Pseudo Label Optimization for Semi-Supervised Medical Image Segmentation

Zexuan Ji, *Member, IEEE*, Shunlong Ye, and Xiao Ma, *Student Member, IEEE*

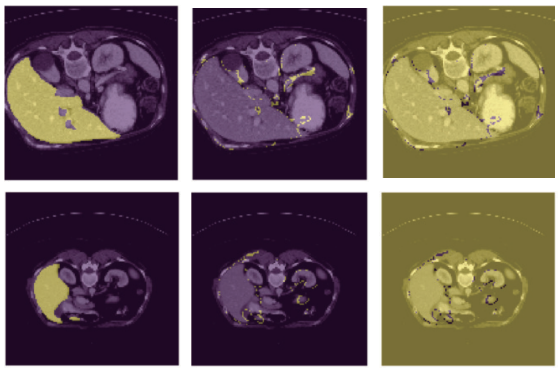
***Abstract*— Semi-supervised learning has emerged as a promising approach to leverage both labeled and unlabeled data due to the limited data annotations in medical image segmentation. Existing semi-supervised methods predominantly focus on high confidence pseudo labels, often neglecting the vast number of low confidence pseudo labels and the potential for improving pseudo labels quality. This paper introduces a novel approach that systematically leverages low confidence pseudo labels to address the limitations of conventional semi-supervised techniques. At the image level, we employ a superpixel algorithm and information entropy voting to ensure spatial coherence, while at the feature level, we utilize triplet loss to distinguish between similar and dissimilar regions. Furthermore, to enhance the overall quality of pseudo labels, we integrate a mutual correction framework, which supports iterative refinement and progressively improves segmentation outcomes. The proposed method achieves state-of-the-art results on two public datasets with different labeled data ratio and shows improvements on all baselines, demonstrating its effectiveness in improving segmentation performance and its potential applicability to a wide range of medical imaging tasks. Code is available at `https://github.com/yeshunlong/PLBOpt`.**

***Index Terms*— Medical image segmentation, semi-supervised learning, pseudo label, superpixel, triplet loss, mutual correction.**

I. Introduction

Medical image segmentation is crucial for various clinical applications [1]–[3], but acquiring labeled data for training deep learning models is challenging due to the scarcity of annotated medical images [4], [5]. This scarcity arises from the significant time and expertise required for precise annotation, often necessitating the involvement of highly specialized radiologists or pathologists. Consequently, this highlights the importance of semi-supervised learning in this field [6]–[8], as it enables the utilization of large amounts of unlabeled data to enhance model performance and accuracy, thereby alleviating the dependency on extensive labeled datasets [9]–[12].

Traditional semi-supervised medical image segmentation models utilize pseudo labels to leverage unlabeled data [13], [14]. However, in practice, these models often directly use predictions with the highest confidence [15], [16], neglecting the distribution of pseudo labels and resulting in suboptimal utilization. As depicted in Fig. 1, when pseudo labels are categorized based on confidence levels, we observe that low confidence pseudo labels delineate boundaries to a certain extent, indicating their significance in segmentation tasks. These low confidence pseudo labels often contain valuable boundary information that can improve the model's understanding of complex anatomical structures. Therefore, these low confidence pseudo labels should be prioritized for optimal utilization, which forms the core motivation of our approach. By refining these labels and incorporating them into the training process, we can further increase the quantity and quality of utilized pseudo labels, thereby enhancing the model's accuracy and robustness in medical image segmentation.



(a) label pixels (b) low conf. pixels (c) high conf. pixels

Fig. 1. Visualization of pseudo labels with different confidence levels. (a) Image label mask. (b) Low confidence pseudo labelsmask. (c) High confidence pseudo labels mask.

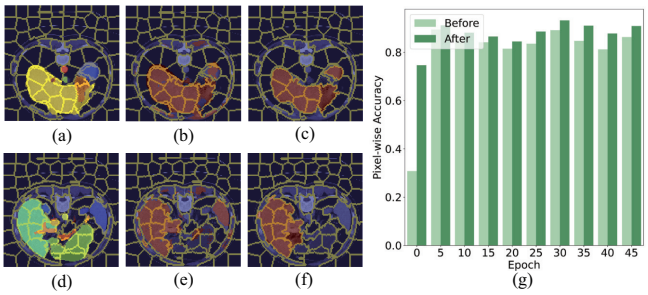


Fig. 2. Visualization of pseudo labels refinement results. (Organ: liver) (a-c) Original image, pseudo label, and refined pseudo labels (slice 1). (d-f) Original image, pseudo label, and refined pseudo label. (slice 2) (g) Correct pseudo label pixel ratio comparison. Yellow lines in (a-f) indicate superpixel boundary.

This work was supported by National Science Foundation of China under Grants No. 62072241. *(Corresponding author: Shunlong Ye.)*

Z. Ji, S. Ye and X. Ma are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jizexuan@njust.edu.cn; yeshunlong@njust.edu.cn; maxiao@njust.edu.cn).

In existing efforts to enhance the utilization of pseudo labels, most approaches focus on optimizing their distribution by predicting confidence levels [17], [18] or measuring uncertainty [19], [20]. However, these methods primarily operate only at the feature level and, while they do increase the number of pseudo labels used, they do not fundamentally improve the quality of these labels. Our approach is centered on enhancing both the quantity and quality of pseudo labels, aiming to provide a more comprehensive solution. As illustrated in Fig. 2, our method significantly improves the ability to capture anatomical structures and variations, particularly in complex regions like the liver. This refinement process not only reduces noise and inaccuracies but also enhances the model's generalization across different slices, leading to improved segmentation accuracy and reliability in clinical applications.

In this study, we introduce a novel training approach for semi-supervised medical image segmentation that aims to both increase the quantity and enhance the quality of pseudo labels. Specifically, our method is designed to systematically refine pseudo labels, thereby addressing the limitations of traditional approaches that primarily operate at the feature level. Our approach consists of three core components: superpixel refinement combined with information entropy voting at the image level, triplet loss applied at the feature level, and a mutual correction framework that iteratively improves pseudo labels.

From the perspective of improving utilization, we introduce a superpixel refinement process that ensures spatial coherence within the image. The superpixel algorithm segments the image into smaller, homogeneous regions, which are then evaluated using information entropy voting. This voting mechanism quantifies the uncertainty within each superpixel, allowing us to adjust the boundaries of the pseudo labels based on the entropy values. If the entropy within a superpixel exceeds a predefined threshold, the boundary is either contracted or expanded depending on the relative entropy of adjacent regions. This approach not only increases the number of usable pseudo labels but also ensures that these labels are more spatially consistent with the underlying image structures.

At the feature level, we employ triplet loss to further refine the pseudo labels by enhancing the model's ability to distinguish between similar and dissimilar regions. We hierarchically divide the pseudo labels into multiple confidence levels and use a feature extractor to compute the cosine similarity between boundary and internal pixels. By identifying the most similar boundary pixels (hard positives) and the most dissimilar internal pixels (hard negatives), we can effectively train the model to focus on boundary details. The triplet loss function penalizes the model if the feature representation of a boundary pixel is more similar to an internal pixel than to another boundary pixel, thereby promoting more accurate segmentation.

From the perspective of improving quality, we propose a mutual correction framework that systematically corrects errors in pseudo labels by leveraging the discrepancies between two identical subnetworks. These subnetworks are trained in parallel, and their outputs are compared to identify regions where the pseudo labels exhibit significant confidence differences. If one subnetwork assigns a high confidence to a region where the other subnetwork assigns low confidence, this discrepancy triggers a reevaluation of the pseudo label. Through this iterative process, the mutual correction framework progressively refines the pseudo labels, resulting in higher accuracy and consistency in the final segmentation results.

