uncertainty-weighted aggregation strategy further improves the quality of pseudo-labels by considering the confidence. Then, the student network is supervised by more reliable aggregation labels under the guidance of the conservative-to-radical teacher networks.

C. Boundary-Guided Contrastive Learning

Current semi-supervised segmentation methods [9], [22], [23] have achieved satisfactory accuracy for segmenting main regions of organs or lesions. But they are unable to accurately identify the boundary regions of the foreground under segmentation. Since the boundary regions usually occupy a

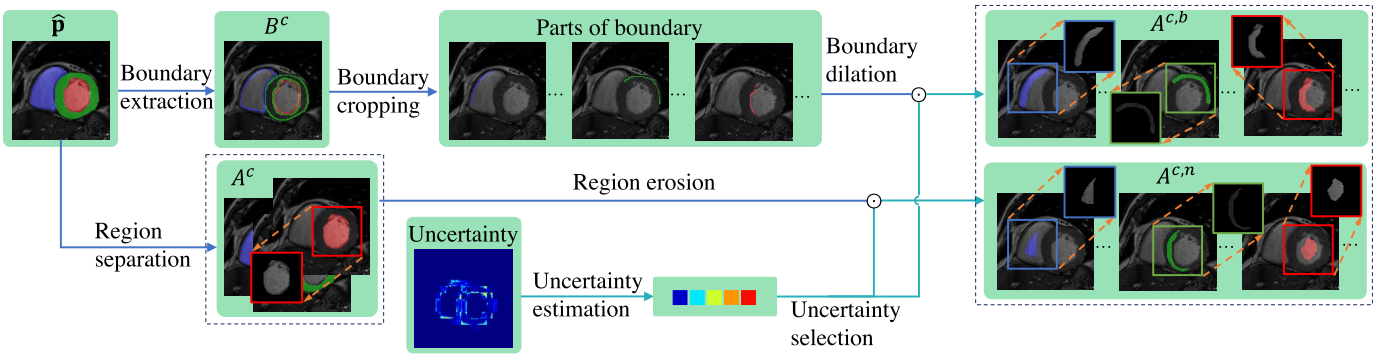

Fig. 3. The detail of boundary-guided patch sampling. The canny operator is utilized for the boundary extraction of different class regions. Boundary patches and non-boundary patches are randomly selected with the guidance of patch uncertainty.

very small percentage of objects and tend to appear blurry, it is difficult for the network to learn massive and meaningful semantic information. To tackle this problem, we design a boundary-guided contrastive learning strategy, which focuses on learning the relationship between boundary and non-boundary regions. The contrastive learning aims to distinguish similar samples (*positive*) from dissimilar ones (*negative*), where an anchor point in a projected embedding space is randomly sampled. To extract embeddings for boundary-guided contrastive learning, a projection head $\mathcal{G}_S$ is introduced after the encoder $\mathcal{E}_S$ of the student network, as illustrated in Fig. 2. In this section, we first introduce the strategy of selecting *positive* and *negative* samples, followed by an explanation of their use in contrastive learning.

1) Boundary-guided Patch Sampling: For a given labeled image $\mathbf{x}^l$, our framework employs the student network to generate the pixel-wise prediction $\mathbf{p}^l$. For an unlabeled image $\mathbf{x}^u$, our framework yields two outputs, with $\mathbf{q}^u$ from the student network and $\bar{\mathbf{p}}^u$ from teacher networks. The reason for using the aggregated prediction $\bar{\mathbf{p}}^u$ rather than $\mathbf{q}^u$ in the subsequent patch selection is that the conservative-to-radical teacher networks generate more robust results. For simplicity, $\mathbf{p}^l$ and $\bar{\mathbf{p}}^u$ are represented by $\mathbf{p}$ thereafter. $\hat{\mathbf{p}}$ represents the binarized segmentation result of $\mathbf{p}$. Let A^c denote the anchor patch that contains the whole region (all pixels) of class c according to $\hat{\mathbf{p}}$. The patch containing an object (or some part of it) of class c can be considered as a *positive* key, while all patches of other $(C-1)$ classes are treated as *negative* keys. Note that all patches are ensured to contain only one class of pixels with the guidance of pixel-wise prediction. However, most of the *positive* and *negative* keys randomly selected in this way may not be representative, which prevents the network from learning discriminative representations, *e.g.*, boundary regions.

Therefore, we design a boundary-guided patch sampling strategy to improve the ability of representation for our semi-supervised framework. Specifically, the boundary regions of the input image are obtained with the Canny operator,

$$\{B^c\}_{c=1}^{C} = Canny(\hat{\mathbf{p}}), \tag{5}$$

where B^c represents the set of coordinates for boundary pixels within the region of class c. With the guidance of B^c, we sample different representative patches, namely the **b**oundary patches $A^{c,b}$ containing a specific proportion of the whole boundary region, and the **n**on-boundary patches $A^{c,n}$ located within the region of class c far away from the boundary. As illustrated in Fig. 3, the patch sampling process can be summarized in three main steps. (1) The boundary B^c is extracted from the binarized prediction $\hat{\mathbf{p}}$ using the Canny detector. A continuous section (one connected component) of B^c is randomly cropped and dilated to generate the boundary patch. This boundary patch is then multiplied by the binary mask for class c to ensure that all pixels retained in the patch belong to class c. (2) The patch A^c, containing all pixels belonging to class c, is obtained by cropping the original image after multiplying it by the binary mask in $\hat{\mathbf{p}}$ that includes class c. (3) The region in A^c is eroded to obtain the non-boundary patches. Note that different kernels and iterations are used for the dilation and erosion operations to ensure diversity in the patches. The final sampled patches are obtained by multiplying the patches from the original images with the binary masks of A^c, $A^{c,b}$, and $A^{c,n}$.

The segmentation results from the conservative-radical teacher networks may not be accurate, potentially causing errors when selecting representative patches based on the boundary of $\hat{\mathbf{p}}$. This can mislead the student network to learn the wrong representation. Effective sampling of numerous patches is of utmost importance. We can sample patches based on their class confidence to reduce the negative impact of inaccurate segmentation results. Therefore, we further design an average patch confidence based on the pixel-wise uncertainty. For the patches $A^{c,b}$ and $A^{c,n}$, they are sent through the student network to obtain the pixel-wise prediction $\mathbf{p}$. The average uncertainty of each patch is defined as the average entropy of all pixels in the prediction. Since patches with high uncertainty are more likely to contain misclassified regions, the patches with low uncertainty are kept as a candidate set for contrastive learning. For each image, the uncertainties for all patches are first estimated for each class. The R patches with the lowest uncertainties are selected to form the candidate set for both boundary and non-boundary regions, denoted by $S^{c,b}$ and $S^{c,n}$ for class c.

2) Patch-based Contrastive Learning: For a given anchor patch A^c, all the patches of class c in $S^{c,b}$ and $S^{c,n}$ are considered as *positive* keys, including $2R$ samples. The

patches of the rest $(C-1)$ classes are considered as *negative* keys, encompassing $2(C-1)R$ samples. These *positive* and *negative* keys are sent to the student encoder and projection head to extract the feature representations. Let A_i^c and $\mathbf{z}_i^c$ denote the anchor patch of class c in image $\mathbf{x}_i$ and its feature representation. $\mathbf{v}_{i,k}^{c+}$ and $\mathbf{v}_{i,r}^{c-}$ represent the feature of the k-th *positive* and the r-th *negative* key for class c, respectively. The boundary-guided contrastive loss is optimized across all classes for N images, defined as,

$$\mathcal{L}_b = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C} \times \log \frac{e^{sim(\mathbf{z}_i^c, \mathbf{v}_{i,k}^{c+})/\tau}}{e^{sim(\mathbf{z}_i^c, \mathbf{v}_{i,k}^{c+})/\tau} + \sum_{r=1}^{2(C-1)R} e^{sim(\mathbf{z}_i^c, \mathbf{v}_{i,r}^{c-})/\tau}}, \quad (6)$$

where $sim(\mathbf{z}, \mathbf{v}) = \mathbf{z}^T\mathbf{v}/\|\mathbf{z}\|\|\mathbf{v}\|$ measures the cosine similarity, and τ denotes the temperature scaling factor used to adjust the scale of the similarity measurement.

