Adaptive Learning of High-Value Regions for Semi-Supervised Medical Image Segmentation

Tao Lei[1*], Ziyao Yang[1], Xingwu Wang[2], Yi Wang[1], Xuan Wang[3], Feiman Sun[4], Asoke K. Nandi[5]
[1]Shaanxi University of Science and Technology, China
[2]Beijing University of Posts and Telecommunications, China
[3]Northwestern Polytechnical University, China
[4]Southwest Jiaotong University, China [5]Brunel University of London, UK
[*]Corresponding author:leitao@sust.edu.cn

Abstract

Existing semi-supervised learning methods typically mitigate the impact of unreliable predictions by suppressing low-confidence regions. However, these methods fail to explore ***which regions hold higher learning value*** *and* ***how to design adaptive learning strategies for these regions****. To address these issues, we propose a novel adaptive learning of high-value regions (ALHVR) framework. By exploiting the diversity of predictions from dual-branch networks, the prediction regions are classified into three groups: reliable stable region, reliable unstable region, and unreliable stable region. For high-value regions (reliable unstable region and unreliable stable region), different training strategies are designed. Specifically, for reliable unstable region, we propose a confidence-guided cross-prototype consistency learning (CG-CPCL) module, which enforces prototype consistency constraints in the feature space. By leveraging confidence information, the high-confidence predictions from one network selectively supervise the low-confidence predictions from the other, thus helping the model learn inter-class discrimination more stably. Additionally, for unreliable stable region, we design a dynamic teacher competition teaching (DTCT) module, which dynamically selects the most reliable pixels as teachers by evaluating the unperturbed predictions from both networks. These selected pixels are then used to supervise perturbed predictions, thereby enhancing the model's learning capability in unreliable region. Experimental results show that our method outperforms state-of-the-art approaches on three public datasets. Code is available at https://github.com/ziziyao/ALHVR.*

1. Introduction

Medical images contain various types of tissues with complex anatomical structures, leading to ambiguous boundaries between target regions and surrounding tissues. This issue increases the segmentation difficulty for the model at boundary regions. Some semi-supervised methods [23, 34] estimate confidence based on predictions from a single network to distinguish between high-confidence and low-confidence pixels. These methods guide the model to focus on high-confidence predictions by discarding low-confidence pixels. However, predictions from a single network are biased, leading to inaccurate pixel classification. Therefore, some approaches evaluate pixel reliability based on dual-network predictions. For instance, MCF [27] treats inconsistent predictions between networks as errors and attempts to correct them, but it fails to handle cases where both predictions are incorrect. MLRP [24] employs a pixel-wise comparison strategy to select relatively reliable predic-

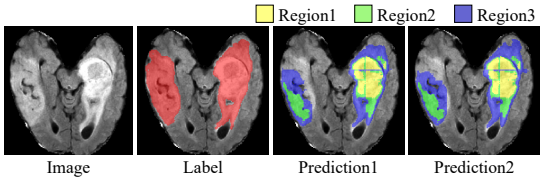

Figure 1. Training on the Brats dataset using the CPS framework, where the predictions of a given sample are classified based on a confidence threshold. Region 1 represents the reliable stable region, Region 2 represents the reliable unstable region, and Region 3 represents the unreliable stable region.

Dataset	Region 1	Region 2	Region 3	Dice
ACDC	✓	✓	✓	85.34
	✓			87.13
		✓		88.46
			✓	88.20
	✓	✓		87.41
	✓		✓	82.51
		✓	✓	**88.58**

Table 1. Experimental results of training on the ACDC dataset with 10% labeled data using the CPS framework under different masked region settings.

tions from two networks as pseudo-labels. However, mutual learning between low-confidence predictions may introduce biases, leading to suboptimal learning. These methods fail to identify or utilize high-value regions, particularly those with high segmentation difficulty. Furthermore, they lack adaptive learning strategies tailored to distinct regions, limiting the model's ability to learn in crucial segmentation regions, such as tissue boundaries.

In order to resolve the aforementioned issues, this paper combines the diversity of dual-network predictions with confidence thresholds to achieve more precise pixel classification. Taking CPS-based methods [6] as an example, we analyze this problem. When two networks predict the same image, the confidence of each pixel relative to a predefined threshold can be categorized into three types: (1) both networks have confidence scores above the threshold, (2) one network has a confidence score above the threshold while the other is below it, and (3) both networks have confidence scores below the threshold. Based on confidence levels and prediction consistency, we define these three regions as the reliable stable region, the reliable unstable region, and the unreliable stable region, respectively. As shown in Fig. 1, different regions are represented by different colors. To further explore the most valuable learning regions, we conducted experiments on the ACDC dataset using the CPS framework. Specifically, we applied masks to occlude different regions, training the model only on unmasked areas. As shown in Tab. 1, the highest Dice score is achieved when pixels in the reliable stable region are masked out during training process. This indicates that the reliable unstable region and the unreliable stable region are high-value regions.

To address the lack of adaptive learning strategies for regions of varying value in existing methods, this paper proposes a semi-supervised medical image segmentation method based on adaptive learning for high-value regions (ALHVR). By analyzing the confidence distributions of model predictions, high-value regions are identified, and region-specific training strategies are designed to improve segmentation performance. Specifically, we propose the confidence-guided cross-prototype consistency learning (CG-CPCL) module and the dynamic teacher competition teaching (DTCT) module, which are designed to train the reliable unstable region and the unreliable stable region, respectively. The main contributions of this paper are as follows:

(1) We conduct an investigation into the high-value regions in semi-supervised medical image segmentation, categorizing them into reliable unstable region and unreliable stable region. Furthermore, we design adaptive learning strategies to enhance segmentation performance in these areas.