This paper introduces three main innovations:

- Firstly, we improve the quantity of usable pseudo labels by incorporating both image level (superpixel refinement and information entropy voting) and feature level (triplet loss) techniques, thereby enabling a more effective utilization of pseudo labels.
- Secondly, we enhance the quality of pseudo labels through the introduction of a mutual correction framework. This framework uses dual subnetworks to iteratively refine predictions, leading to more accurate pseudo labels and overall segmentation quality.
- Lastly, the proposed approach achieves state-of-the-art (SOTA) results on two publicly available datasets for multi-organ segmentation with varying labeled data ratios, confirming the method's efficacy and reliability. Moreover, the improvements on all baselines further demonstrate the versatility of our approach.

II. Related Work

A. Semi-supervised medical image segmentation

Semi-supervised medical image segmentation is a rapidly advancing field, driven by the need to utilize both labeled and unlabeled data effectively. This approach is particularly vital due to the scarcity of annotated medical images, which poses a significant challenge for training deep learning models [5]. The utilization of pseudo labels has become an essential strategy to mitigate this challenge, as it allows models to leverage the vast amount of unlabeled data available [9].

Several methods have been proposed to effectively utilize pseudo labels in different network architectures and scenarios. For instance, consistency learning under transformations has been explored to ensure that the model predictions remain stable under various image transformations, thereby enhancing the robustness of pseudo labels [7]. Another innovative approach involves cross teaching between convolutional neural networks (CNNs) and transformers, which leverages the complementary strengths of these architectures to improve segmentation performance [9]. Dual-task consistency is another technique employed to utilize pseudo labels effectively. This method ensures that the model maintains consistent predictions across different tasks, which in turn enhances the reliability of the pseudo labels [21]. Attention-based mechanisms have also been integrated into semi-supervised frameworks to focus on critical regions of the images [10]. Multi-task learning approaches have shown promise in leveraging pseudo labels by simultaneously optimizing for multiple related tasks. This not only improves segmentation accuracy but also enhances the model's ability to generalize to different medical imaging scenarios [11]. Additionally, causality-inspired semi-supervised learning has been proposed to incorporate causal inference

principles into the segmentation process, further refining the utilization of pseudo labels [8]. Furthermore, few-shot learning techniques have been adapted to the semi-supervised paradigm to handle scenarios with extremely limited labeled data. These methods aim to learn effective segmentation models from very few examples by making efficient use of pseudo labels [3].

B. Pseudo labels utilization and optimization

In the realm of semi-supervised medical image segmentation, the utilization and optimization of pseudo labels play a pivotal role in bridging the gap between labeled and unlabeled data. The current methodologies can be broadly classified into two main approaches: leveraging a larger number of pseudo labels and enhancing the quality of these labels.

To utilize more pseudo labels, various strategies have been developed. One common approach is to generate pseudo labels for the unlabeled data based on the model's predictions and then use these labels to retrain the model iteratively. This method is particularly effective in expanding the training dataset and improving the model's robustness. For instance, curriculum pseudo labeling, which introduces pseudo labels gradually according to their confidence levels, ensures a smoother learning process and better model performance [22], [23]. Another notable technique is the use of federated learning frameworks that incorporate pseudo labels denoising to enhance segmentation performance across distributed datasets without compromising data privacy [24], [25].

On the other hand, improving the quality of pseudo labels is crucial for achieving higher accuracy in segmentation tasks. Quality enhancement methods often involve refining the pseudo labels by addressing uncertainty and noise. Uncertainty-aware methods generate pseudo labels that account for prediction confidence, thereby filtering out unreliable labels and retaining high quality annotations for model training [20], [26]. Similarly, approaches like pseudo labeling with confirmation bias mitigation ensure that the pseudo labels are not only generated but also validated for consistency, reducing the risk of propagating errors through the training process [27].

Despite these advancements, existing methods face two significant challenges. Firstly, most approaches tend to focus on either increasing the number of pseudo labels or enhancing their quality, rarely addressing both aspects simultaneously. This single-faceted focus can limit the overall effectiveness of the semi-supervised learning process. Secondly, the majority of current methods concentrate at feature level optimization of pseudo labels, which might not fully capture the complexities and nuances of data at image level. For example, methods like prototype-based pseudo labeling and contrastive learning are primarily feature-driven and may overlook detailed boundary information crucial for precise segmentation.

III. Method

In this section, we present our proposed method for pseudo labels optimization in semi-supervised medical image segmentation. The method consists of three main components: pseudo labels refinement, triplet loss computation, and mutual correction framework. These components work together to enhance the utilization and quality of pseudo labels, thus improve the model's segmentation performance.

A. Pseudo labels refinement

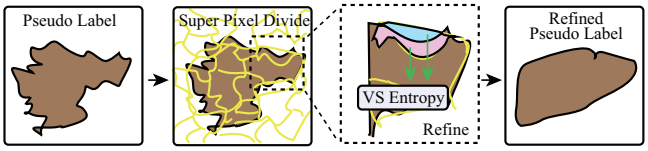


Fig. 3. Illustration of pseudo labels refinement process.

To formally describe the semi-supervised medical image segmentation problem, we denote the annotated dataset as $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_l}$, where $\mathbf{x}_i$ is the input image and $\mathbf{y}_i$ is its corresponding ground truth label. The unannotated dataset is denoted as $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=1}^{N_u}$, and the pseudo labels generated for $\mathcal{D}_u$ are denoted as $\hat{\mathbf{y}}_i$.

As illustrated in Fig. 3, to optimize pseudo labels, we first segment the pseudo labels into superpixels using the SLIC [28] algorithm, yielding $\mathcal{S} = \{S_i\}_{i=1}^{N_s}$, where S_i represents the ith superpixel. For each superpixel S_i, we calculate the intersection and complement with the pseudo labels mask, denoted as $\mathbf{M}_i^{int} = S_i \cap \hat{\mathbf{y}}$ and $\mathbf{M}_i^{comp} = S_i \setminus \hat{\mathbf{y}}$, respectively. We then compute the entropy of each region:

$$H(\mathbf{M}_i^{int}) = -\sum_{c=1}^{C} p(c|\mathbf{M}_i^{int}) \log p(c|\mathbf{M}_i^{int}) \tag{1}$$

$$H(\mathbf{M}_i^{comp}) = -\sum_{c=1}^{C} p(c|\mathbf{M}_i^{comp}) \log p(c|\mathbf{M}_i^{comp}) \tag{2}$$

where $p(c|\mathbf{M}_i^{int})$ and $p(c|\mathbf{M}_i^{comp})$ are the normalized class probabilities within $\mathbf{M}_i^{int}$ and $\mathbf{M}_i^{comp}$, respectively.

To decide whether to expand or contract the boundary of the pseudo label, we introduce a threshold parameter τ. If $H(\mathbf{M}_i^{int}) > \tau$ and $H(\mathbf{M}_i^{int}) > H(\mathbf{M}_i^{comp})$, it indicates that the current boundary of the pseudo labels is uncertain and should be contracted. Conversely, if $H(\mathbf{M}_i^{comp}) > \tau$ and $H(\mathbf{M}_i^{comp}) > H(\mathbf{M}_i^{int})$, the boundary is expanded. More expriments results on the influence of superpixel segments number and algorithm can be found in section IV.