The global representation $\mathbf{z}_i^c$ is extracted from an anchor patch including the whole region of class c in image $\mathbf{x}_i$. In comparison, local representations $\mathbf{v}_{i,k}^{c+}$ and $\mathbf{v}_{i,r}^{c-}$ are computed from the patches containing boundary and non-boundary regions of class c and the rest $(C-1)$ classes, respectively. The goal of contrastive learning is to maximize the similarity between global and local representations for patches belonging to the same class, while simultaneously minimizing the similarity between patches in different classes. Eq. 6 only computes the (dis)similarities in the same image, *i.e.*, matching **intra-image** patch representations. To learn more robust representations, we further include **inter-image** representation matching by computing (dis)similarities from different images. Specifically, for an anchor $\mathbf{z}_i^c$, the representation of *positive* keys $\mathbf{v}_{i,k}^{c+}$ and $\mathbf{v}_{i',k}^{c+}(i \neq i')$ are extracted from patches within the same image as well as from different images in the batch. Similarly, the *negative* keys are extracted from patches within the same image and different images in the batch.

D. The Overall Learning Objective

Our framework consists of a student network, conservative-radical-teacher networks for leveraging unlabeled data, and the additional projection head for improving performance in boundary regions. The total objective loss function can be formulated as,

$$\mathcal{L} = \mathcal{L}_s + \lambda_1\mathcal{L}_u + \lambda_2\mathcal{L}_b, \quad (7)$$

where $\mathcal{L}_s$ is the supervised loss only for the labeled data, and $\mathcal{L}_u$ is the pseudo-label training loss for unlabeled data. $\mathcal{L}_b$ is the boundary-guided contrastive loss for the whole dataset. λ_1 and λ_2 are two coefficients.

IV. Experiment

A. Datasets and Evaluation

In this study, we assessed the performance of our method and conducted a comparative analysis with several previous works on three public datasets, including whole brain tumor segmentation dataset (BraTS2020), left atrium segmentation dataset (LA) and cardiac segmentation dataset (ACDC).

1) Two-Class 3D Whole Brain Tumor Segmentation: The BraTS2020 [74] dataset consists of 496 subjects. In this study, we used the FLAIR modality for semi-supervised segmentation of whole tumors. These scans were randomly divided into training (380 scans), validation (26 scans), and testing (90 scans) sets. For pre-processing, each instance was normalized by its channel-wise means and standard deviations. Subsequently, intensity rescaling was performed to ensure values fell within the range of [0, 1].

2) Two-Class 3D Left Atrium Segmentation: The LA dataset [75] contains 100 3D gadolinium-enhanced MR imaging (GE-MRI) scans with corresponding 3D left atrium segmentation masks. Following [16], the dataset was divided into 80 scans for training purposes and the remaining 20 scans for evaluation. Furthermore, to prioritize the heart region, all scans were cropped with the center of attention on this specific area. Additionally, normalization was implemented to ensure that the data maintained a zero mean and unit variance.

3) Three-Class 2D Cardiac Segmentation: The ACDC [76] dataset consists of 100 MR-cine T1 3D volumes depicting cardiac anatomy. Each image within the dataset necessitates the segmentation of three distinct categories: the right ventricle, left ventricle cavities, and the myocardium. Following [34], we randomly selected 75 subjects for training, 5 subjects for validation, and 20 subjects for testing. For pre-processing, we rescaled the intensity of each scan to the range [0, 1].

4) Evaluation Metrics: We utilized four widely recognized evaluation metrics: the Dice Similarity Coefficient (DSC), the Jaccard Index (Jaccard), the 95% Hausdorff Distance (95HD), and the Average Surface Distance (ASD).

B. Implementation Details

In this study, our method was implemented by PyTorch on two NVIDIA GeForce 3090 GPUs. 3D-UNet [77] and UNet [78] were employed as the segmentation backbones for 3D and 2D datasets. The projection head for contrastive learning was basically shallow FC layers [79], consisting of two linear layers with batch normalization and ReLU. The network was converged using the Adam optimizer with a learning rate of 1e-4. Following previous work [22], [35], the student network and contrastive module $\mathcal{G}_S$ are randomly initialized and pre-trained for 100 epochs using labeled data. The teacher network is then initialized with the pre-trained model, and the entire framework undergoes 200 epochs of semi-supervised training. In each semi-supervised iteration, the training batch size was set to 16, with a combination of labeled and unlabeled images in a half-and-half ratio [15]. For 3D volumes, the sub-volumes of size $112 \times 112 \times 112$ were randomly cropped as input to the network. At the inference stage, we introduced the sliding-window strategy to generate the final segmentation results. For 2D images, the inputs were resized to 256×256. To employ different teacher networks from conservative to radical, the value range of momentum coefficient parameter α_m was determined by $\kappa_1 = 0$ and $\kappa_2 = 0.015$. We set the coefficient λ_1 to 0.5 to facilitate the balance between supervised and unsupervised learning, and the

TABLE I

QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART METHODS ON BRATS2020, LA AND 2017 ACDC DATASETS. ↑ INDICATES THAT THE LARGER VALUES ARE BETTER AND ↓ INDICATES THAT SMALLER VALUES ARE BETTER

Method	% scans used		BraTS2020 (3D)				LA (3D)				2017 ACDC (2D)			
	Labeld	Unlabeld	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓
UAMT [16] (MICCAI'19)	5	95	48.92	37.98	20.25	6.68	78.12	65.04	28.86	8.57	48.57	39.62	20.03	7.89
SASSNet [26] (MICCAI'20)			51.46	43.62	23.61	7.56	79.73	67.08	25.36	7.12	57.83	46.51	19.89	7.72
Tri-U-MT [34] (MICCAI'21)			53.64	44.06	19.82	7.39	80.10	66.73	23.41	7.96	58.66	46.92	19.53	7.64
DTC [27] (AAAI'21)			56.44	45.23	17.69	6.45	80.21	67.92	23.88	7.15	56.62	45.17	21.98	7.56
CoraNet [36] (MedIA'22)			57.37	45.86	20.08	6.12	80.71	68.21	19.63	5.14	59.33	47.85	16.69	6.03
MC-Net+ [17] (MedIA'22)			58.45	46.83	21.13	7.33	83.26	71.75	15.08	3.35	62.91	52.62	7.55	2.42
URPC [15] (MedIA'22)			60.03	50.29	18.32	7.17	82.17	70.23	16.42	3.86	62.03	52.16	7.86	2.71
PLCT [9] (MedIA'23)			65.47	55.16	16.67	6.89	87.25	78.34	12.32	2.64	77.96	67.03	6.68	2.57
MCF [23] (CVPR'23)			79.64	68.48	14.86	4.59	83.86	72.04	14.51	3.07	79.74	68.22	6.96	2.43
CAML [30] (MICCAI'23)			77.62	66.21	15.26	5.12	87.07	77.98	12.68	3.49	78.91	68.36	6.33	2.29
DCNet [31] (MICCAI'23)			78.31	67.63	17.43	4.36	86.23	75.26	11.47	2.58	70.32	60.80	8.54	4.16
PatchCL [22] (CVPR'23)			77.95	66.84	15.47	4.87	86.12	78.05	11.97	3.15	79.43	68.25	70.4	2.46
SC-SSL [37] (TMI'23)			77.12	65.89	14.97	4.68	85.92	77.94	12.26	3.07	76.51	66.15	10.18	4.46
BCP [35] (CVPR'23)			79.53	68.42	14.98	4.67	87.39	78.34	11.35	2.52	**87.03**	**76.21**	**4.75**	**1.69**
BoCLIS (Ours)			**82.62**	**71.24**	**12.03**	**3.17**	**88.93**	**79.62**	**10.14**	**2.19**	83.15	72.92	5.08	1.97
UAMT [16] (MICCAI'19)	10	90	80.88	68.74	17.63	6.86	85.62	75.36	16.28	4.46	81.32	70.58	13.17	3.77
SASSNet [26] (MICCAI'20)			82.16	70.85	14.86	4.15	85.51	75.29	18.56	5.13	84.26	74.21	6.08	1.76
Tri-U-MT [34] (MICCAI'21)			82.70	71.41	15.26	3.62	85.52	75.59	14.27	4.51	83.71	73.99	7.54	2.73
DTC [27] (AAAI'21)			81.86	70.31	16.31	3.67	84.68	74.03	13.53	3.44	82.43	71.15	8.82	3.15
CoraNet [36] (MedIA'22)			81.29	69.93	13.97	3.96	83.36	71.92	17.85	4.39	84.17	74.08	6.18	2.41
MC-Net+ [17] (MedIA'22)			83.75	72.18	13.55	3.34	87.41	78.07	10.66	3.06	86.55	77.13	7.01	2.13
URPC [15] (MedIA'22)			84.08	72.24	11.56	3.28	84.08	72.24	11.56	3.28	84.72	74.26	5.12	1.64
PLCT [9] (MedIA'23)			83.51	71.86	13.74	3.62	89.33	80.94	7.42	2.75	86.48	76.73	6.69	2.32
MCF [23] (CVPR'23)			83.47	71.94	13.42	3.14	87.73	78.46	8.62	2.81	86.95	77.43	5.04	1.86
CAML [30] (MICCAI'23)			84.11	73.63	12.15	3.38	89.46	80.98	9.73	2.67	87.46	78.51	5.01	1.39
DCNet [31] (MICCAI'23)			83.21	71.78	11.98	3.54	87.81	78.52	8.92	3.04	87.64	78.83	4.89	1.26
PatchCL [22] (CVPR'23)			83.06	71.52	11.74	3.42	89.05	80.94	7.71	2.95	87.21	78.46	5.15	1.68
SC-SSL [37] (TMI'23)			83.07	71.72	13.16	3.47	86.83	77.91	9.42	2.51	83.14	73.31	9.47	2.63
BCP [35] (CVPR'23)			83.97	72.19	11.47	3.37	89.39	80.97	7.36	7.38	88.02	79.13	4.68	1.43
BoCLIS (Ours)			**86.04**	**75.61**	**9.13**	**2.46**	**90.16**	**82.50**	**6.91**	**1.92**	**88.96**	**79.72**	**4.23**	**1.21**
UAMT [16] (MICCAI'19)	20	80	84.86	74.63	12.21	2.19	88.25	79.14	9.86	3.13	85.62	76.65	9.31	1.53
SASSNet [26] (MICCAI'20)			84.67	73.98	9.41	2.64	88.07	79.01	12.63	3.68	86.98	77.34	5.36	2.52
Tri-U-MT [34] (MICCAI'21)			85.02	74.63	8.83	3.16	88.15	79.08	8.37	3.20	87.04	78.09	5.62	1.63
DTC [27] (AAAI'21)			84.82	74.55	12.69	3.43	87.97	78.63	10.16	2.64	86.13	76.87	6.28	2.31
CoraNet [36] (MedIA'22)			84.37	73.76	9.05	2.62	87.73	78.40	11.29	3.56	86.32	77.03	6.45	2.21
MC-Net+ [17] (MedIA'22)			85.17	74.89	9.72	3.01	90.06	82.45	6.57	2.18	88.35	79.11	5.76	1.73
URPC [15] (MedIA'22)			85.61	75.26	8.91	2.55	89.86	81.19	9.37	3.48	87.18	78.30	5.29	1.64
PLCT [9] (MedIA'23)			85.38	75.16	8.72	2.94	90.07	82.41	7.11	2.45	88.26	79.07	5.84	2.13
MCF [23] (CVPR'23)			85.54	75.22	8.36	2.61	88.75	80.61	6.83	2.31	88.47	79.92	5.61	1.79
CAML [30] (MICCAI'23)			85.95	75.51	9.73	3.64	90.57	83.06	6.35	1.96	89.06	80.49	5.21	1.56
DCNet [31] (MICCAI'23)			86.02	75.64	8.50	2.73	89.92	81.54	8.46	2.89	89.31	80.57	4.86	1.19
PatchCL [22] (CVPR'23)			85.76	75.41	8.33	2.58	89.66	81.32	7.21	2.57	89.06	80.33	5.14	1.61
SC-SSL [37] (TMI'23)			85.47	75.23	8.85	3.11	88.59	80.53	7.01	2.39	86.52	77.81	6.73	1.86
BCP [35] (CVPR'23)			85.86	75.48	8.27	2.51	90.41	82.98	6.52	2.04	89.12	80.42	4.42	1.29
BoCLIS (Ours)			**87.71**	**78.24**	**8.15**	**2.12**	**90.91**	**83.29**	**6.08**	**1.64**	**90.10**	**81.11**	**4.14**	**1.06**
nn-UNet [80](Nat.Methods'21)	100	0	89.29	81.07	7.97	1.83	92.74	84.96	4.12	1.27	92.12	84.83	1.73	0.52