(2) We propose CG-CPCL to address the challenge where unstable predictions within a region prevent the model from accurately recognizing key region features. CG-CPCL enhances cross-consistency between features and prototypes by constructing class prototypes while incorporating confidence constraints to ensure that low-confidence predictions learn exclusively from high-confidence predictions. This approach effectively guides the model to learn more reliable feature representations in reliable unstable region, thereby improving its ability to identify key regions.

(3) We design DTCT to tackle the issue where unreliable predictions within a region lead to the accumulation of erroneous signals, causing prediction bias. DTCT employs pixel-level reliability assessment to dynamically and competitively select more reliable pixels from unperturbed predictions as teacher predictions to supervise perturbed predictions. This method provides the model with more valuable and reliable learning signals, effectively reducing prediction bias.

(4) The proposed framework integrates both learning modules and is evaluated on three public datasets: ACDC, AbdomenCT-1K, and Brats. Experimental results demonstrate that our framework outperforms current state-of-the-art semi-supervised learning methods.

2. Related Work

2.1. Semi-Supervised Medical Image Segmentation

The key challenge in semi-supervised image segmentation is to effectively exploit unlabeled data to improve model performance. To address this, researchers have proposed various approaches, among which consistency regularization and pseudo-labeling methods have become mainstream directions.

Firstly, consistency regularization methods are based on the smoothness assumption, which states that a model's predictions should remain similar under different levels of perturbations, including data-level, model-level, and feature-level perturbations. At the data level, common perturbations involve transformations such as rotation and color jittering. To generate more diverse samples, copy-paste techniques are popular, such as CutMix [35] and ClassMix [20]. However, existing copy-paste methods overlook the distribution mismatch between labeled and unlabeled data. To tackle this issue, BCP [2] introduces bidirectional copy-paste between labeled and unlabeled data to generate new samples. At the model level, perturbations are typically introduced by modifying model parameters or structures. For example, MT [25] employs an exponential moving average (EMA) to update the teacher model's parameters with those of the student model, CPS [6] initializes two networks with different parameters and trains them independently. Additionally, MCF [27] and CC-Net [10] introduce variations in encoder or decoder structures, ensuring that different net-

work architectures produce consistent predictions. At the feature level, CCT [21] injects noise or drops activations to enforce consistency between decoders under perturbations. PS-MT [14] uses T-VAT from the teacher model for more accurate student perturbations. Since these consistency regularization methods rely on individual perturbations, ABD [8] dynamically adjusts image displacement, improving the model's adaptability to complex data variations and robustness.

Secondly, the pseudo-labeling methods are based on the clustering assumption. It expands the dataset by generating pseudo-labels for unlabeled data using a pre-trained model. To improve the quality of pseudo-labels, ST++ [32] prioritizes reliable unlabeled images based on prediction stability and performs selective retraining. DMT [9] overcomes the limitation of self-training by identifying incorrect pseudo-labels through prediction differences between multiple models. Additionally, some methods combine pseudo-labeling with consistency regularization. For example, FixMatch [23] and RCPS [37] encourage the model to produce consistent predictions between strongly augmented data and pseudo-labels generated from weakly augmented data. While these methods enhance model performance through data diversification or improved pseudo-labels' quality, they overlook the varying complexity of predictions across different image regions, and fail to learn more discriminative features from high-value regions.

2.2. Uncertainty-Guided Semi-Supervised Medical Image Segmentation

In semi-supervised medical image segmentation, uncertainty analysis is a key strategy for effectively exploiting pseudo-labels. Existing methods typically estimate uncertainty using Monte Carlo Dropout [11]. Additionally, some approaches employ confidence scores or entropy measures for uncertainty guidance. Specifically, to enable the model to learn more reliable targets, UA-MT [34] utilizes Monte Carlo Dropout for uncertainty estimation and trains the model only with low-uncertainty pixels. URPC [16] refines uncertainty by computing the KL divergence of predictions, allowing the model to progressively learn from multi-scale reliable predictions. To further improve pseudo-label quality, UniMatch [33] evaluates the confidence of model predictions using the maximum class probability and retains only high-confidence pixels from weakly augmented data as pseudo-labels. MLRP [24] assesses the reliability of different subnet predictions via entropy and uses more reliable predictions as pseudo-labels to supervise the other subnetwork's training. To fully exploit pixel information with varying confidence levels, UFC [13] distinguishes high-performing and under-performing pixels based on confidence differences between the student and teacher models. Both DC-Net [5] and DD-Net [26] decouple predictions using confidence thresholds and optimize different pixels with distinct loss functions, effectively improving model performance. However, although these methods assess and train pixel reliability using various strategies, they rely solely on simple consistency losses (e.g. cross-entropy loss and mean squared error loss) and fail to design adaptive training strategies based on the characteristics of the data.

3. Method

3.1. Overview of the Framework

In semi-supervised image segmentation tasks, the training dataset typically consists of a small amount of labeled data and a large amount of unlabeled data. Mathematically, the training dataset is defined as $D = \{D_L, D_U\}$, where $D_L = \{(x_i^L, y_i^L)\}_{i=1}^N$ represents the labeled dataset containing N samples, and $D_U = \{x_i^U\}_{i=1}^M$ represents the unlabeled dataset containing M samples, with $N \ll M$. We define the input image as $x_i \in \mathbb{R}^{H\times W\times Z}$, where H, W and Z denote the height, width, and depth of the image, respectively. The ground truth corresponding to the labeled data is defined as $y_i \in \{0, 1, \ldots, C-1\}^{H\times W\times Z}$, where C represents the number of classes.

The training process of ALHVR is shown in Fig. 2, which consists of two parallel networks $f_A(\cdot)$ and $f_B(\cdot)$ with the same structure but independently updating parameters. Each network's decoder takes the encoded feature F_1 and its perturbed version F_2 as inputs to generate predictions from different perspectives. It is worth noting that feature perturbations are generated by randomly adding noise or dropout.