B. Triplet loss

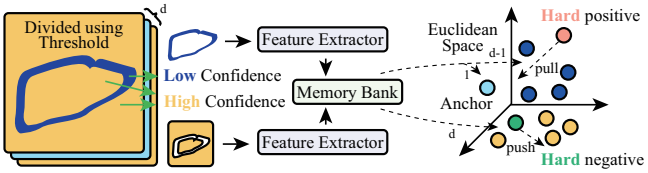


Fig. 4. Illustration of triplet loss computation process.

As shown in Fig. 4, assume the pseudo labels as $\hat{y}$ associated confidence levels as $c(\hat{y})$, we hierarchically divide these pseudo labels into k ($k = 3$ in our method) confidence

levels $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\}$ such that each subset $\hat{y}_i$ satisfies the condition:

$$\forall \hat{y}_i,\ \hat{y}_j \in \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\} \text{ and } i < j,\ \text{if } c(\hat{y}_i) < c(\hat{y}_j) \quad (3)$$

Each level pseudo labels must satisfy the following constraints: within the same domain, features of boundary pixels should be more similar to each other, while features of boundary pixels and internal pixels should be less similar. This ensures that the model learns to distinguish boundary details effectively.

Assume information within the same domain is extracted by a feature extractor $f(\cdot)$ for each pixel x. The similarity between two feature vectors $f(x_i)$ and $f(x_j)$ is measured using the cosine similarity:

$$\text{sim}(f(x_i), f(x_j)) = \frac{f(x_i) \cdot f(x_j)}{\|f(x_i)\| \|f(x_j)\|} \quad (4)$$

To query the most similar and most dissimilar feature vectors for boundary refinement, we define:

- Anchor $x_{i,a}$: A boundary pixel.
- Hard positive $x_{i,p}$: The boundary pixel with the highest cosine similarity within the same domain.
- Hard negative $x_{i,n}$: The internal pixel with the lowest cosine similarity within the same domain.

Thus, the hard positive and hard negative can be identified as follows:

$$x_{i,p} = \arg \max_{x_j \in \text{boundary}} \text{sim}(f(x_{i,a}), f(x_j)) \quad (5)$$

$$x_{i,n} = \arg \min_{x_j \in \text{internal}} \text{sim}(f(x_{i,a}), f(x_j)) \quad (6)$$

During this process, positive and negative information extracted by the feature extractor $f(\cdot)$ is stored in two memory bank $\mathcal{M}_p$ and $\mathcal{M}_n$, respectively. The memory bank holds feature vectors for efficient comparison and retrieval during the training process. It is updated using a queue mechanism with limited capacity C. When a new feature vector $f(x)$ is added, the oldest entry in the memory bank is removed if the capacity is exceeded. The update process can be described as:

$$\mathcal{M} \leftarrow \begin{cases} \mathcal{M} \cup \{f(x)\} & \text{if } |\mathcal{M}| < C \\ (\mathcal{M} \setminus \{\text{oldest entry}\}) \cup \{f(x)\} & \text{if } |\mathcal{M}| = C \end{cases} \quad (7)$$

Then, the triplet loss can be defined as follows:

$$\mathcal{L}_{triplet} = \sum_{i=1}^{N} \Big[\|f(x_{i,a}) - f(x_{i,p})\|^2 - \|f(x_{i,a}) - f(x_{i,n})\|^2 + \alpha \Big]_+ \quad (8)$$

where $x_{i,a}$ is an anchor boundary pixel, $x_{i,p}$ is the hard positive boundary pixel, $x_{i,n}$ is the hard negative internal pixel, $f(\cdot)$ is the feature extractor, $[z]_+ = \max(z, 0)$, and α is the margin hyperparameter.

This hierarchical pseudo labels division and triplet loss framework collectively enhance the model's ability to accurately capture and refine boundary details in semi-supervised medical image segmentation.

C. Mutual correction framework

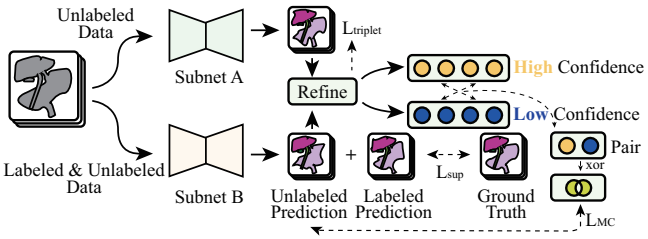


Fig. 5. Overview of the mutual correction framework with pseudo labels refinement and triplet loss computation.

Inspired by MCF [29], we employ two structurally identical subnetworks, Subnet A and Subnet B, to address potential errors in pseudo labeling, as depicted in Fig. 5. The principle is that if the confidence levels of pseudo labels for the same input are significantly different between the two networks, it indicates potential errors in the corresponding region predictions, prompting a reevaluation.

Let $\mathbf{p}_i^{(A)}$ and $\mathbf{p}_i^{(B)}$ represent the confidence scores for the pseudo labels generated by Subnet A and Subnet B, respectively. We denote high confidence by H and low confidence by L. For each pixel location j, if both networks show high confidence (H, H) or low confidence (L, L), it suggests either consistent certainty or uncertainty, providing limited additional information. However, a high confidence in one network and low confidence in the other $(H, L$ or $L, H)$ indicates a discrepancy that warrants further scrutiny.

The consistency requirement can be mathematically expressed as:

$$C(\mathbf{p}_i^{(A)}(j), \mathbf{p}_i^{(B)}(j)) = \begin{cases} 1 & \text{if } (\mathbf{p}_i^{(A)}(j) > \tau \text{ and } \mathbf{p}_i^{(B)}(j) \leq \tau) \\ & \text{or } (\mathbf{p}_i^{(A)}(j) \leq \tau \text{ and } \mathbf{p}_i^{(B)}(j) > \tau) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where τ is the confidence threshold. The mutual correction loss $\mathcal{L}_{MC}$ is calculated as:

$$\mathcal{L}_{MC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} C(\mathbf{p}_i^{(A)}(j), \mathbf{p}_i^{(B)}(j)) \left\| \hat{\mathbf{y}}_i^{(A)}(j) - \hat{\mathbf{y}}_i^{(B)}(j) \right\|^2 \quad (10)$$

where $\hat{\mathbf{y}}_i^{(A)}(j)$ and $\hat{\mathbf{y}}_i^{(B)}(j)$ are the pseudo labels generated by Subnet A and Subnet B at location j, respectively. The function $C(\mathbf{p}_i^{(A)}(j), \mathbf{p}_i^{(B)}(j))$ identifies locations with significant confidence discrepancies, emphasizing the correction of these regions. The mutual correction framework aims to enhance the robustness and accuracy of segmentation predictions by leveraging the disagreement between the two subnetworks to identify and rectify potential errors.

D. Overview of the proposed method

The overall loss function consists of three components: the triplet loss $\mathcal{L}_{triplet}$, the mutual correction loss $\mathcal{L}_{MC}$, and the supervised loss $\mathcal{L}_{sup}$. The coefficients for the triplet loss and mutual correction loss are updated using an exponential ramp-up strategy [30] to balance their contributions effectively. The total loss is formulated as:

$$\mathcal{L}_{total} = \lambda_{triplet} \cdot \mathcal{L}_{triplet} + \lambda_{MC} \cdot \mathcal{L}_{MC} + \mathcal{L}_{sup} \tag{11}$$

The training process is summarized in Algorithm 1, where the subnetworks Subnet A and Subnet B are iteratively updated to refine pseudo labels and enhance the model's segmentation performance.

Algorithm 1 Pseudocode for the entire training process.