coefficient λ_2 was set to 0.25 to improve the representation ability of the model. Weak augmentation, such as random rotation and crop, was employed, whereas strong augmentation is achieved through changes in brightness. Each unlabeled image is subjected to both weak and strong augmentations. Half of the teacher networks are randomly assigned the weak augmentation image, while the remaining half are assigned the strong augmentation image.

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

In this section, we evaluated several recent semi-supervised approaches for segmentation tasks. The evaluated methods can be categorized as (1) pseudo-label based: UAMT [16], Tri-U-MT [34], CoraNet [36], MCF [23], PatchCL [22], SC-SSL [37] and BCP [35]; (2) consistency regularization based: SASS-Net [26], DTC [27], MC-Net+ [17], URPC [15], PLCT [9], CAML [30], DCNet [31]. Note, PatchCL [22] and PLCT [9] are based on contrastive learning. Furthermore, we introduced nnU-Net [80] as a benchmark for performance comparison, which was trained in a fully supervised manner and serves as an upper bound. We reproduced the baseline results based on the official codes provided by the papers. Some methods, such as UAMT [16], SASSNet [26] and DTC [27], were only implemented for the 3D architecture in their official codes, so we reimplemented their method on 2D architectures to ensure that they can work properly on ACDC datasets. Each method was executed five times and the average outcomes were recorded to ensure the reliability and consistency.

1) Performance on the BraTS2020 Dataset: The segmentation performance for the whole brain tumor is presented in Table I (left half). To ensure a fair comparison, experiments were conducted with three different labeling ratios. It is evident that the result of our method consistently outperforms all baselines across all settings. Specifically, in terms of DSC, the most common metric used to evaluate segmentation performance, our method achieves the best results at 5%, 10%, and 20% labeling ratios. For instance, with 20% labeled data, our method shows a 1.69% improvement (87.71% vs. 86.02%) compared to the second-best method, DCNet. As the labeled data is reduced to 10% and 5%, the performance gains

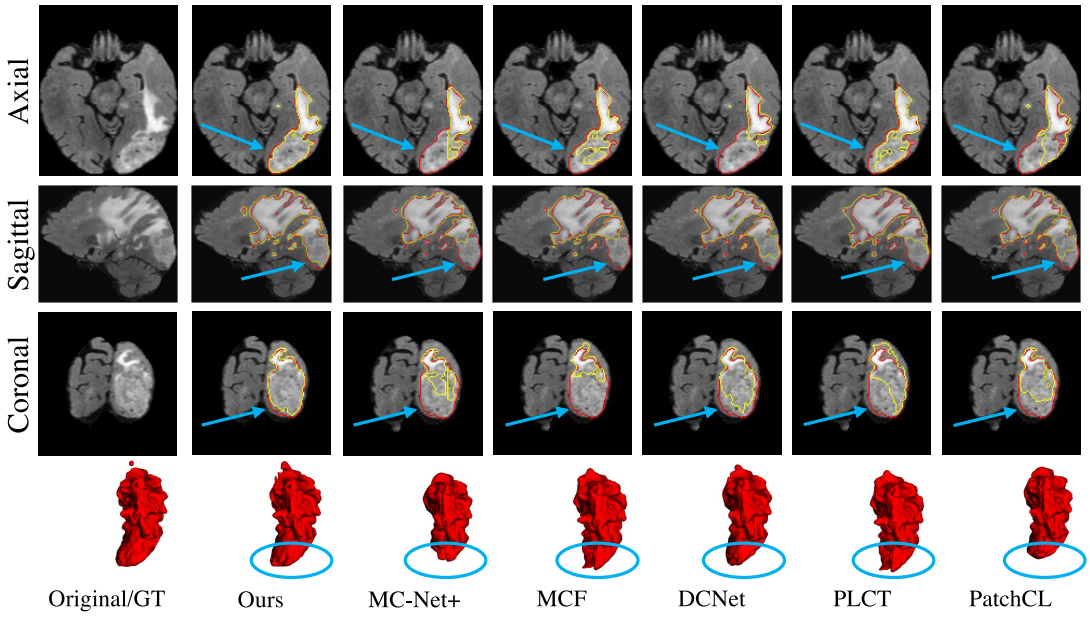

Fig. 4. Visual comparisons between the proposed method and baseline methods (third to seventh column) from one image on BraTS2020. During training, 5% of the training samples were annotated. Red and yellow contours denote ground-truth and prediction boundaries, respectively. Last row (in red): view of the 3D segmentation lesions.

increase from 2.82% to 4.31%, highlighting our method's ability to effectively leverage unlabeled data, especially in low-labeling scenarios. For the other metrics, our method achieves the highest Jaccard values of 71.24%, 75.61%, and 78.24%, while recording the lowest 95HD values of 12.03, 9.13, and 8.15, as well as the lowest ASD values of 3.17, 2.46, and 2.12, respectively. As illustrated in Fig. 4, while other baselines fail to accurately identify ambiguous boundary regions (indicated by blue arrows in the 2D views), our method precisely segments the entire tumor region. The whole segmentation results (indicated by ellipses in 3D views) in the last row further confirm the superior accuracy of our method (second column) in detecting both ambiguous and boundary regions.