The original and perturbed features from both networks are fed into their corresponding decoders, $f_A^d(\cdot)$ and $f_B^d(\cdot)$, to generate the final prediction outputs:

$$\hat{y}_{a1} = f_A^d(F_{a1}), \hat{y}_{a2} = f_A^d(F_{a2}), \tag{1}$$

$$\hat{y}_{b1} = f_B^d(F_{b1}), \hat{y}_{b2} = f_B^d(F_{b2}), \tag{2}$$

where the decoder outputs include the prediction results for both labeled and unlabeled data, $\hat{y} = \hat{y}^L \cup \hat{y}^U$.

For labeled data, a combination of cross-entropy loss and dice loss is used as the supervised learning loss function:

$$L_s^a = L_{ce}(\hat{y}_{a1}, y) + L_{dice}(\hat{y}_{a1}, y), \tag{3}$$

$$L_{\mathrm{s}}^b = L_{ce}(\hat{y}_{b1}, y) + L_{dice}(\hat{y}_{b1}, y). \tag{4}$$

For unlabeled data, an adaptive training strategy is proposed based on high-value regions. Specifically, for reliable unstable region, the CG-CPCL module is designed to introduce a cross-prototype consistency loss. Additionally, to further enhance prototype consistency and reduce intra-class variance, the cosine similarity loss between prototypes and features is incorporated as a soft loss term to optimize intra-class feature clustering. The specific definition is as follows:

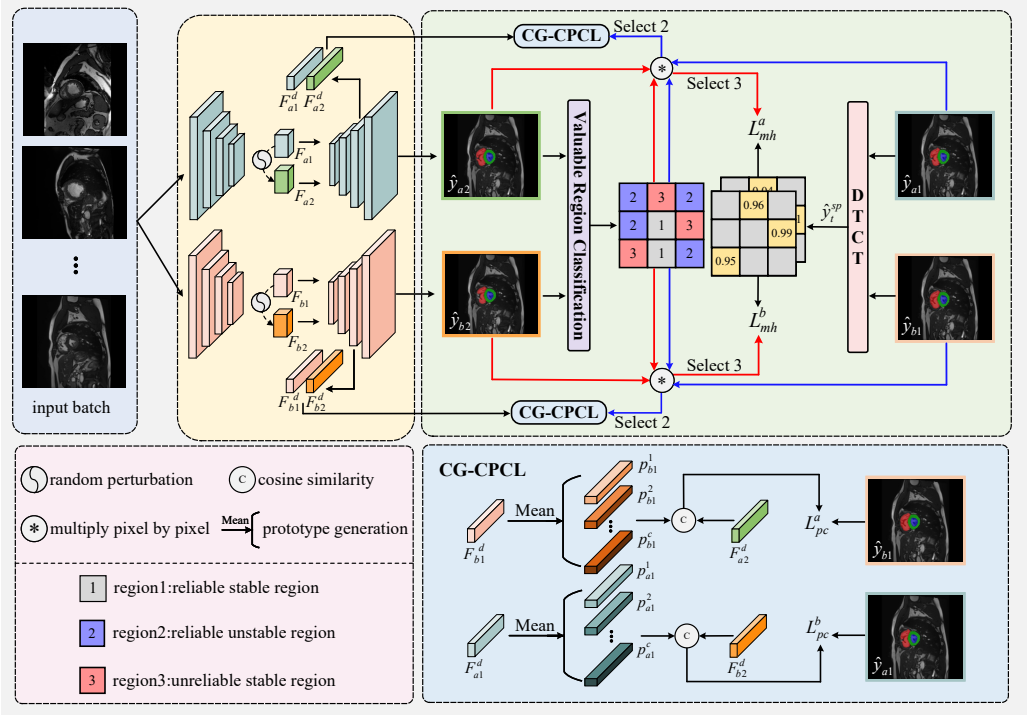

Figure 2. The process of our proposed ALHVR framework. This method first classifies the perturbation predictions of the network output into value regions, then applies an adaptive training strategy to high-value regions, where CG-CPCL is used to train Region 2 and DTCT is used to train Region 3. The supervised learning process and the pseudo-label conversion process are omitted in the figure.

$$L^a_{pc} = L^a_{ce}((F^d_{a2}, p^c_{b1}), \hat{y}^p_{b1}) + L^a_{\cos}(F^d_{a2}, p^c_{b1}), \quad (5)$$

$$L^b_{pc} = L^b_{ce}((F^d_{b2}, p^c_{a1}), \hat{y}^p_{a1}) + L^b_{\cos}(F^d_{b2}, p^c_{a1}), \quad (6)$$

where F^d_{a2} and F^d_{b2} are perturbed features from the penultimate layer of the decoder, p^c_{a1} and p^c_{b1} denotes the c-th class prototype computed from the original features of the decoder, $\hat{y}^p_{a1}$ and $\hat{y}^p_{b1}$ denote the pseudo-labels generated from $\hat{y}_{a1}$ and $\hat{y}_{b1}$, respectively. $L_{ce}(\cdot)$ represents the cross-entropy loss, and $L_{\cos}(\cdot)$ denotes the cosine similarity loss.

For unreliable stable region, the mean squared error loss is introduced through DTCT. To further constrain the model's probability distribution and reduce uncertainty in these unreliable regions, we incorporate the entropy of the model's predictions as a regularization term, thereby enhancing the model's robustness in such regions. This is formally defined as:

$$L^a_{mh} = L^a_{mse}(\hat{y}_{a2}, (\hat{y}_{a1}, \hat{y}_{b1})) + H_{a1}(\hat{y}_{a1}) + H_{a2}(\hat{y}_{a2}), \quad (7)$$

$$L^b_{mh} = L^b_{mse}(\hat{y}_{b2}, (\hat{y}_{a1}, \hat{y}_{b1})) + H_{b1}(\hat{y}_{b1}) + H_{b2}(\hat{y}_{b2}), \quad (8)$$

where $L_{mse}(\cdot)$ represents the mean squared error loss, and $H(\cdot)$ denotes the entropy.