Input: Annotated dataset $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_l}$, unannotated dataset $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=1}^{N_u}$, hyperparameters $\lambda_{triplet}$ and λ_{MC}
Output: Trained subnetworks Subnet A and Subnet B

Initialize Subnet A and Subnet B with random weights, positive memory bank $\mathcal{M}_p$, negative memory bank $\mathcal{M}_n$ and featrue extractor $f(\cdot)$
repeat
  Sample a mini-batch from $\mathcal{D}_l$
  $\mathbf{y}^{(B)} = \text{Subnet B}(\mathbf{x}_l)$
  Compute supervised loss $\mathcal{L}_{sup}$
  $\mathcal{L}_{sup} = \text{SupervisedLoss}(\mathbf{y}^{(B)}, \mathbf{y}_l)$
  Sample a mini-batch from $\mathcal{D}_u$
  $\hat{\mathbf{y}}^{(A)} = \text{Subnet A}(\mathbf{x}_u)$
  $\hat{\mathbf{y}}^{(B)} = \text{Subnet B}(\mathbf{x}_u)$
  Segment pseudo labels into superpixels
  $\mathcal{S} = \text{SLIC}(\mathbf{x}_u)$
  Refine pseudo labels using entropy comparison
  for each pseudo label $\hat{\mathbf{y}}_i$ in $\hat{\mathbf{y}}^{(A)}$ and $\hat{\mathbf{y}}^{(B)}$
    $\hat{\mathbf{y}}_i$ = PseudoLabelRefinement($\hat{\mathbf{y}}_i, \mathcal{S}$)
  end for
  Save refined pseudo labels to memory bank
  $x_{i,a}$ = $f(\hat{\mathbf{y}}^{(A)}) \cup f(\hat{\mathbf{y}}^{(B)})$
  MemorybankUpdate($\mathcal{M}_p, x_{i,a}$)
  MemorybankUpdate($\mathcal{M}_n, x_{i,a}$)
  Construct hard positive and hard negative triplets
  $x_{i,p}$ = CosineSimilarityQuery($\mathcal{M}_p, x_{i,a}$)
  $x_{i,n}$ = CosineDissimilarityQuery($\mathcal{M}_n, x_{i,a}$)
  Compute triplet loss $\mathcal{L}_{tri}$
  $\mathcal{L}_{triplet} = \text{Tri}(\hat{\mathbf{y}}^{(A)}, x_{i,p}, x_{i,n}) + \text{Tri}(\hat{\mathbf{y}}^{(B)}, x_{i,p}, x_{i,n})$
  Compute mutual correction loss $\mathcal{L}_{MC}$
  $\mathcal{L}_{MC} = \text{MutualCorrectionLoss}(\hat{\mathbf{y}}^{(A)}, \hat{\mathbf{y}}^{(B)})$
  Combine all losses to form the total loss
  $\mathcal{L}_{total} = \lambda_{triplet} \cdot \mathcal{L}_{triplet} + \lambda_{MC} \cdot \mathcal{L}_{MC} + \mathcal{L}_{sup}$
  Backpropagate the total loss and update Subnet A and Subnet B
until convergence

IV. Expriments

In this section, we present the experimental setup, including the datasets used, evaluation metrics. We then provide a comprehensive analysis of the results obtained and compare our method with SOTA models on two publicly available datasets.

A. Dataset and evaluation metrics

We conduct comprehensive experiments to evaluate the performance of our proposed approach on two widely-used and publicly available human organ segmentation datasets: Synapse [31] and AMOS [32].

The Synapse dataset encompasses 13 foreground classes, including spleen (Sp), right kidney (RK), left kidney (LK), gallbladder (Ga), esophagus (Es), liver (Li), stomach (St), aorta (Ao), inferior vena cava (IVC), portal & splenic veins (PSV), pancreas (Pa), right adrenal gland (RAG), and left adrenal gland (LAG). It consists of 30 axial contrast-enhanced abdominal CT scans, distributed randomly as 20, 4, and 6 scans for training, validation, and testing, respectively. The AMOS dataset comprises 360 scans, with a split of 216, 24, and 120 scans for training, validation, and testing, respectively, and includes three new classes: duodenum (Du), bladder (Bl), and prostate/uterus (P/U). We adopt DHC [33] as our baseline and compare our results with SOTA models reported in recent years using different labeled data ratio. In addtion, we employ three-fold cross-validation to evaluate our method on Synapse dataset considering the limited number of labeled data.

We evaluate the segmentation performance using the below metrics:

- The Dice coefficient is defined as
$$Dice = \frac{2 \times |X \cap Y|}{|X| + |Y|},$$
where X represents the predicted segmentation mask, Y is the ground truth mask, and $\cap$ denotes the intersection of the two sets. A higher Dice coefficient indicates better agreement between the predicted and ground truth segmentations, with a perfect score of 1 indicating complete overlap.
- The Average Surface Distance (ASD) metric quantifies the average distance between the surfaces of the predicted segmentation and the ground truth. It is defined as
$$ASD = \frac{1}{|X|} \sum_{x \in X} d(x, Y),$$
where $d(x, Y)$ represents the shortest Euclidean distance from point x on the predicted surface to the nearest point on the ground truth surface, and $|X|$ denotes the number of surface points in the predicted segmentation. A lower ASD value indicates better spatial agreement between the predicted and ground truth surfaces.

B. Results

We present the results of our method, as shown in Tables I, II, III, and IV. Our method achieves SOTA performance on most organ segmentation tasks in both datasets, indicating the reliability and effectiveness of our approach. Specifically, on the Synapse dataset with 10% labeled data (Table I), our method outperforms existing methods in terms of average Dice

TABLE I
COMPARISON OF DICE AND ASD ON THE SYNAPSE DATASET WITH 10% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓	Dice Of Each Class												
			Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PSV	PA	RAG	LAG
V-Net (Fully) [34]	62.0 ± 1.2	10.3 ± 3.9	84.6	77.2	73.8	73.3	38.2	94.6	68.4	72.1	71.2	58.2	48.5	17.9	29.0
UA-MT [6]	18.0 ± 2.9	57.6 ± 7.9	27.1	7.1	17.0	24.4	0.0	80.6	15.6	39.3	16.7	4.4	2.0	0.0	0.0
URPC [35]	24.0 ± 10.0	75.0 ± 24.1	46.8	44.7	36.5	0.2	0.0	**81.3**	12.5	48.0	36.3	0.0	5.5	0.0	0.0
CPS [13]	20.1 ± 0.6	59.1 ± 1.2	33.8	29.5	29.2	**41.2**	0.0	46.9	15.5	48.6	27.6	0.0	4.7	0.0	0.0
SS-Net [36]	17.5 ± 10.9	66.1 ± 20.7	45.6	11.6	42.3	2.4	0.0	74.5	6.0	32.6	2.8	0.0	0.0	3.8	5.8
DST [37]	20.9 ± 1.5	61.3 ± 1.4	43.3	32.8	16.0	24.9	0.0	75.8	22.4	27.6	19.4	3.8	5.0	0.3	0.0
DePL [38]	21.0 ± 2.0	58.4 ± 7.6	34.2	32.2	17.3	27.2	0.0	65.7	16.8	40.8	29.3	2.8	6.8	0.0	0.0
Adsh [39]	20.9 ± 1.7	55.8 ± 4.8	36.0	47.7	32.0	37.9	0.0	53.0	25.0	45.4	26.8	0.2	3.7	0.0	0.0
CReST [40]	17.3 ± 0.5	38.3 ± 3.9	32.7	35.3	29.5	24.7	0.3	42.5	19.7	45.2	18.9	4.2	6.3	4.8	1.7
SimiS [41]	25.0 ± 2.3	43.9 ± 1.7	42.0	38.6	27.2	19.7	0.0	74.2	16.5	**51.7**	35.0	13.6	5.4	0.0	1.8
Basak et al. [42]	25.3 ± 3.6	50.0 ± 6.5	40.9	42.3	19.2	35.2	0.0	75.7	19.2	44.7	32.8	5.0	10.4	3.5	0.0
CLD [43]	22.4 ± 1.0	49.7 ± 3.6	39.3	43.9	25.6	12.8	0.0	73.3	14.3	14.1	25.7	8.8	6.1	0.2	1.1
DHC [33]	28.6 ± 2.1	25.0 ± 3.4	49.7	49.1	28.1	23.3	0.0	46.9	14.3	29.9	44.0	15.7	**14.3**	5.5	8.9
Ours	**32.4 ± 2.2**	**19.5 ± 2.4**	**54.6**	**58.9**	**48.5**	40.8	**0.4**	59.7	**25.7**	32.9	**49.4**	**17.0**	13.0	**7.5**	**12.1**