TABLE II
QUANTITATIVE COMPARISONS BETWEEN OUR METHOD (WITH AND WITHOUT CP PREPROCESSING) AND THE BCP METHOD ON THE 2017 ACDC DATASET. ↑ INDICATES THAT THE LARGER VALUES ARE BETTER AND ↓ INDICATES THAT SMALLER VALUES ARE BETTER

Method	% scans used		2017 ACDC (2D)			
	Labeld	Unlabeld	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓
BCP [30] (CVPR'23)	5	95	87.03	76.21	4.75	1.69
BoCLIS w/o CP			83.15	72.92	5.08	1.97
BoCLIS w/ CP			**87.63**	**76.91**	**4.49**	**1.23**
BCP [30] (CVPR'23)	10	90	88.02	79.13	4.68	2.63
BoCLIS w/o			88.96	79.72	4.23	1.21
BoCLIS w/ CP			**89.10**	**79.83**	**4.18**	**1.19**

2) Performance on the LA Dataset: As shown in Table I (middle half), our method demonstrates clear superiority by achieving significantly better performance across all settings compared to other methods. With only 5% labeled data, our method achieves an impressive DSC of 88.93%, marking a substantial improvement of 1.54% over the BCP. Interestingly, pseudo-label based methods, such as UAMT and Tri-U-MT, show lower performance than consistency-based methods when only 5% of the training data is labeled. This reduced performance may be due to the inability of a single teacher network to provide reliable and accurate pseudo-labels to the student network for effective supervision, especially when there is an extremely limited amount of labeled data for training. To address this, our method employs conservative-to-radical teacher networks, which better utilize the additional information from the unlabeled data. As a result, it achieves the highest DSC of 90.16% and Jaccard index of 82.50% in the 10% labeled data setting. Furthermore, when the labeled data ratio increases to 20%, our model achieves results comparable to nn-UNet (which is trained with 100% labeled data), with a DSC of 90.91%, compared to the upper bound model's DSC of 92.74%.

3) Performance on the ACDC Dataset: Our method was further evaluated and compared to other approaches on a 3-class 2D segmentation task. The results of all methods are presented in Table I (right half). Similar outcomes were observed on the ACDC dataset, where our method demonstrated the best performance compared to other approaches in the setting of 10% and 20% labeled data. Although the performance of our method in the 5% labeled setting is lower than that of BCP, with a DSC of 87.03%, Jaccard of 76.21%, 95HD of 4.75, and ASD of 1.69, our method outperforms BCP in the 10% and 20% labeled settings. One possible reason BCP performs better than our method on the ACDC dataset under the 5% annotation setting is that BCP leverages the Copy-Paste [81] data preprocessing strategy, which can construct the precise distribution of the entire dataset with as few as three annotated samples (5% labeled) in ACDC dataset and align the empirical

distributions of labeled and unlabeled features [35]. However, as the annotated data increases to 10% or 20%, the advantage of the Copy-Paste strategy diminishes. To validate this hypothesis, we integrated the Copy-Paste (CP) preprocessing strategy into our approach, as shown in Table II. The results indicate that under the 5% annotation setting on the ACDC dataset, our method achieved a significant improvement, with a DSC of 87.63%, a Jaccard of 76.91%, a 95HD of 4.49, and an ASD of 1.23, all surpassing those of BCP. As the annotated data increases to 10%, the performance improvement from CP preprocessing in our method becomes less significant.

It is noteworthy that our method still maintains obviously superiority compared to the two contrastive learning based methods, *i.e.*, PLCT [9] and PatchCL [22]. The reason is that PLCT [9] only considers feature contrast at the pixel level, while PatchCL [22] extends the comparison to the patch level, but still ignores the relationship between patches in boundary regions and patches in non-boundary regions. As a result, previous semi-supervised methods fail to accurately delineate the boundary regions, while in our method, boundary regions are effectively captured (in Fig. 4) with the help of boundary-guided patch sampling.

Another surprising observation is that as the labeled data decreases to 10% and 5%, the performance gains increase, highlighting that our method allows to effectively exploit unlabeled data for performance improvement, especially in smaller labeled scenario.

4) Performance on the Boundary Region: DSC, the most common segmentation evaluation metric, considers the entire overlap between the segmentation result and ground-truth, which is insensitive to boundary regions if the overlapping area is large. To more comprehensively evaluate the segmentation performance of different methods in boundary regions, we defined the pixel band obtained from ground-truth after extending the edge by 10 pixels for DSC evaluation. Across the BraTS2020, LA, and 2017 ACDC datasets, our method consistently outperformed the strongest baseline PatchCL by 20.47%, 16.75%, and 17.18% in terms of DSC, respectively, when trained with only 5% labeled images (Fig. 5). The impact of boundary-guided patch sampling strategy on boundary region segmentation performance has also been further explored. This significant performance gain further demonstrates that our proposed boundary-guided contrastive learning enables the network to learn representative features in boundary regions.

D. Sensitivity Study

In this section, we performed comprehensive experiments to explore the sensitivity of the components in our framework.

1) Sensitivity of Teacher Network Numbers M: The number of pseudo-labels for each input is decided by the hyper-parameter M, which determines the quality of aggregation labels and plays a vital role in stable training. As shown in Fig. 6, we set M from 2 to 8 to investigate its effects. It can be observed that appropriately increasing the number of teacher networks can stably improve segmentation performance. Note that if the M is greater than 4, our method is insensitive to it.

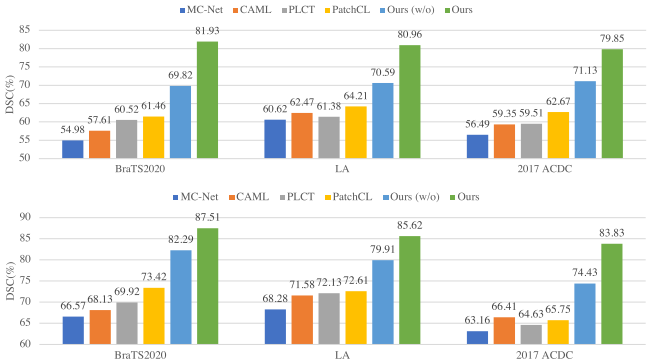

Fig. 5. Performance on boundary regions from our method with or without boundary-guided patch sampling strategy and other baselines on BraTS2020, LA 2017 ACDC datasets, with 5% (upper) and 10% (lower) labeled images used for training.

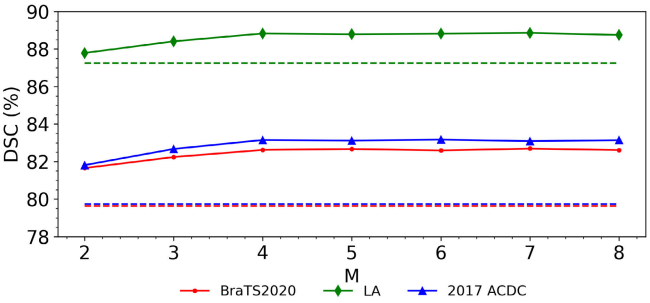

Fig. 6. Performance of our method on BraTS2020, LA and 2017 ACDC datasets with different teacher network numbers of M, where 5% labeled images were used for model training. Dashed lines represent the performance of the strongest baselines.

2) Sensitivity of Weight Balance λ_1 and λ_2: We validate the sensitivity of the weights λ_1 and λ_2 in Eq. 7, which are used to control the trade-off between the semi-supervised learning and the contrastive learning in the total loss. To evaluate their impact, we uniformly vary the values of λ_1 and λ_2 within the range of [0, 1] with an increment of 0.25. We conduct the experiments on three datasets with 5% labeled scans for training. From Table III, it is evident that the performance is relatively stable across different weight values of λ_1 and λ_2.