The unsupervised loss is defined as:

$$L^a_u = L^a_{pc} + L^a_{mh}, L^b_u = L^b_{pc} + L^b_{mh}. \quad (9)$$

The total loss function for each network is:

$$L_a = L^a_s + \lambda L^a_u, L_b = L^b_s + \lambda L^b_u, \quad (10)$$

where λ is a time-dependent Gaussian warming up function [12], used to balance the supervised and unsupervised losses.

3.2. Valuable Region Classification Method

Existing semi-supervised learning methods typically use fixed thresholds to separate high-confidence and low-confidence pixels. However, this binary approach provides a coarse distinction of pixel prediction difficulty, neglecting the potential information in low-confidence pixels. To address these limitations, an adaptive threshold is employed to refine pixel classification. Specifically, we directly derive the adaptive threshold from the confidence maps produced by two networks, ensuring better adaptation to current data variations. The confidence maps are generated using perturbed predictions: $con_{a2} = max(\hat{y}_{a2}), con_{b2} = max(\hat{y}_{b2})$. By computing the pixel-wise average of the two confidence maps and sorting the values in ascending order, we obtain the sorted average confidence vector con.

Finally, the lowest K of the sorted confidence values are selected, and their mean is computed as the adaptive threshold:

$$\gamma = \frac{1}{K} \sum_{i=1}^{K} con(i), \quad (11)$$

where $K = \beta \cdot (H \times W \times Z)$, and β is a hyperparameter within the range $(0, 1)$, controlling the proportion of

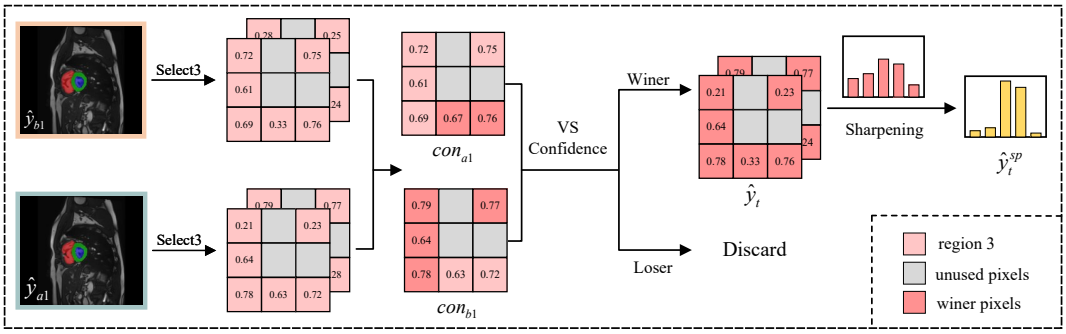

Figure 3. The structure of the DTCT module. By comparing the confidence of pixels in the Region 3 from the unperturbed predictions, the more reliable prediction is selected as the teacher. A probability sharpening function is then applied to adjust the probability distribution, generating more reliable pseudo-labels to supervise the perturbed predictions.

selected pixels.

Using the confidence threshold, we classify predicted pixels from the two networks interactively. By comparing the confidence values of corresponding pixels with the threshold, the pixels are categorized into three regions including reliable stable region Ω_1, reliable unstable region Ω_2 and the unreliable stable region Ω_3 (as described in Chapter 1). These regions can be defined as:

$$\Omega_1 = \{con_{a2} \geq \gamma \ \ \& \ \ con_{b2} \geq \gamma\}, \tag{12}$$

$$\begin{aligned}\Omega_2 = \{&(con_{a2} > \gamma \ \ \& \ \ con_{b2} < \gamma) \mid \\ &(con_{a2} < \gamma \ \ \& \ \ con_{b2} > \gamma)\},\end{aligned} \tag{13}$$

$$\Omega_3 = \{con_{a2} < \gamma \ \ \& \ \ con_{b2} < \gamma\}. \tag{14}$$

The reliable stable region is easily learnable by both networks, while the reliable unstable region reflects areas with unclear predictions, and the unreliable stable region represents difficult-to-learn areas. Thus, focusing on the reliable unstable region and unreliable stable region, rather than the reliable stable region, can better enhance the model's learning ability and provide higher learning value, as confirmed by the experimental results in Tab. 1.

3.3. Confidence-Guided Cross-Prototype Consistency Learning

Due to complex semantic information in reliable unstable region and differences in model learning capabilities, different models produce divergent predictions for the same region. To address this, we propose the CG-CPCL module, which introduces prototypes as category center references, providing a global and stable reference point. Similar to UPCoL [15], we approximate the probability distribution of voxels by calculating cosine similarity between features and prototypes, introducing global constraints to prevent over-reliance on local features. Additionally, we use confidence-based conditional guidance to ensure the model learns only from reliable targets.