TABLE II
COMPARISON OF DICE AND ASD ON THE AMOS DATASET WITH 2% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓	Dice Of Each Class													
			Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PA	RAG	LAG	Du	Bl	P/U
V-Net (Fully) [34]	76.5	2.0	92.2	92.2	93.3	65.5	70.3	95.3	82.4	91.4	85.0	74.9	58.6	58.1	65.6	64.4	58.3
UA-MT [6]	33.9	22.4	**62.5**	**61.7**	59.8	17.5	13.8	73.4	39.4	34.6	32.4	26.5	12.1	6.5	15.3	32.4	21.7
URPC [35]	38.3	37.5	60.8	57.7	56.5	**34.6**	0.0	**78.4**	**41.4**	53.3	49.6	**40.4**	0.0	0.0	**30.1**	42.5	**30.6**
CPS [13]	31.1	39.9	49.8	36.3	48.6	27.1	0.0	68.1	30.8	46.2	46.8	26.4	0.0	0.0	18.8	42.6	26.2
SS-Net [36]	17.4	59.0	37.7	20.1	26.3	9.0	3.3	57.1	25.1	28.4	28.2	0.0	0.0	0.0	0.0	26.5	0.2
DST [37]	30.6	31.8	49.6	34.6	40.3	27.3	2.2	55.4	30.1	45.7	41.7	24.8	20.3	2.4	18.3	41.9	24.5
DePL [38]	29.6	40.0	53.4	37.0	43.7	21.6	0.0	67.4	25.7	46.6	43.1	19.5	0.3	0.0	21.2	41.1	24.3
Adsh [39]	30.3	42.4	53.9	45.1	51.2	28.5	0.0	62.1	27.0	41.4	42.7	25.0	0.0	0.0	20.3	35.8	21.6
CReST [40]	26.9	27.1	37.7	50.1	42.2	5.7	8.8	46.7	29.1	35.0	38.2	19.9	7.1	11.3	15.1	35.6	21.0
SimiS [41]	36.8	26.1	57.8	58.6	58.6	22.9	0.0	70.9	38.0	52.0	47.0	32.4	20.2	**11.5**	18.1	39.9	25.5
B. et al. [42]	29.8	35.5	50.7	47.7	44.1	21.1	0.0	61.8	27.7	38.1	40.4	21.8	9.6	9.5	14.6	36.5	24.5
CLD [43]	36.2	27.6	55.8	55.8	59.1	23.9	0.0	69.9	38.2	50.1	44.5	32.3	18.9	9.2	18.8	42.2	24.9
DHC [33]	38.2	20.3	62.1	59.5	57.8	25.0	20.5	66.0	38.2	51.3	47.9	26.8	**26.4**	7.0	17.8	43.2	24.8
Ours	**41.0**	**17.6**	61.7	60.2	**60.8**	31.6	**21.6**	69.6	37.6	**58.8**	**52.8**	35.5	24.4	8.5	20.0	**45.1**	26.5

TABLE III
COMPARISON OF DICE AND ASD ON THE SYNAPSE DATASET WITH 20% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓	Dice Of Each Class												
			Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PSV	PA	RAG	LAG
V-Net (Fully) [34]	62.0 ± 1.2	10.3 ± 3.9	84.6	77.2	73.8	73.3	38.2	94.6	68.4	72.1	71.2	58.2	48.5	17.9	29.0
UA-MT [6]	20.2 ± 2.2	71.6 ± 7.4	48.2	31.7	22.2	0.0	0.0	81.2	29.1	23.3	27.5	0.0	0.0	0.0	0.0
URPC [35]	25.6 ± 5.1	72.7 ± 15.5	66.7	38.2	56.8	0.0	0.0	85.3	33.9	33.1	14.8	0.0	5.1	0.0	0.0
CPS [13]	33.5 ± 3.7	41.2 ± 9.1	62.8	55.2	45.4	35.9	0.0	**91.1**	31.3	41.9	49.2	8.8	14.5	0.0	0.0
SS-Net [36]	35.0 ± 2.8	50.8 ± 6.5	62.7	67.9	60.9	34.3	0.0	89.9	20.9	61.7	44.8	0.0	8.7	4.2	0.0
DST [37]	34.4 ± 1.6	37.6 ± 2.9	57.7	57.2	46.4	43.7	0.0	89.0	33.9	43.3	46.9	9.0	21.0	0.0	0.0
DePL [38]	36.2 ± 0.9	36.0 ± 0.8	62.8	61.0	48.2	54.8	0.0	90.2	36.0	42.5	48.2	10.7	17.0	0.0	0.0
Adsh [39]	35.2 ± 0.5	39.6 ± 4.6	55.1	59.6	45.8	52.2	0.0	89.4	32.8	47.6	53.0	8.9	14.4	0.0	0.0
CReST [40]	38.3 ± 3.4	22.8 ± 9.0	62.1	64.7	53.8	43.8	8.1	85.9	27.2	54.4	47.7	14.4	13.0	18.7	4.6
SimiS [41]	40.0 ± 0.6	32.9 ± 0.5	62.3	69.4	50.7	61.4	0.0	87.0	33.0	59.0	57.2	29.2	11.8	0.0	0.0
Basak et al. [42]	33.2 ± 0.6	43.7 ± 2.5	57.4	53.8	48.5	46.9	0.0	87.8	28.7	42.3	45.4	6.3	15.0	0.0	0.0
CLD [43]	41.0 ± 1.2	32.1 ± 3.3	62.0	66.0	59.3	61.5	0.0	89.0	31.7	62.8	49.4	28.6	18.5	0.0	5.0
DHC [33]	48.6 ± 0.9	10.7 ± 2.6	62.8	69.5	59.2	66.0	13.2	85.2	**36.9**	**67.9**	**61.5**	**37.0**	30.9	**31.4**	10.6
Ours	**50.0 ± 0.2**	**9.6 ± 1.1**	**83.5**	**78.9**	**76.7**	**74.2**	**21.4**	83.9	35.4	61.4	40.7	33.0	**33.3**	6.4	**19.6**

and average ASD, achieving an average Dice of 32.4 and an average ASD of 19.5. It demonstrates superior performance in segmenting organs such as Sp, RK, LK, and Ga, with notable improvements in Dice scores compared to the baseline. Similarly, on the AMOS dataset with 2% labeled data (Table II), our method surpasses other methods, achieving an average Dice of 41.0 and an average ASD of 17.6. This performance is particularly evident in organs like Es, Ao, and IVC, where our method achieves significantly higher Dice scores.