3) Sensitivity of Momentum Coefficient Range $[\kappa_1, \kappa_2]$: The update space of each teacher network is adjusted by the hyper-parameter α_m. As the parameters α_m continue to increase within the range of $[\kappa_1, \kappa_2]$, the update paces of the teacher model change from conservative to radical. We first discuss the setting of gradually increasing α_m. Specifically, we set each α_m in Eq. 1 by equally dividing the value range of the $[\kappa_1, \kappa_2]$ according to the number of teacher models, with $M = 4$ and $\kappa_1 = 0$ by default. Each α_m is individually to applied the corresponding teacher network $f_{\zeta_i}^T$, resulting in a robust and accurate aggregation label. In Table IV (upper half), the DSC performance of our method is presented with different value range of α_m. The results demonstrate that in different settings of the labeled data ratio, the DSC performance remains comparable across varing κ_2 values. This suggests that our method is relatively insensitive to changes in the momentum coefficient within a reasonable range. To further investigate

TABLE III
SENSITIVITY OF HYPER-PARAMETERS λ_1 AND λ_2 IN DSC ON BRATS2020, LA AND ACDC DATASET, RESPECTIVELY

λ_1	0.25				0.5				0.75				1			
λ_2	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
BraTS2020	81.22	81.91	81.97	81.74	**82.62**	82.36	81.96	81.87	81.67	81.73	82.11	82.06	81.52	81.48	81.68	81.37
LA	87.45	87.86	87.92	88.04	**88.93**	88.76	88.21	88.34	87.63	87.95	88.26	88.04	87.66	87.54	87.71	87.82
ACDC	82.01	82.24	82.35	82.42	83.15	**83.21**	82.86	82.43	82.45	82.64	82.76	82.51	82.15	82.37	82.46	82.24

TABLE IV
SENSITIVITY ANALYSIS OF THE MOMENTUM COEFFICIENT SETTING STRATEGY AND VALUE RANGE ON THE BRATS2020 DATASET. DSC IS USED AS THE METRIC TO EVALUATE SEGMENTATION PERFORMANCE

α_m setting strategy	κ_2	5% Labeled	10% Labeled	20% Labeled
Increasing	0.005	82.18	85.71	87.52
Increasing	0.01	82.43	85.83	**87.77**
Increasing	0.015	**82.62**	**86.04**	87.71
Increasing	0.02	82.37	85.96	87.63
Random	0.005	80.77	84.73	86.68
Random	0.01	**80.82**	**85.04**	86.84
Random	0.015	80.55	84.96	**87.03**
Random	0.02	80.49	84.81	86.85

TABLE V
SENSITIVITY ANALYSIS OF THE PATCH SELECTION STRATEGY AND REPRESENTATION MATCHING SCHEME ON THE ACDC DATASET. DSC IS USED AS THE METRIC TO EVALUATE SEGMENTATION PERFORMANCE

Selection strategy	Scheme	Memory	5% Labeled	10% Labeled	20% Labeled
Random	Inter	10	81.91	87.92	89.12
Random	Inter	15	**82.06**	**88.07**	89.24
Random	Inter	20	82.05	87.98	**89.31**
Uncertainty-based	Inter	10	82.97	88.75	89.92
Uncertainty-based	Inter	15	**83.15**	**88.96**	90.10
Uncertainty-based	Inter	20	83.12	88.84	**90.13**
Uncertainty-based	Intra	10	82.25	88.12	89.47
Uncertainty-based	Intra	15	**82.47**	88.31	**89.62**
Uncertainty-based	Intra	20	82.41	**88.35**	89.54

the effectiveness of the strategy with gradually increasing α_m, we perform a comparison by setting the α_m randomly within the value range of $[\kappa_1, \kappa_2]$. As illustrated in Table IV (lower half), the results demonstrate our designed gradually increasing strategy obtains gains on performance.

4) Sensitivity of Patch Representation: Here, we investigate our patch sampling method with three noteworthy aspects: **(1) Patch selection strategy:** For the patches $A^{c,b}$ and $A^{c,n}$, their average patch uncertainties are used to select representative patches. As a baseline, random patch sampling selects patches without considering their uncertainty. In Table V, it is obvious that the average patch uncertainty strategy is consistently better than the random patch strategy. **(2) Number of selected patches:** We set the memory space of the *positive* and *negative* key sets to 10, 15, and 20, respectively. It is noticed that increasing the number of selected patches from 10 to 15 results in a marginal improvement. Furthermore, the performance is relatively stable when the number is increased further from 15 to 20. This suggests that our method is insensitive to the number of patches selected within a reasonable range. **(3) Representation matching scheme:** For an anchor representation $\mathbf{z}_i^c$, the *positive* keys $\mathbf{v}_{i,k}^{c+}$ and the *negative* keys $\mathbf{v}_{i,r}^{c-}$ are computed from within the same image for matching intra-image patch representations, while the *positive* keys $\mathbf{v}_{i',k}^{c+}(i \neq i')$ and the *negative* keys $\mathbf{v}_{i',r}^{c-}$ as the supplements for matching inter-image patch representations. To investigate the impact of the matching scheme on segmentation performance, we employed two different approaches in our experiments: one limiting the anchor representation, *positive* keys, and *negative* keys to originate from the same image (intra-matching), and the other allowing these representations and keys to come from different images (inter-matching). It is evident that inter-level matching can consistently improve the segmentation performance.

TABLE VI
SENSITIVITY ANALYSIS OF THE METRIC LEARNING ON THE BRATS2020 AND 2017 ACDC DATASETS. DSC IS USED AS THE METRIC TO EVALUATE SEGMENTATION PERFORMANCE

Metric loss	BraTS2020 (3D)			2017 ACDC (2D)		
	5% labeled	10% labeled	20% labeled	5% labeled	10% labeled	20% labeled
Center [83]	75.42	79.68	80.04	74.58	80.95	82.33
Triplet [82]	76.81	80.23	81.37	75.17	81.72	83.46
Contrastive (ours)	**82.62**	**86.04**	**87.71**	**83.15**	**88.96**	**90.10**

5) Sensitivity of Different Metric Learning: We validate the effectiveness of the boundary-guided contrastive loss by comparing it with other metric losses, specifically triplet loss [82] and center loss [83]. **Triplet loss:** For an anchor representation $\mathbf{z}_i^c$, we calculate its Euclidean distance to both the *positive* keys $\mathbf{v}_{i,k}^{c+}$, $\mathbf{v}_{i',k}^{c+}$ and the *negative* keys $\mathbf{v}_{i,r}^{c-}$, $\mathbf{v}_{i',r}^{c-}$. **Center loss:** The cluster center is estimated as the average of all in-class representations, and then the distance between each sample and its corresponding class center is calculated. As shown in Table VI, the boundary-guided contrastive loss performs better than triplet loss and center loss on the BraTS2020 and 2017 ACDC datasets. Compared to triplet loss and center loss, our method directly optimizes dis(similarity) relationships between samples and learns more representative feature representations, achieving a better balance between inter-class separation and intra-class compactness.

E. Analysis of the Boundary Refinement Method

To further evaluate the effectiveness of our framework in boundary regions, we quantitatively and qualitatively compare our method with other baselines using a model-agnostic post-processing method, SegFix [63], on the BraTS2020 and 2017 ACDC datasets. Table VII presents the performance of our method and other semi-supervised baselines using SegFix. While SegFix consistently improves the DSC of the baselines, our method outperforms the baselines with SegFix,

TABLE VII
QUANTITATIVE RESULTS OF THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS WITH BOUNDARY REFINEMENT ON THE BRATS2020 AND 2017 ACDC DATASETS. DSC IS USED AS THE EVALUATION METRIC TO ASSESS PERFORMANCE

Method	BraTS2020 (3D)			2017 ACDC (2D)		
	5% labeled	10% labeled	20% labeled	5% labeled	10% labeled	20% labeled
MCF + SegFix	80.01	84.37	86.04	80.26	87.41	88.97
CAML + SegFix	78.65	84.86	86.29	79.54	87.82	89.29
PatchCL + SegFix	78.92	83.75	86.36	80.52	87.74	89.36
Ours	82.62	86.04	**87.71**	83.15	88.96	90.10
Ours + SegFix	**82.71**	**86.16**	87.64	**83.27**	**89.04**	**90.23**

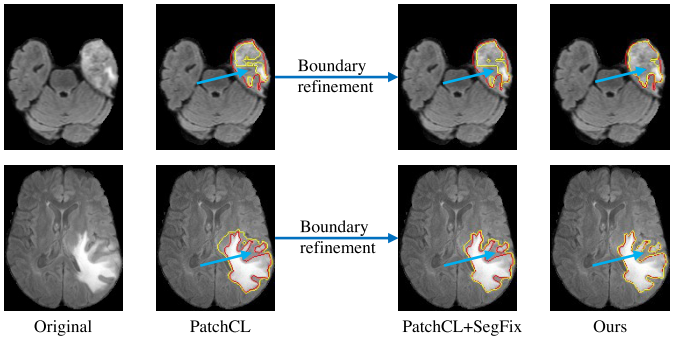

Fig. 7. Visual comparisons between the proposed method and PatchCL with boundary refinement from different subjects on the BraTS2020 dataset. During training, 5% of the training samples were annotated. Red and yellow contours denote ground-truth and prediction boundaries, respectively.

achieving DSC increases of 3.7% and 2.63% on the two datasets with 5% labeled data. One reason is that SegFix primarily focuses on refining pixels near the boundary of predictions, which does not lead to substantial improvements in overall segmentation performance. As illustrated in Fig. 7, misclassified regions (indicated by blue arrows) far from the boundary cannot be effectively refined by SegFix.