Specifically, we first generate class prototypes for the unlabeled data within the current mini-batch. We upsample the features from the second-to-last layer of the decoder using trilinear interpolation to match the size of the pseudo-labels. Then, we use the pseudo-labels to select the corresponding class features, and generate class prototypes by calculating the mean of these features. This process is described as follows:

$$p^c_{a1} = \frac{\sum F^d_{a1} \cdot \mathbb{I}(\hat{y}^p_{a1} = c)}{\sum (\hat{y}^p_{a1} = c)}, \tag{15}$$

where F^d_{a1} represents the original features from the decoder, and $\mathbb{I}(\cdot)$ denotes the indicator function. For simplicity, the summation process omits the dimensional details of the features and pseudo-labels. Use the same method to calculate the p^c_{b1}. After obtaining the class prototypes, we approximate the model's prediction probability by calculating the cosine similarity between the features and the prototypes:

$$sim_a = cos(F^d_{a2}, p^c_{b1}), sim_b = cos(F^d_{b2}, p^c_{a1}), \tag{16}$$

where $cos(a, b) = \frac{a \cdot b}{\|a\|_2 \cdot \|b\|_2}$. To ensure that low-confidence predictions in the Ω_2 region only learn from reliable targets, we need to perform pixel selection on the confidence maps predicted by each network. This generates a position mask that contains the interactive information from both networks, which will be used as a guiding signal for the model's learning:

$$Mask_a = con_{a2} < \gamma \ \ \& \ \ con_{b2} > \gamma, \tag{17}$$

$$Mask_b = con_{a2} > \gamma \ \ \& \ \ con_{b2} < \gamma. \tag{18}$$

Then, the position mask can be used to calculate the cross-prototype consistency loss, and the formula is as follows:

$$L^a_{ce} = L_{ce}(sim_a, \hat{y}^p_{b1}) \odot Mask_a, \tag{19}$$

where $\odot$ represents the pixel-wise multiplication operation. Calculate L^b_{ce} in the same way. Moreover, to further enhance the model's ability to learn class features, the cosine

similarity loss between the features and prototypes is also calculated:

$$L^a_{\cos} = (1 - sim_a) \odot Mask_a. \quad (20)$$

We calculate $L^b_{\cos}$ similarly. Finally, the total loss for each model training the reliable unstable region is a linear combination of L_{ce} and $L_{\cos}$, as shown in Eqs. (5) and (6).

3.4. Dynamic Teacher Competition Teaching

The unreliable stable region is an area where both models struggle to make accurate predictions and typically exhibits high uncertainty. The module improves the adaptability of the model to complex regions by reducing the influence of noise and uncertainty in the training process. The structure of the DTCT module is shown in Fig. 3.

To generate the new teacher prediction, we perform pixel-wise confidence comparison based on the confidence maps of predictions without feature perturbation. We retain the pixels with higher confidence as the new prediction while discarding the pixels with lower confidence. Specifically, we first generate the confidence maps of the two predictions without feature perturbation: $con_{a1} = max(\hat{y}_{a1}), con_{b1} = max(\hat{y}_{b1})$. Then, we compare the generated confidence maps and select the prediction probabilities corresponding to the pixels with higher confidence as the new teacher prediction. Next, we apply a probability sharpening function [29] to convert this into a soft pseudo-label. This process can be defined as:

$$\hat{y}_t = \begin{cases} \hat{y}_{a1}, con_{a1} \geq con_{b1} \\ \hat{y}_{b1}, con_{a1} < con_{b1} \end{cases}, \quad (21)$$

$$\hat{y}^{\text{sp}}_t = \frac{\hat{y}_t^{1/T}}{\hat{y}_t^{1/T} + (1 - \hat{y}_t)^{1/T}}, \quad (22)$$

where T is a hyperparameter that controls the output distribution of the model. Since the DTCT module is only used to train the pixels in the unreliable stable region, a location mask needs to be generated first to determine the region location of Ω_3:

$$Mask = con_{a2} < \gamma \ \ \& \ \ con_{b2} < \gamma. \quad (23)$$

Then, the newly generated reliable pseudo-labels are used to supervise the perturbed version of the predictions, as shown in the following formula:

$$L^a_{mse} = L_{mse}(\hat{y}_{a2}, \hat{y}^{sp}_t) \odot Mask. \quad (24)$$

We can calculate L^b_{mse} in the same way. To enhance the model's ability to perceive uncertainty in the unreliable stable region, guide the model to focus on regions with higher uncertainty, and effectively reduce the uncertainty in these regions, we quantify uncertainty using entropy and add it as a regularization term to the loss function:

$$H_{a1} = H(\hat{y}_{a1}) \odot Mask, H_{a2} = H(\hat{y}_{a2}) \odot Mask, \quad (25)$$

Method	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
U-Net	7(10%)	0	81.59	70.74	8.07	2.35
U-Net	14(20%)	0	84.83	75.33	7.61	2.09
U-Net	70(All)	0	91.10	84.04	4.90	1.14
MT[25][NIPS'2017]	7(10%)	63(90%)	83.65	73.15	13.45	3.61
CPS[6][CVPR'2021]			85.34	75.50	8.78	2.37
MC-Net[28][MICCAI'2021]			86.07	76.58	11.48	3.37
MCF[27][CVPR'2023]			85.18	75.22	10.73	2.78
AC-MT[31][MIA'2023]			86.36	76.86	9.64	2.59
UG-MCL[36][AIIM'2023]			86.84	77.60	7.77	1.94
EVIL[7][CBM'2024]			87.25	78.18	5.31	1.35
AAU[1][MIA'2024]			86.46	76.99	8.00	2.06
MLRP[24][MIA'2024]			87.10	78.16	4.91	1.30
ours			**90.56**↑3.31	**83.21**	**2.57**	**0.78**
MT[25][NIPS'2017]	14(20%)	56(80%)	86.55	77.49	6.86	2.15
CPS[6][CVPR'2021]			87.03	78.13	6.66	2.12
MC-Net[28][MICCAI'2021]			86.58	77.68	15.99	4.93
MCF[27][CVPR'2023]			87.32	78.43	6.53	2.00
AC-MT[31][MIA'2023]			87.99	79.34	7.85	2.14
UG-MCL[36][AIIM'2023]			88.40	79.86	9.12	2.44
EVIL[7][CBM'2024]			88.34	79.90	6.46	1.68
AAU[1][MIA'2024]			87.71	78.92	8.51	2.14
MLRP[24][MIA'2024]			88.00	79.54	5.41	1.58
ours			**91.09**↑2.69	**84.05**	**2.71**	**0.73**

Table 2. Comparisons with state-of-the-art semi-supervised segmentation methods on the ACDC dataset.

where $H(\hat{y}) = -\hat{y} \log \hat{y}$. We can use the same method to calculate the H_{b1} and H_{b2}. Finally, the total loss for training the unreliable stable region in each model is a linear combination of L_{mse} and H, as shown in Eqs. (7) and (8).