Furthermore, with 20% labeled data on the Synapse dataset (Table III), our method continues to demonstrate robust performance with an average Dice of 50.0 and an average ASD of 9.6. This indicates that our method can effectively utilize refined pseduo label to improve segmentation accuracy. Notably, the performance improvements are significant in the segmentation of organs such as Li, St, and Ao. On the AMOS dataset with 5% labeled data (Table IV), our method achieves an average Dice of 50.1 and an average ASD of 8.6, outperforming other methods significantly. The results highlight the model's capability to accurately segment complex structures with minimal labeled data, as seen with high Dice scores in organs like RK, LK, and Ga.

TABLE IV
COMPARISON OF DICE AND ASD ON THE AMOS DATASET WITH 5% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓	Dice Of Each Class														
			Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PA	RAG	LAG	Du	Bl	P/U
V-Net (Fully) [34]	76.5	2.0	92.2	92.2	93.3	65.5	70.3	95.3	82.4	91.4	85.0	74.9	58.6	58.1	65.6	64.4	58.3
UA-MT [6]	42.1	15.4	59.8	64.9	64.0	35.3	34.1	77.7	37.8	61.0	46.0	33.3	26.9	12.3	18.1	29.7	31.6
URPC [35]	44.9	27.4	67.0	64.2	67.2	36.1	0.0	83.1	45.5	67.4	54.4	**46.7**	0.0	29.4	**35.2**	44.5	33.2
CPS [13]	41.0	20.3	56.1	60.3	59.4	33.3	25.4	73.8	32.4	65.7	52.1	31.1	25.5	6.2	18.4	40.7	35.8
SS-Net [36]	33.8	54.7	65.4	68.3	69.9	37.8	0.0	75.1	33.2	68.0	56.6	33.5	0.0	0.0	0.0	0.2	0.2
DST [37]	41.4	21.1	58.9	63.3	63.8	37.7	29.6	74.6	36.1	66.1	49.9	32.8	13.5	5.5	17.6	39.1	33.1
DePL [38]	41.9	20.4	55.7	62.4	57.7	36.6	31.3	68.4	33.9	65.6	51.9	30.2	23.3	10.2	20.9	43.9	37.7
Adsh [39]	40.3	24.5	56.0	63.6	57.3	34.7	25.7	73.9	30.7	65.7	51.9	27.1	20.2	0.0	18.6	43.5	35.9
CReST [40]	46.5	14.6	66.5	64.2	65.4	36.0	32.2	77.8	43.6	68.5	52.9	40.3	24.7	19.5	26.5	43.9	36.4
SimiS [41]	47.2	11.5	**77.4**	72.5	68.7	32.1	14.7	86.6	46.3	**74.6**	54.2	41.6	24.4	17.9	21.9	**47.9**	28.2
B. et al. [42]	38.7	31.7	68.8	59.0	54.2	29.0	0.0	83.7	39.3	61.7	52.1	34.6	0.0	0.0	26.8	45.7	26.2
CLD [43]	46.1	15.8	67.2	68.5	71.4	41.0	21.0	76.1	42.4	69.8	52.1	37.9	24.7	23.4	22.7	38.1	35.2
DHC [33]	49.5	13.8	68.1	69.6	71.1	**42.3**	37.0	76.8	43.8	70.8	57.4	43.2	**27.0**	28.7	29.1	41.4	**36.7**
Ours	**50.1**	**8.6**	69.3	**76.8**	**77.6**	31.6	**45.1**	**87.1**	**47.3**	68.8	**58.0**	46.2	24.4	**37.7**	32.1	29.1	16.9

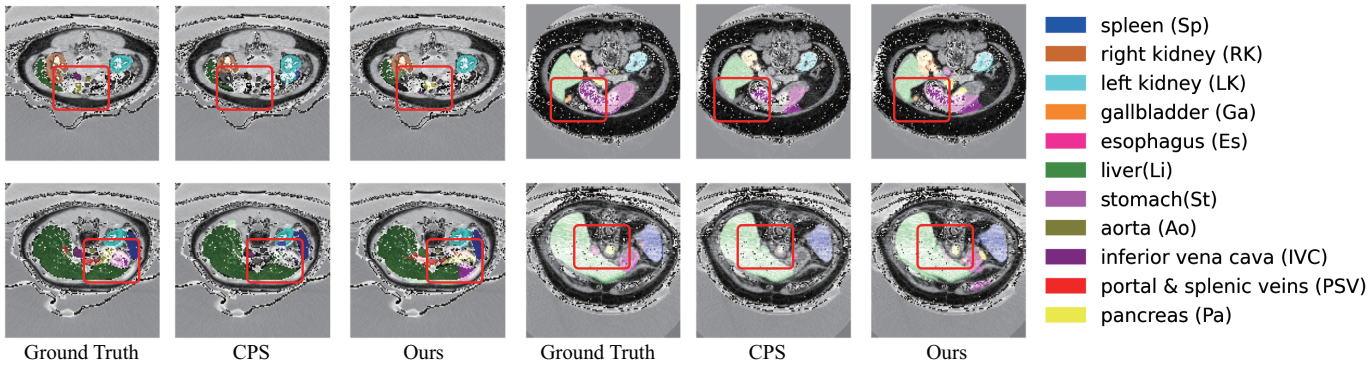


Fig. 6. Visualization of segmentation results on the Synapse dataset. The regions marked by red rectangles indicate the regions where the segmentation results are clearly different.

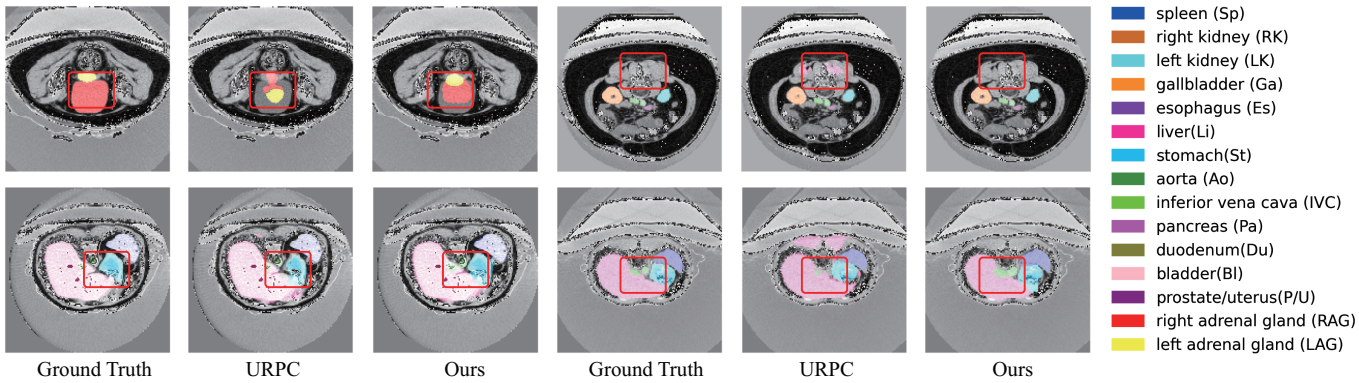


Fig. 7. Visualization of segmentation results on the AMOS dataset. The regions marked by red rectangles indicate the regions where the segmentation results are clearly different.

By analyzing Fig. 6 and Fig. 7, our method demonstrates excellent capability in accurately segmenting organ contours, showcasing superior attention to detail and the ability to differentiate between fine contours and blurry boundaries compared to other models. This indicates that our method effectively learns meaningful features through the process of optimizing pseudo labels and enhancing pseudo labels quality. Moreover, the visualizations reveal that our approach consistently improves the sharpness and accuracy of segmentation across different datasets.