F. Impact of the Boundary-Guided Contrastive Learning in Pseudo-Labels

One important claim is that our boundary-guided contrastive learning significantly improves the quality of the pseudo-labels. We present the quantitative and visualization results comparing the pseudo-labels generated by our method with those from the ablated version of our method without boundary-guided contrastive learning, using different subjects from the BraTS2020 dataset. Under the 5% annotation setting, the DSC between the pseudo-labels generated by our method and the ground truth is 86.31%, the Jaccard is 74.42%, the 95HD is 8.96, and the ASD is 2.29. In comparison, the ablated version without boundary-guided contrastive learning achieved a DSC of 82.14%, a Jaccard of 70.79%, a 95HD of 12.41, and an ASD of 3.46. As shown in Fig. 8, the pseudo-labels produced by our approach (second row) demonstrate superior accuracy in identifying boundary regions (indicated by blue arrows), outperforming the ablated version without boundary-guided contrastive learning (third row). This may explain why our method achieves better performance than other baselines, particularly in segmenting ambiguous and boundary regions. It is important to note that the ground-truth was not accessible during model training and is only provided for boundary comparison.

TABLE VIII
QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART METHODS ON PANCREAS-NIH DATASET. ↑ INDICATES THAT THE LARGER VALUES ARE BETTER AND ↓ INDICATES THAT SMALLER VALUES ARE BETTER

Method	% scans used		Pancreas-NIH (3D)			
	Labeld	Unlabeld	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓
MC-Net+ [17] (MedIA'22)	5	95	68.85	54.26	15.27	4.28
PLCT [9] (MedIA'23)			76.72	65.59	12.67	3.64
CAML [30] (MICCAI'23)			74.95	64.25	11.25	3.79
PatchCL [22] (CVPR'23)			77.26	66.07	9.41	3.52
BoCLIS (Ours)			**80.31**	**68.87**	**7.43**	**2.87**
MC-Net+ [17] (MedIA'22)	10	90	74.12	60.09	12.24	3.26
PLCT [9] (MedIA'23)			79.53	67.02	7.64	2.72
CAML [30] (MICCAI'23)			78.91	66.72	8.16	3.68
PatchCL [22] (CVPR'23)			80.37	68.69	8.26	3.07
BoCLIS (Ours)			**82.24**	**70.39**	**6.97**	**2.52**
nn-UNet [80](Nat.Methods'21)	100	0	83.46	71.29	5.24	1.26

TABLE IX
QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART METHODS ON DRIVE DATASET. ↑ INDICATES THAT THE LARGER VALUES ARE BETTER AND ↓ INDICATES THAT SMALLER VALUES ARE BETTER

Method	% scans used		DRIVE (2D)			
	Labeld	Unlabeld	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓
MC-Net+ [17] (MedIA'22)	10	90	69.56	55.98	10.27	0.97
PLCT [9] (MedIA'23)			71.28	57.25	9.14	0.89
CAML [30] (MICCAI'23)			72.39	58.20	8.79	0.72
PatchCL [22] (CVPR'23)			74.62	63.68	6.89	0.57
BoCLIS (Ours)			**78.42**	**67.03**	**2.31**	**0.12**
MC-Net+ [17] (MedIA'22)	20	80	75.30	60.54	8.23	0.30
PLCT [9] (MedIA'23)			77.73	63.73	6.98	0.51
CAML [30] (MICCAI'23)			79.08	65.54	4.98	0.18
PatchCL [22] (CVPR'23)			79.96	66.71	5.09	0.48
BoCLIS (Ours)			**82.96**	**70.89**	**1.10**	**0.03**
nn-UNet [80](Nat.Methods'21)	100	0	84.02	71.81	0.73	0.01

G. Analysis of Challenging Boundary Regions

We further verify the effectiveness of the proposed method on data with very thin structures and more challenging boundary regions. To evaluate segmentation performance on such data, we conduct experiments on the Pancreas-NIH [84] and DRIVE [85] datasets. The Pancreas-NIH dataset consists of 82 cases, with 58 used as the training set, 4 as the validation set, and 20 as the test set. Table VIII presents the segmentation results, demonstrating that our method consistently surpasses all baselines across various settings. For example, under the 5% setting, our method achieves a DSC of 80.31%, a Jaccard of 68.87%, a 95HD of 7.43 and an ASD of 2.87, marking a significant DSC improvement of 3.05% over the second-best method, PatchCL. To more comprehensively evaluate the segmentation performance of different methods in boundary regions, we defined the pixel band obtained from ground-truth after extending the edge by 10 pixels for DSC evaluation. Our method consistently outperformed the second-best baseline PatchCL by 17.98% and 12.13% in terms of DSC on boundary regions, respectively, when trained with only 5% and 10% labeled images (Fig. 9). The visualization results shown in

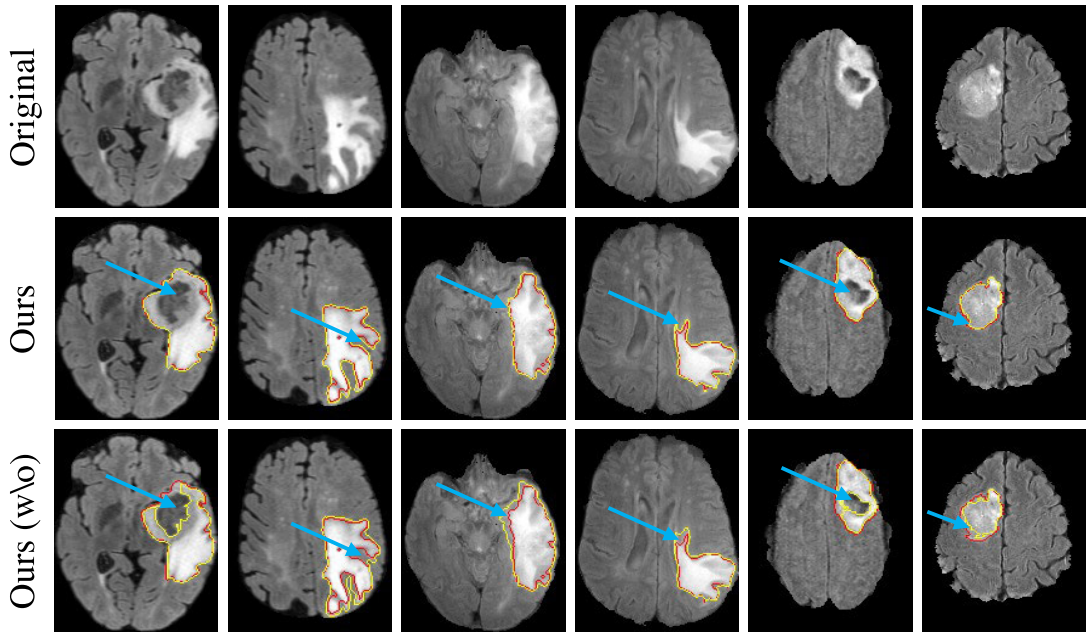

Fig. 8. Visual comparisons of pseudo-labels between the proposed method (second row) and the ablated version of the proposed method without boundary-guided contrastive learning (third row) from different subjects on the BraTS2020 dataset. During training, 5% of the training samples were annotated. Red and yellow contours denote ground-truth and pseudo-label boundaries, respectively.

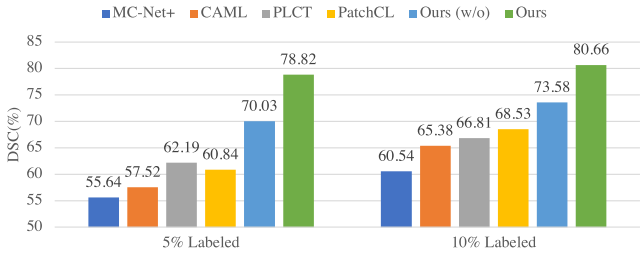

Fig. 9. Performance on boundary regions from our method with or without boundary-guided patch sampling strategy and other baselines on Pancreas-NIH dataset, with 5% and 10% labeled images used for training.