4. Experiments

4.1. Datasets and Evaluation Metrics

The **ACDC dataset** [4] consists of cardiac dynamic MRI scans and corresponding ground truth from 100 patients, categorized into three classes: right ventricle, left ventricle, and myocardium. Following the experimental setup in CTCT [17], we used scans from 70 patients for training, 10 patients for validation, and 20 patients for testing.

The **AbdomenCT-1K dataset** [18] contains over 1,000 (1K) CT scans from 12 medical centers, categorized into four classes: liver, kidneys, spleen, and pancreas. Following the experimental setup in BFFC [38], we selected 30 scans for training and 11 scans for testing.

The **Brats dataset** [3] consists of preoperative MRI scans from 335 glioma patients. Following the experimental setup in CPCL [30], we used 250 samples for training, 25 samples for validation, and 60 for testing.

Metrics. We evaluate the method using four metrics categorized into region-sensitive and edge-sensitive types. Region-sensitive metrics include Dice and Jaccard, while edge-sensitive metrics include 95HD and ASD.

4.2. Implementation Details

All experiments were implemented using the PyTorch 1.7 on NVIDIA GeForce RTX 3090 GPU. The model is trained using the SGD optimizer with a weight decay factor of 0.0001, a momentum of 0.9 and an initial learning rate of 0.01. For segmentation on the ACDC and AbdomenCT-1K datasets, we use U-Net [22] as the backbone, with a

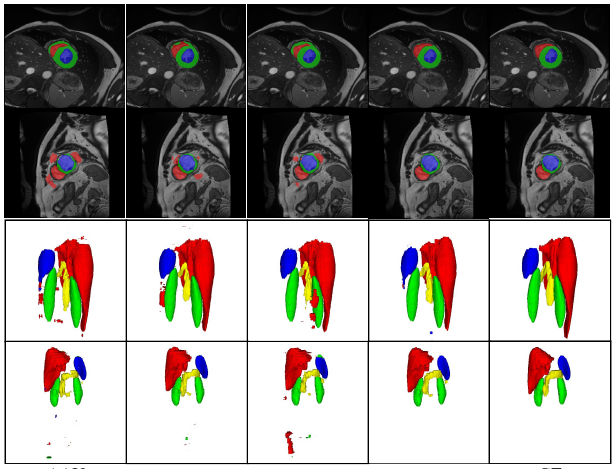

Figure 4. Visualization of some semi-supervised methods. The first two rows show the segmentation results on the ACDC dataset trained with 10% labeled data. The last two rows show the segmentation results on the AbdomenCT-1K dataset trained with 5% labeled data.

batch size of 16 and 30k training iterations. For Brats, we use V-Net [19] as the backbone, with a batch size of 4 and 15k training iterations. In the total loss function of the network, the Gaussian warming up function is defined as $\lambda(t) = 0.1 \times e^{-5(1-t/t_{\max})^2}$, where t denotes the current number of iterations and $t_{\max}$ represents the maximum number of iterations. In the probability sharpening function, the hyperparameter T is usually empirically set to 0.1. For calculating the confidence threshold, the hyperparameter β values were configured as 0.8, 0.7, and 0.7 in the ACDC, AbdomenCT-1K, and Brats datasets, respectively (set by Ablation Studies).

4.3. Comparison with State-of-the-Art Methods

Firstly, we evaluate the proposed method on the ACDC dataset, comparing it with MT [25], CPS [6], MC-Net [28], MCF [27], AC-MT [31], UG-MCL [36], EVIL [7], AAU [1], and MLRP [24]. Additionally, we use U-Net trained with 10%, 20%, and 100% of labeled data as the baseline and upper bound for reference. As shown in Tab. 2, our method achieves significant improvements across all four evaluation metrics compared to the baseline. Specifically, with 10% labeled data, the Dice score increases from 81.59% to 90.56%. With 20% labeled data, it improves from 84.83% to 91.09%, approaching the upper bound of 91.10% achieved with full supervision. Compared to state-of-the-art methods, ALHVR consistently outperforms others. With 10% labeled data, it surpasses the second-best method (EVIL) by 3.31% in Dice. When using 20% labeled data, it outperforms UG-MCL by 2.69% in Dice.

To further validate the generalization capability of the proposed method across different datasets, we evaluate its performance on the five-class AbdomenCT-1K dataset using the same comparison methods as in the ACDC experiment. As shown in Tab. 3, with 5% labeled data, ALHVR improves the Dice score by 17.36% compared to the baseline. When increasing the labeled data to 20%, the Dice score improves by 10.02%. Compared to state-of-the-art methods, ALHVR consistently achieves the best performance in segmenting four organs. Specifically, with 5% and 10% labeled data, it surpasses the second-best methods (MC-Net and MLRP) by 3.78% and 1.13% in Dice, respectively. Furthermore, as illustrated in Fig. 4 , our method provides more precise segmentation of fine structures and boundary regions across different organs.