C. Abalation study

Table V presents the results of our ablation study on the Synapse dataset. We compare the performance of our method with and without the components of superpixel refinement with information entropy voting, triplet loss, and mutual correction. The results demonstrate that each component contributes significantly to the overall performance improvement of our method.

Specifically, incorporating only the superpixel refinement with information entropy voting increases the average Dice score from 28.6 to 31.1 and reduces the average ASD from

TABLE V
ABLATION STUDY ON THE SYNAPSE DATASET WITH 10% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓
Baseline (DHC [33])	28.6 ± 2.1	25.0 ± 3.4
Only Refinement	31.1 ± 3.2	21.4 ± 3.0
Only Triplet Loss	29.7 ± 4.6	24.0 ± 9.0
Only Mutual Correction	29.8 ± 4.9	24.5 ± 7.5
Full	**32.4 ± 2.2**	**19.5 ± 2.4**

25.0 to 21.4, indicating a substantial improvement in segmentation accuracy and boundary delineation. The introduction of triplet loss alone yields an average Dice score of 29.7 and an average ASD of 24.0, showcasing its effectiveness in enhancing feature similarity constraints. Similarly, the mutual correction mechanism achieves an average Dice score of 29.8 and an average ASD of 24.5, highlighting its role in rectifying pseudo labels errors. Combining all components results in the highest performance, with an average Dice score of 32.4 and an average ASD of 19.5, thereby demonstrating the synergistic effect of the proposed methods. These quantitative improvements underscore the importance of each component in the overall architecture and its contribution to achieving SOTA results.

D. Performance comparison with different superpixel

TABLE VI
PERFORMANCE COMPARISON WITH DIFFERENT SUPERPIXEL METHODS ON THE SYNAPSE DATASET WITH 10% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Avg. Dice ↑	Avg. ASD ↓
Baseline (DHC [33])	28.6 ± 2.1	25.0 ± 3.4
Compact Watershed [44]	28.5 ± 2.7	22.0 ± 3.3
Felzenszwalb [45]	29.7 ± 3.3	29.2 ± 14.0
Quickshift [46]	31.3 ± 1.8	**17.8 ± 3.6**
SLIC [28]	**32.4 ± 2.2**	19.5 ± 2.4

Table VI reveals the impact of various superpixel segmentation methods on the Synapse dataset. SLIC achieves the highest performance with an average Dice score of 32.4 and an average ASD of 19.5. While all methods outperform the baseline, SLIC stands out as the most effective.

TABLE VII
PERFORMANCE COMPARISON WITH DIFFERENT SUPERPIXEL NUMBER ON SYNAPSE DATASET WITH 10% LABELED DATA AND AMOS DATASET WITH 2% LABELED DATA. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	Segment Number	Avg. Dice ↑	Avg. ASD ↓
Synapse	$H \times W$ / 10	30.6 ± 4.37	25.2 ± 7.66
	$H \times W$ / 50	**32.4 ± 2.2**	**19.5 ± 2.4**
	$H \times W$ / 100	29.8 ± 4.92	24.5 ± 7.5
AMOS	$H \times W$ / 10	39.6	21.7
	$H \times W$ / 50	**41.0**	**17.6**
	$H \times W$ / 100	39.1	18.2

Table VII illustrates the performance of our method on the Synapse and AMOS datasets with varying superpixel numbers. For both datasets, the model achieves optimal performance with $H \times W$ / 50 segments. This indicates that the number of superpixels plays a crucial role in the segmentation accuracy and pseudo labels refinement process. A moderate number of superpixels can effectively capture the structural information of the organs and facilitate the refinement of pseudo labels. In contrast, an excessive number of superpixels may introduce noise and hinder the model's ability to learn meaningful features.

E. Accssment of efficacy on different baselines

Comparing with different baselines on the Synapse and AMOS datasets (Table VIII), our method consistently outperforms the baselines, demonstrating its versatility and effectiveness. On Synapse, our method significantly improves the average Dice score and reduces the average ASD across most baselines, such as enhancing CLD from 22.4 to 32.0 in Dice and reducing ASD from 49.7 to 18.0. Similarly, on AMOS, the model enhances performance, exemplified by DHC's Dice score increases from 38.2 to 41.0 and ASD reduces from 20.3 to 17.6.

F. Input perturbation analysis

As shown in Table IX, our method demonstrates superior robustness against perturbations on the Synapse dataset, achieving higher segmentation accuracy under both weak and strong noise conditions. Specifically, with weak noise, our method attains an average Dice score of 27.8 and an average ASD of 22.1, outperforming the baseline (23.9, 29.0). Under strong noise, our method achieves 24.2 and 34.2, compared to the baseline's 20.6 and 30.5. This resilience is attributed to the superpixel and mutual correction framework, effectively mitigating the impact of perturbations.

V. DISCUSSION

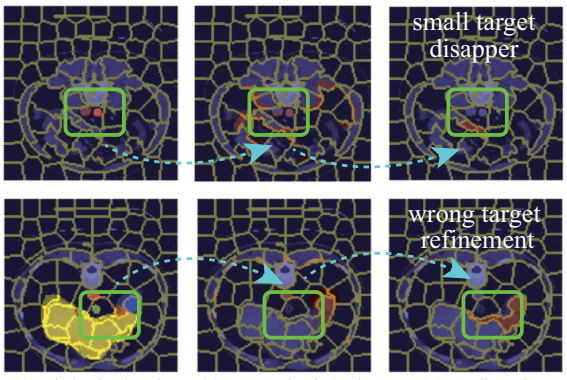


Fig. 8. Failure cases of our method on the Synapse dataset. (a) Original image. (b) Pseudo label. (c) Refined pseudo label.

Analyzing Fig. 8, we can identify several issues with our method. The primary issue lies in the reliance on pseudo labels refinement using superpixel methods during the early stages of training when the model has not yet converged. Initially, this refinement significantly improves the quality of pseudo labels.

TABLE VIII
PERFORMANCE COMPARISON WITH DIFFERENT BASELINES ON THE SYNAPSE DATASET WITH 10% LABELED DATA AND AMOS DATASET WITH 2% LABELED DATA. ↗ AND ↘ INDICATE INCREASE AND DECREASE, RESPECTIVELY.

Method	Synapse				AMOS			
	Before		After		Before		After	
	Avg. Dice ↑	Avg. ASD ↓	Avg. Dice ↑	Avg. ASD ↓	Avg. Dice ↑	Avg. ASD ↓	Avg. Dice ↑	Avg. ASD ↓
URPC [35]	24.0 ± 10.0	75.0 ± 24.1	24.8 ± 10.3 ↗	74.0 ± 25.2 ↘	38.3	37.5	38.9 ↗	30.8 ↘
CPS [13]	20.1 ± 0.6	59.1 ± 1.2	21.4 ± 1.1 ↗	57.2 ± 2.1 ↘	31.1	39.9	31.6 ↗	40.6 ↗
DST [37]	20.9 ± 1.5	61.3 ± 1.4	21.6 ± 0.4 ↗	51.7 ± 9.7 ↘	30.6	31.8	32.2 ↗	40.5 ↗
DePL [38]	21.0 ± 2.0	58.4 ± 7.6	21.6 ± 0.8 ↗	53.8 ± 2.3 ↘	29.6	40.0	30.6 ↗	35.7 ↘
Adsh [39]	20.9 ± 1.7	55.8 ± 4.8	22.0 ± 1.1 ↗	51.1 ± 7.2 ↘	30.3	42.4	34.9 ↗	26.5 ↘
CReST [40]	17.3 ± 0.5	38.3 ± 3.9	19.3 ± 4.9 ↗	31.7 ± 2.0 ↘	26.9	27.1	31.8 ↗	25.6 ↘
SimiS [41]	25.0 ± 2.3	43.9 ± 1.7	28.7 ± 2.8 ↗	25.8 ± 3.3 ↘	36.8	26.1	39.9 ↗	22.5 ↘
CLD [43]	22.4 ± 1.0	49.7 ± 3.6	32.0 ± 2.3 ↗	18.0 ± 1.8 ↘	36.2	27.6	37.5 ↗	26.0 ↘
DHC [33]	28.6 ± 2.1	25.0 ± 3.4	32.4 ± 2.2 ↗	19.5 ± 2.4 ↘	38.2	20.3	41.0 ↗	17.6 ↘