Fig. 10 indicate that our method is able to accurately segment the boundary regions, even when the foreground regions are more imbalanced. We further investigate the performance of the proposed method on thinner structures. The DRIVE dataset consists of 40 retinal images, with 18 used as the training set, 2 as the validation set, and 20 as the test set. The retinal vessel image contains many thin vessels and the segmentation result is illustrated in Table IX. It is obvious that our method is competitive with other methods by achieving the best results with the highest DSC of 78.42%, Jaccard of 67.03%, 95HD of 2.31 and ASD of 0.12, in the 10% labeled setting. We further visualize the vessel segmentation results, including those of PatchCL, PLCT, CAML, MC-Net+ and our method, as shown in Fig. 11. We can observe that our method detects more thin vessel pixels with low contrast than other baselines.

H. Ablation Study

The further detailed ablation studies are performed on the BraTS2020 dataset to show the effectiveness of each component we designed. Note that the average patch uncertainty selection strategy and $M = 4$ are set by default in the following experiments. Fig. X reveals that, (1) the most significant performance improvements (the DSC gains are 24.96% and 23.13%, respectively) are achieved by introducing the conservative-to-radical teacher networks, denoted as CRT, (*i.e.*, improving the quality of aggregation labels) compared to the single-teacher network; (2) the proposed uncertainty-weighted aggregation strategy, referred UWA, can consistently improve the segmentation performance, achieving DSC gains of 0.73% and 0.65%; (3) introducing patch-level contrastive learning module, denoted as PLC, results in DSC improvements of 7.54% and 6.06%. Note that our proposed method considers the (dis)similarities between local and global representations as well as intra- and inter-images, which greatly promotes the

TABLE X
ABLATION STUDY OF OUR DESIGNED MODULES ON THE BRATS2020 DATASET. CRT, UWA, PLC AND BGS DENOTE THE CONSERVATIVE-TO-RADICAL TEACHER NETWORKS, UNCERTAINTY-WEIGHTED AGGREGATION, PATCH-LEVEL CONTRASTIVE LEARNING AND BOUNDARY-GUIDED PATCH SAMPLING, RESPECTIVELY

% scans used		Designs				Metrics			
Labeled	Unlabeled	CRT	UWA	PLC	BGS	DSC(%) ↑	Jaccard(%) ↑	95HD(mm) ↓	ASD(mm) ↓
5	95					45.61	35.72	24.86	8.62
		✓				68.39	59.03	16.83	6.97
		✓	✓			70.57	60.58	15.82	5.69
				✓		58.63	47.28	17.64	6.58
				✓	✓	64.46	53.41	16.72	6.31
		✓		✓		76.84	66.41	15.41	5.59
		✓	✓	✓		78.11	67.52	14.96	4.63
		✓		✓	✓	81.89	70.70	12.89	3.75
		✓	✓	✓	✓	**82.62**	**71.24**	**12.03**	**3.17**
10	90					53.82	44.16	19.23	7.61
		✓				75.21	64.98	16.95	6.72
		✓	✓			76.96	66.52	16.72	5.64
				✓		61.43	51.20	17.27	6.87
				✓	✓	66.12	55.79	16.04	6.53
		✓		✓		81.97	70.79	13.37	4.75
		✓	✓	✓		83.02	72.51	12.05	3.62
		✓		✓	✓	85.39	73.68	10.29	3.09
		✓	✓	✓	✓	**86.04**	**75.61**	**9.31**	**2.46**

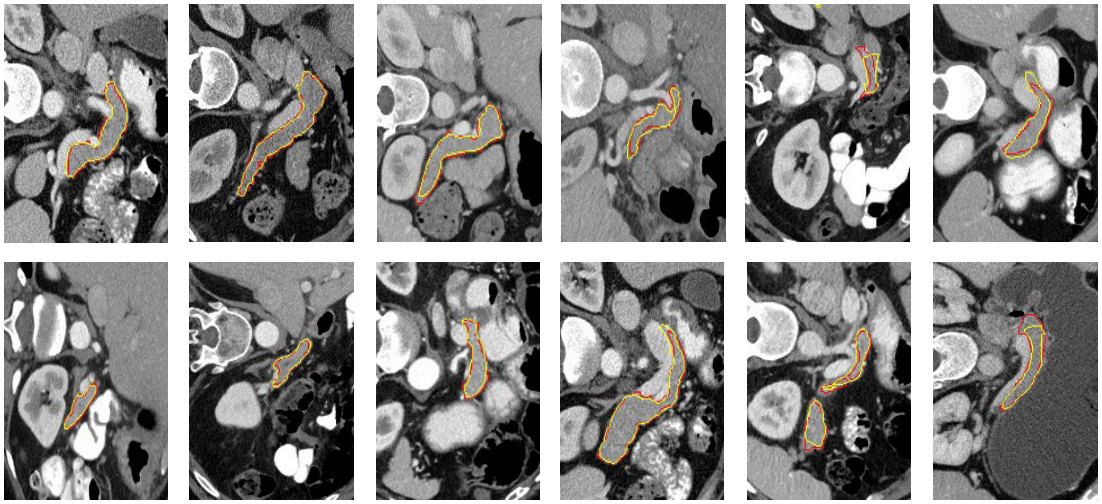

Fig. 10. Visualization results of the proposed method from different subjects on the Pancreas-NIH dataset. During training, 5% of the training samples were annotated. Red and yellow contours denote ground-truth and prediction boundaries, respectively.

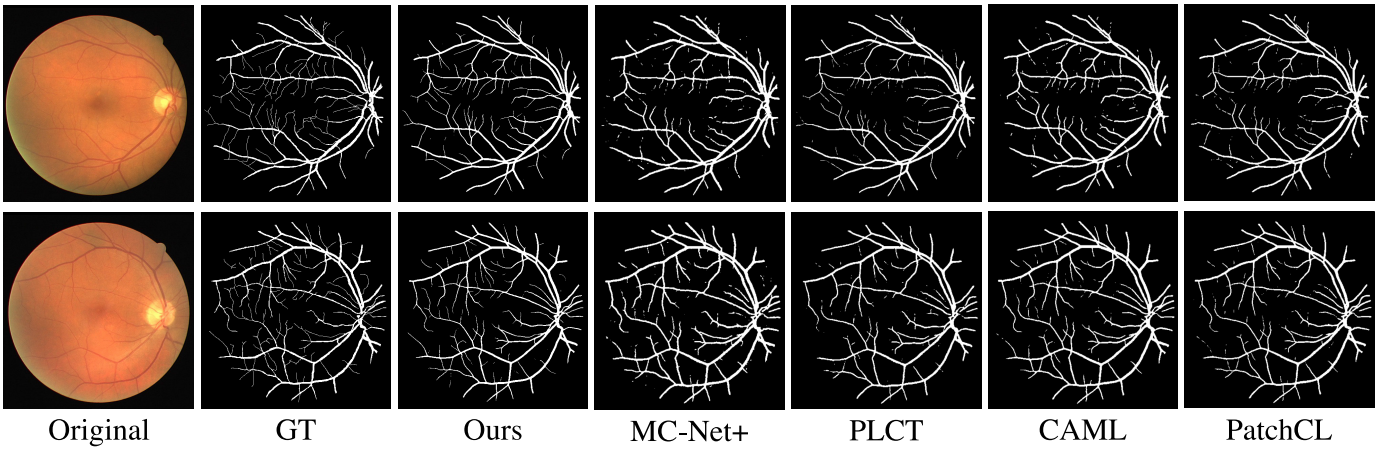

Fig. 11. Visual comparisons between the proposed method and baseline methods (fourth to seventh column) from different subjects on the DRIVE dataset. During training, 10% of the training samples were annotated.

utilization of unlabeled data to achieve better performance; (4) further designing the boundary-guided patch sampling, denoted as BGS, obtains the gains of 4.51% and 3.02% in DSC. The results suggest that strengthening the comparison between boundary region and whole region representations is beneficial to the segmentation performance of the model.

I. Limitations and Future Work

However, our method has some limitations. (1) Model design still requires multiple teacher networks, introducing extra computation. (2) The range of momentum is manually determined based on experimental results. (3) For the patch selection strategy, current uncertainty estimates are based on previous work [22] and are limited, which will affect the efficiency of selecting representative patches. (4) More model architectures, such as Transformer, are deserved to adapt to our framework. However, the limited amount of available data in medical imaging may be the reason why the Transformer architecture has not been widely adopted in current semi-supervised learning methods [86]. In future work, we will be dedicated to addressing the above limitations. For instance, the development of an automated strategy for setting the momentum range can effectively optimize the training process. Additionally, we notice that Zhang et al. [21] proposes the Best-model Moving Average (BMA) strategy to update the teacher model, which could be further introduced into our framework.