Method	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
U-Net	1(5%)	0	67.39	57.14	46.87	11.37
U-Net	3(10%)	0	75.66	66.33	22.44	6.89
U-Net	30(All)	0	86.93	79.27	11.55	2.45
MT[25][NIPS'2017]	1(5%)	29(95%)	71.82	61.19	42.03	9.23
CPS[6][CVPR'2021]			78.14	68.32	26.25	6.68
MC-Net[28][MICCAI'2021]			80.97	71.43	32.98	7.48
MCF[27][CVPR'2023]			80.79	71.30	26.13	7.81
AC-MT[31][MIA'2023]			76.18	65.99	36.12	8.89
UG-MCL[36][AIIM'2023]			74.06	63.58	26.93	7.06
EVIL[7][CBM'2024]			80.68	71.47	23.72	6.33
AAU[1][MIA'2024]			80.65	71.03	32.68	9.04
MLRP[24][MIA'2024]			80.30	70.67	20.94	5.43
ours			**84.75**↑3.78	**76.35**	21.15	5.59
MT[25][NIPS'2017]	3(10%)	27(90%)	78.75	69.10	25.35	6.88
CPS[6][CVPR'2021]			82.71	74.11	17.30	4.90
MC-Net[28][MICCAI'2021]			82.94	74.13	19.35	4.59
MCF[27][CVPR'2023]			83.34	74.63	16.97	4.72
AC-MT[31][MIA'2023]			82.32	72.97	26.52	7.54
UG-MCL[36][AIIM'2023]			81.43	72.56	20.29	5.67
EVIL[7][CBM'2024]			84.24	75.75	18.14	5.25
AAU[1][MIA'2024]			81.42	71.82	29.69	8.25
MLRP[24][MIA'2024]			84.55	75.99	13.82	4.05
ours			**85.68**↑1.13	**77.35**	21.96	6.22

Table 3. Comparisons with state-of-the-art semi-supervised segmentation methods on the AbdomenCT-1K dataset.

In addition to organ segmentation tasks, we evaluated our method on the Brats dataset using the same comparison methods as in the ACDC experiment. As shown in Tab. 4, with only 10% labeled data, ALHVR improves the Dice score by 5.91% compared to the baseline. Furthermore, it outperforms state-of-the-art methods, achieving a 1.18% higher Dice score than the second-best method (MLRP).

4.4. Ablation Studies

In this section, we conduct ablation studies to validate the effectiveness of each part and analyze the hyperparameters of the proposed method. The experiments were conducted on the ACDC (labeled ratio:10%), AbdomenCT-1K (labeled ratio:5%), and Brats (labeled ratio:10%) datasets.

4.4.1. Impact of Different Modules

In Tab. 5, we examine the effectiveness of the major components of ALHVR, CG-CPCL, and DTCT, on different datasets. When CG-CPCL training (learning only reliable unstable region) was used, the performance of all datasets was significantly improved compared with the baseline

Method	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
V-Net	25(10%)	0	79.23	68.40	15.83	4.34
V-Net	250(All)	0	86.49	77.23	8.68	1.98
MT[25][NIPS'2017]	25(10%)	225(90%)	80.49	69.99	11.71	3.46
CPS[6][CVPR'2021]			82.70	72.30	12.99	3.46
MC-Net[28][MICCAI'2021]			82.51	72.98	14.57	3.78
MCF[27][CVPR'2023]			81.96	71.76	13.96	3.99
AC-MT[31][MIA'2023]			82.09	71.65	15.91	4.81
UG-MCL[36][AIIM'2023]			82.85	72.69	11.12	2.28
EVIL[7][CBM'2024]			82.60	72.46	11.44	2.76
AAU[1][MIA'2024]			82.43	72.33	13.83	4.18
MLRP[24][MIA'2024]			83.96	74.00	15.16	4.21
ours			**85.14**↑1.18	**75.75**	**7.79**	**2.00**

Table 4. Comparisons with state-of-the-art semi-supervised segmentation methods on the Brats dataset.

Datasets	CG-CPCL	DTCT	Dice↑	Jaccard↑	95HD↓	ASD↓
ACDC			86.61	77.14	7.30	2.15
	✓		89.32	81.31	3.45	1.02
		✓	89.19	80.97	3.26	0.94
	✓	✓	**90.56**	**83.21**	**2.57**	**0.78**
AbdomenCT-1K			79.19	69.21	29.77	7.56
	✓		83.98	75.91	9.59	2.81
		✓	83.79	75.25	21.39	5.28
	✓	✓	**84.75**	**76.35**	21.15	5.59
Brats			83.20	73.73	11.91	3.01
	✓		84.62	74.81	8.92	2.36
		✓	84.85	75.22	9.96	2.61
	✓	✓	**85.14**	**75.75**	**7.79**	**2.00**

Table 5. Ablation results of different components of ALHVR. We conducted experiments to test the impact of CG-CPCL and DTCT on model performance.

method, especially on the AbdomenCT-1K dataset, where the Dice index improved by 4.79%. In the ACDC dataset, when cross pseudo supervision similar to CPS was applied (see Tab. 1 for experimental results), CG-CPCL increased by 0.86%. The reliable unstable region usually has feature differences within the class, which is prone to misjudgment. Therefore, the class prototype is constructed, and cross-prototype consistency is used to enhance the model's discrimination. In addition, the confidence-based conditional guidance mechanism effectively reduces the performance difference between models.

Similarly, with DTCT training (learning only the unreliable stable region), the performance of all datasets improved significantly, with the Brats dataset improving by 1.65%. On the ACDC dataset, CG-CPCL improved by 0.99% when cross pseudo supervision similar to CPS was used. The unreliable stable region has high uncertainty, and a single model is prone to bias. To this end, we improved the accuracy of the model in this area by integrating predictions from multiple models, competitively generating teacher predictions, and selecting the most credible predictions as pseudo-labels.

When CG-CPCL and DTCT are combined, the model can achieve optimal segmentation performance by using differentiated learning strategies to optimize both reliable unstable and unreliable stable regions. Compared with the baseline model, the Dice indexes of multiple datasets improved by 3.95%, 5.56%, and 1.94%, respectively. On the ACDC dataset, ALHVR increased the Dice indicator from

β	0.5	0.6	0.7	0.8	0.9
ACDC	89.88	89.94	90.26	**90.56**	89.85
AbdomenCT-1K	84.16	84.39	**84.75**	84.49	83.71
Brats	84.71	85.02	**85.14**	84.84	84.50

Table 6. Ablation studies on the hyperparameter β with 10% , 5% and 10% labeled data.