TABLE IX
INPUT PERTURBATION ANALYSIS ON THE SYNAPSE DATASET WITH 10% LABELED DATA, WHERE WEAK NOISE ORIGINATES FROM RANDOM VARIATIONS IN HUE, SATURATION, AND CONTRAST, WHILE STRONG NOISE IS GENERATED BY THE ADDITION OF RANDOM GAUSSIAN NOISE.

Method	Noise level	Avg. Dice ↑	Avg. ASD ↓
Baseline (DHC [33])	weak	23.9 ± 2.2	29.0 ± 2.4
	strong	20.6 ± 1.9	30.5 ± 2.4
Ours	weak	27.8 ± 2.2	22.1 ± 3.2
	strong	24.2 ± 5.2	34.2 ± 15.4

However, as the model converges, the effectiveness of pseudo labels refinement becomes constrained by the limitations of the superpixel segmentation technique. For small targets, an insufficient number of superpixels can result in the disappearance of these targets during the refinement process, thus degrading the quality of pseudo labels (as illustrated in the first example in Fig. 8). Moreover, the precision of pseudo labels refinement remains constant throughout the training process. While this leads to substantial improvements during the early stages, it becomes less beneficial once the model has converged. In some cases, it can even introduce misleading boundary results (as shown in the second example in Fig. 8).

Additionally, defining and selecting the most similar and most dissimilar pixel points for triplet loss poses a noteworthy challenge. Improper selection can render the triplet loss ineffective for model optimization and may cause instability in training. Furthermore, the mutual correction mechanism depends on the predictions of two subnetworks. If both subnetworks produce inaccurate predictions in certain regions, the correction mechanism might fail to function effectively and could potentially introduce additional errors. Setting an appropriate threshold to determine the prediction differences between the two subnetworks is also crucial.

In future, we plan to explore more advanced pseudo labels optimization algorithms, employ more precise methods to evaluate the impact of pseudo labels refinement on label quality, and investigate adaptive confidence level categorization methods to further enhance the model's performance.

VI. CONCLUSION

In this paper, we propose a novel training approach for semi-supervised medical image segmentation, which improves performance by addressing the underutilization of low confidence pseudo labels and enhancing pseudo labels quality. Our method integrates superpixel refinement with information entropy voting to leverage low confidence pseudo labels at the image level. At the feature level, triplet loss is employed to enforce similarity constraints among boundary and internal features, maximizing the utilization of pseudo labels. Additionally, mutual correction utilizes dual subnetworks to iteratively identify and rectify pseudo labels errors, thereby improving overall pseudo labels quality. Our method achieves SOTA results on two publicly available datasets for multi-organ segmentation, validating its effectiveness and reliability. Furthermore, our ablation study and performance comparison with different superpixel methods demonstrate the importance of each component in the overall architecture and the impact of superpixel segmentation on segmentation accuracy. Our method also exhibits improvements on different baselines and robustness against input perturbations, highlighting its potential for real-world applications.

REFERENCES

[1] M. Van Eijnatten, R. van Dijk, J. Dobbe, G. Streekstra, J. Koivisto, and J. Wolff, "Ct image segmentation methods for bone used in medical additive manufacturing," *Medical engineering & physics*, vol. 51, pp. 6–16, 2018.

[2] H. R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "Deep learning and its application to medical image segmentation," *Medical Imaging Technology*, vol. 36, no. 2, pp. 63–71, 2018.

[3] A. R. Feyjie, R. Azad, M. Pedersoli, C. Kauffman, I. B. Ayed, and J. Dolz, "Semi-supervised few-shot learning for medical image segmentation," *arXiv preprint arXiv:2003.08462*, 2020.

[4] C. You, W. Dai, F. Liu, Y. Min, H. Su, X. Zhang, X. Li, D. A. Clifton, L. Staib, and J. S. Duncan, "Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels," *arXiv preprint arXiv:2209.13476*, 2022.

[5] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin, "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation," *Computers in Biology and Medicine*, p. 107840, 2023.

[6] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 605–613.

[7] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. De Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 810–818.

[8] J. Miao, C. Chen, F. Liu, H. Wei, and P.-A. Heng, "Caussl: Causality-inspired semi-supervised learning for medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 426–21 437.

[9] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 820–833.

[10] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer, 2018, pp. 370–378.

[11] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. Van Tulder, and M. De Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer, 2019, pp. 457–465.

[12] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE transactions on medical imaging*, vol. 41, no. 3, pp. 608–620, 2021.

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

[14] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 614–623.

[15] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7236–7246.

[16] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 786–19 797.

[17] H. Yao, R. Chen, W. Chen, H. Sun, W. Xie, and X. Lu, "Pseudo-label-based unreliable sample learning for semi-supervised hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[18] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4248–4257.

[19] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9092–9101.

[20] L. Lu, M. Yin, L. Fu, and F. Yang, "Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation," *Biomedical Signal Processing and Control*, vol. 79, p. 104203, 2023.

[21] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 8801–8809.

[22] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[23] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 6912–6920.

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

[25] H. Wu, B. Zhang, C. Chen, and J. Qin, "Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning," *IEEE Transactions on Medical Imaging*, 2023.

[26] J. Wu, G. Wang, R. Gu, T. Lu, Y. Chen, W. Zhu, T. Vercauteren, S. Ourselin, and S. Zhang, "Upl-sfda: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[27] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[29] Y. Wang, B. Xiao, X. Bi, W. Li, and X. Gao, "Mcf: Mutual correction framework for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 651–15 660.

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

[31] B. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. R. Langerak, and A. Klein, "2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[32] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 722–36 732, 2022.

[33] H. Wang and X. Li, "Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 582–591.

[34] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[35] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 318–329.

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

[37] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 424–32 437, 2022.

[38] X. Wang, Z. Wu, L. Lian, and S. X. Yu, "Debiased learning from naturally imbalanced pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 647–14 657.

[39] L.-Z. Guo and Y.-F. Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8082–8094.

[40] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 857–10 866.

[41] H. Chen, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, M. Savvides, and B. Raj, "An embarrassingly simple baseline for imbalanced semi-supervised learning," *arXiv preprint arXiv:2211.11086*, 2022.

[42] H. Basak, S. Ghosal, and R. Sarkar, "Addressing class imbalance in semi-supervised image segmentation: A study on cardiac mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 224–233.

[43] Y. Lin, H. Yao, Z. Li, G. Zheng, and X. Li, "Calibrating label distribution for class-imbalanced barely-supervised knee segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 109–118.

[44] L. Najman and M. Schmitt, "Watershed of a continuous function," *Signal processing*, vol. 38, no. 1, pp. 99–112, 1994.

[45] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, pp. 167–181, 2004.

[46] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*. Springer, 2008, pp. 705–718.