V. Conclusion

In this paper, we have proposed a novel semi-supervised method based on conservative-to-radical teacher networks with patch-level boundary-guided contrastive learning. The conservative-to-radical update strategy is designed to maintain teacher model diversity, while the integration of uncertainty-weighted aggregation helps the teacher networks effectively utilize the extra semantic information inherent in unlabeled data. These networks can significantly improve the quality of the pseudo-labels, as the reliable supervision for the student network. To overcome the formidable challenge of boundary segmentation, we design a boundary-guided patch sampling strategy and patch-level boundary-guided contrastive learning to guide our framework to learn more

discriminative representations in boundary regions. The proposed conservative-to-radical teacher networks and the boundary-guided patch sampling strategy are easily applicable to other segmentation networks, which indicates the usability and scalability of our method. A comprehensive evaluation on three public datasets has shown our method achieves the best performance under different limited annotation settings.

REFERENCES

[1] W. Zhou et al., "Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images," *Nature Commun.*, vol. 12, no. 1, p. 1259, Feb. 2021.

[2] J.-X. Zhuang, J. Cai, J. Zhang, W.-S. Zheng, and R. Wang, "Class attention to regions of lesion for imbalanced medical image recognition," *Neurocomputing*, vol. 555, Oct. 2023, Art. no. 126577.

[3] Y. Yang, R. Wang, T. Zhang, and J. Su, "Semi-supervised medical image segmentation via feature-perturbed consistency," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2023, pp. 1635–1642.

[4] X. Wang, B. Zhang, L. Yu, and J. Xiao, "Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3114–3123.

[5] 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.

[6] C. Chen, K. Zhou, Z. Wang, and R. Xiao, "Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 346–353, Feb. 2023.

[7] Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, "Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11350–11359.

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

[9] 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.

[10] H. Ye et al., "CAR: Class-aware regularizations for semantic segmentation," in *Proc. ECCV*, Jan. 2022, pp. 518–534.

[11] T. Nguyen-Duc, T. Le, R. Bammer, H. Zhao, J. Cai, and D. Phung, "Cross-adversarial local distribution regularization for semi-supervised medical image segmentation," in *Proc. MICCAI*, Jan. 2023, pp. 183–194.

[12] F. Lyu, M. Ye, J. F. Carlsen, K. Erleben, S. Darkner, and P. C. Yuen, "Pseudo-label guided image synthesis for semi-supervised COVID-19 pneumonia infection segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 797–809, Mar. 2023.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1195–1204.

[14] S. Adiga V., J. Dolz, and H. Lombaert, "Anatomically-aware uncertainty for semi-supervised image segmentation," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 103011.

[15] X. Luo et al., "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102517.

[16] 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.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 605–613.

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

[18] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2019.

[19] J. Chen, J. Zhang, K. Debattista, and J. Han, "Semi-supervised unpaired medical image segmentation through task-affinity consistency," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 594–605, Mar. 2023.

[20] P. Qiao et al., "Semi-supervised CT lesion segmentation using uncertainty-based data pairing and SwapMix," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1546–1562, May 2023.

[21] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102656.

[22] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19786–19797.

[23] Y. Wang, B. Xiao, X. Bi, W. Li, and X. Gao, "MCF: Mutual correction framework for semi-supervised medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15651–15660.

[24] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6819–6829.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, no. 11, pp. 1–27, 2008.

[26] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Oct. 2020, pp. 552–561.

[27] 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.

[28] 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. 12674–12684.

[29] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4258–4267.

[30] S. Gao, Z. Zhang, J. Ma, Z. Li, and S. Zhang, "Correlation-aware mutual learning for semi-supervised medical image segmentation," in *Proc. MICCAI*, Jan. 2023, pp. 98–108.

[31] F. Chen, J. Fei, Y. Chen, and C. Huang, "Decoupled consistency for semi-supervised medical image segmentation," in *Proc. MICCAI*, Jan. 2023, pp. 551–561.

[32] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[33] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. ICLR*, Jan. 2016, pp. 1–13.

[34] K. Wang et al., "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 450–460.

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

[36] 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.

[37] J. Miao et al., "SC-SSL: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 4, pp. 1347–1364, Apr. 2024.

[38] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. ICLR*, Jan. 2021, pp. 1–20.

[39] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4248–4257.

[40] J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, "Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels," in *Proc. NeurIPS*, vol. 34, Jan. 2021, pp. 16686–16699.

[41] C. Wang, H. Xie, Y. Yuan, C. Fu, and X. Yue, "Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 931–942.

[42] R. Gu et al., "Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 245–256, Jan. 2023.

[43] Y. Zhang, X. Zhang, J. Li, R. Qiu, H. Xu, and Q. Tian, "Semi-supervised contrastive learning with similarity co-calibration," *IEEE Trans. Multimedia*, vol. 25, pp. 1749–1759, 2023.

[44] J. Ma, C. Wang, Y. Liu, L. Lin, and G. Li, "Enhanced soft label for semi-supervised semantic segmentation," in *Proc. ICCV*, Oct. 2023, pp. 1185–1195.

[45] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8219–8228.

[46] Y. Zhou, H. Xu, W. Zhang, B. Gao, and P.-A. Heng, "C3-SemiSeg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7036–7045.

[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. Liu, W. Zhang, and J. Wang, "Adaptive multi-teacher multi-level knowledge distillation," *Neurocomputing*, vol. 415, pp. 106–113, Nov. 2020.

[49] Y. Gu, C. Deng, and K. Wei, "Class-incremental instance segmentation via multi-teacher networks," in *Proc. AAAI*, May 2021, vol. 35, no. 2, pp. 1478–1486.

[50] C. Pham, T. Hoang, and T. T. Do, "Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jul. 2023, pp. 6435–6443.

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

[52] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between CNN and transformer," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2022, pp. 820–833.

[53] 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. Intervent. (MICCAI)*, 2021, pp. 297–306.

[54] Z. Zhao, Z. Wang, L. Wang, D. Yu, Y. Yuan, and L. Zhou, "Alternate diverse teaching for semi-supervised medical image segmentation," in *Proc. ECCV*, Oct. 2024, pp. 227–243.

[55] L. Liu and R. T. Tan, "Certainty driven consistency loss on multi-teacher networks for semi-supervised learning," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108140.

[56] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2017, pp. 1285–1294.

[57] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15334–15342.

[58] H. J. Lee, J. U. Kim, S. Lee, H. G. Kim, and Y. M. Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4817–4826.

[59] D. Marin et al., "Efficient segmentation: Learning downsampling near semantic boundaries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2131–2141.

[60] F. Chen et al., "Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3779–3793, Dec. 2023.

[61] W. Kuo, A. Angelova, J. Malik, and T.-Y. Lin, "ShapeMask: Learning to segment novel objects by refining shape priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9206–9215.

[62] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 205–214.

[63] Y. Yuan, J. Xie, X. Chen, and J. Wang, "SegFix: Model-agnostic boundary refinement for segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2020, pp. 489–506.

[64] J. Chu et al., "Pay more attention to discontinuity for medical image segmentation," in *Proc. MICCAI*, vol. 12264, Jan. 2020, pp. 166–175.

[65] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3000–3009.

[66] J. Yuan, Z. Deng, S. Wang, and Z. Luo, "Multi receptive field network for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1883–1892.

[67] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8533–8542.

[68] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102395.

[69] R. Cong et al., "Boundary guided semantic learning for real-time COVID-19 lung infection segmentation system," *IEEE Trans. Consum. Electron.*, vol. 68, no. 4, pp. 376–386, Nov. 2022.

[70] K. Hu, X. Zhang, D. Lee, D. Xiong, Y. Zhang, and X. Gao, "Boundary-guided and region-aware network with global scale-adaptive for accurate segmentation of breast tumors in ultrasound images," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 9, pp. 4421–4432, Sep. 2023.

[71] Y. Wu et al., "BGM-net: Boundary-guided multiscale network for breast lesion segmentation in ultrasound," *Frontiers Mol. Biosci.*, vol. 8, Jul. 2021, Art. no. 698334.

[72] R. Xu et al., "BG-net: Boundary-guided network for lung segmentation on clinical CT images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8782–8788.

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

[74] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Dec. 2014.

[75] 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.

[76] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[77] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[78] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[79] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[80] F. Isensee, P. F. Jaeger, S. A. 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, Feb. 2021.

[81] N. Dvornik, J. Mairal, and C. Schmid, "Modeling visual context is key to augmenting object detection datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 364–380.

[82] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[83] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*, 2016, pp. 499–515.

[84] K. Clark et al., "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[85] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[86] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" 2021, *arXiv:2108.09038*.