88.58% to 90.56%. Since these two types of regions have different distributions and prediction characteristics, a unified training strategy is difficult to account for both. Therefore, our method optimizes the two types of regions with targeted strategies, effectively improving the overall performance of the model and its generalization ability under complex scenarios.

4.4.2. Hyperparameter Value Analysis

We investigate the impact of confidence thresholds calculated based on different pixel proportions on model training. As shown in Tab. 6, the performance of dynamic thresholds on multiple datasets is evaluated using Dice coefficients. The results indicate that the optimal proportions are 0.8 for the ACDC dataset and 0.7 for the AbdomenCT-1K and Brats datasets. When the proportion is set lower, the threshold decreases, causing the model to focus more on high-confidence pixels, which may reduce its ability to handle difficult samples and degrade the quality of pseudo-labels. Conversely, when the threshold increases, the model may overlook subtle features, affecting detail prediction. Therefore, it is necessary to balance the pixel proportion to optimize the model's attention on the reliable unstable region and the unreliable stable region, thereby improving performance and robustness.

5. Conclusion

In this paper, we propose a novel semi-supervised learning framework, ALHVR, which explores high-value learning regions and designs adaptive training strategies. Specifically, CG-CPCL is used to train the reliable unstable region, while DTCT is employed for the unreliable stable region. To enhance stability in the reliable unstable region, CG-CPCL enables low-confidence predictions to learn from high-confidence predictions through confidence-guided cross-pseudo supervision and prototype consistency constraints, improving intra-class consistency and inter-class discriminability. To enhance reliability in the unreliable stable region, DTCT dynamically selects more reliable pixels as teachers using a dynamic teacher selection mechanism and competitive teaching strategy, reducing noise interference while strengthening the model's learning capability. Experimental results on three public datasets demonstrate that ALHVR achieves better performance than existing state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62271296, 62201334), the Shaanxi Province Innovation Capacity Support Program (Grant No. 2025RS-CXTD-012), and the Scientific Research Program funded by the Shaanxi Provincial Education Commission (Grant Nos. 23JP022, 23JP014).

References

[1] Sukesh Adiga, Jose Dolz, and Herve Lombaert. Anatomically-aware uncertainty for semi-supervised image segmentation. *Medical Image Analysis*, 91:103011, 2024. 6, 7, 8

[2] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023. 2

[3] Spyridon (Spyros) Bakas. Brats miccai brain tumor dataset, 2020. https://dx.doi.org/10.21227/hdtd-5j88. 6

[4] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37 (11):2514–2525, 2018. 6

[5] Faquan Chen, Jingjing Fei, Yaqi Chen, and Chenxi Huang. Decoupled consistency for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 551–561. Springer, 2023. 3

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 2, 6, 7, 8

[7] Yingyu Chen, Ziyuan Yang, Chenyu Shen, Zhiwen Wang, Zhongzhou Zhang, Yang Qin, Xin Wei, Jingfeng Lu, Yan Liu, and Yi Zhang. Evidence-based uncertainty-aware semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 170:108004, 2024. 6, 7, 8

[8] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4070–4080, 2024. 3

[9] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130:108777, 2022. 3

[10] Hejun Huang, Zuguo Chen, Chaoyang Chen, Ming Lu, and Ying Zou. Complementary consistency semi-supervised learning for 3d left atrial image segmentation. *Computers in Biology and Medicine*, 165:107368, 2023. 2

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 3

[12] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 4

[13] Tao Lei, Yi Wang, Xingwu Wang, Xuan Wang, Bin Hu, and Asoke K Nandi. Unified feature consistency of underperforming pixels and valid regions for semi-supervised medical image segmentation. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2024. 3

[14] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4258–4267, 2022. 3

[15] Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu, and Yang Yang. Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 662–672. Springer, 2023. 5

[16] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting 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*, pages 318–329. Springer, 2021. 3

[17] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International conference on medical imaging with deep learning*, pages 820–833. PMLR, 2022. 6

[18] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021. 6

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 7

[20] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1369–1378, 2021. 2

[21] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 3

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 6

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 3

[24] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94:103111, 2024. 1, 3, 6, 7, 8

[25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 6, 7, 8

[26] Bing Wang, Taifeng Huang, Shuo Yang, Ying Yang, Junhai Zhai, and Xin Zhang. Dual-decoder data decoupling training for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 100:106984, 2025. 3

[27] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15651–15660, 2023. 1, 2, 6, 7, 8

[28] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part II 24*, pages 297–306. Springer, 2021. 6, 7, 8

[29] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 6

[30] Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3174–3184, 2022. 6

[31] Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai-yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, 88:102880, 2023. 6, 7, 8

[32] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022. 3

[33] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7236–7246, 2023. 3

[34] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann 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*, pages 605–613. Springer, 2019. 1, 3

[35] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[36] Yichi Zhang, Rushi Jiao, Qingcheng Liao, Dongyang Li, and Jicong Zhang. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. *Artificial Intelligence in Medicine*, 138:102476, 2023. 6, 7, 8

[37] Xiangyu Zhao, Zengxin Qi, Sheng Wang, Qian Wang, Xuehai Wu, Ying Mao, and Lichi Zhang. Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):251–261, 2023. 3

[38] Zhongda Zhao, Haiyan Wang, Tao Lei, Xuan Wang, Xiaohong Shen, and Haiyang Yao. Balanced feature fusion collaborative training for semi-supervised medical image segmentation. *Pattern Recognition*, 157:110856, 2025. 6