MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation

Yongchao Wang Bin Xiao* Xiuli Bi Weisheng Li Xinbo Gao
Chongqing University of Posts and Telecommunications
Chongqing, China
d200201023@stu.cqupt.edu.cn, {xiaobin,bixl,liws,gaoxb}@cqupt.edu.cn

Abstract

Semi-supervised learning is a promising method for medical image segmentation under limited annotation. However, the model cognitive bias impairs the segmentation performance, especially for edge regions. Furthermore, current mainstream semi-supervised medical image segmentation (SSMIS) methods lack designs to handle model bias. The neural network has a strong learning ability, but the cognitive bias will gradually deepen during the training, and it is difficult to correct itself. We propose a novel mutual correction framework (MCF) to explore network bias correction and improve the performance of SSMIS. Inspired by the plain contrast idea, MCF introduces two different subnets to explore and utilize the discrepancies between subnets to correct cognitive bias of the model. More concretely, a contrastive difference review (CDR) module is proposed to find out inconsistent prediction regions and perform a review training. Additionally, a dynamic competitive pseudo-label generation (DCPLG) module is proposed to evaluate the performance of subnets in real-time, dynamically selecting more reliable pseudo-labels. Experimental results on two medical image databases with different modalities (CT and MRI) show that our method achieves superior performance compared to several state-of-the-art methods. The code will be available at https://github.com/WYC-321/MCF.

1. Introduction

Making pixel-level annotation is difficult and time-consuming, especially for medical images. Semi-supervised learning is a promising approach for processing images with limited supervised data [2, 3, 14, 17, 23, 30, 31]. In recent years, semi-supervised methods based on consistency regularization [21, 27] have attracted the attention of researchers and are one of the mainstream techniques, especially in SSMIS. These methods usually include two subnets, and the general process is to add random perturbations to the same samples and force the subnets to produce consistent prediction results for these perturbed inputs. Mean Teacher (MT) [22] is a typical method among them and inspired a series of SSMIS work [13, 33]. Although these methods have achieved promising results, they ignore the impact of cognitive biases in the model. Cognitive bias is the phenomenon in which the model persists in mispredictions, caused by overfitting to wrong supervision signals [12]. There is evidence that cognitive biases reduce the performance of consistency-regularization-based methods [1].

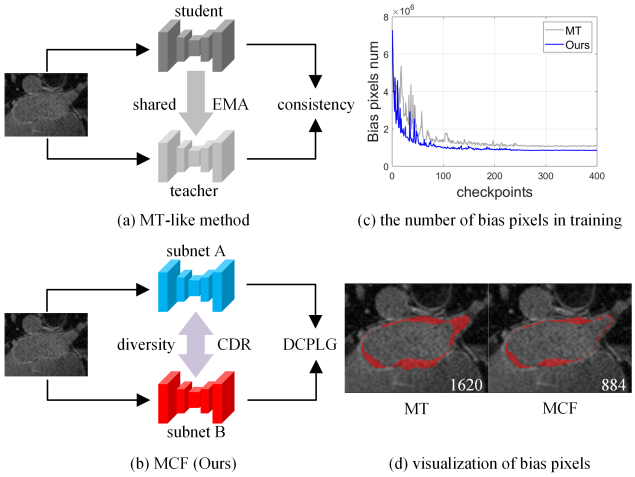


Figure 1. A brief description of the process of MT-like methods (a) and our proposed MCF (b). (c) Biased prediction of MT and MCF during training, MCF reduces more biased predictions. (d) The red masks represent the model's wrong predictions, and the white numbers represent the number of wrong pixels.

We take MT-like methods as an example to analyze this issue. As shown Fig. 1 (a) the methods based on the MT framework have three features: **(1)** The model consists

*Corresponding author

of teacher and student networks with a shared structure. **(2)** The student network parameters θ are updated through stochastic gradient descent, while the teacher network parameters θ' are updated from the student network using Exponential Moving Average (EMA) as Eq. (1):

$$\theta' = \alpha\theta' + (1-\alpha)\theta \tag{1}$$

where α is the EMA decay that controls the updating rate. **(3)** Consistency regularization is implemented to encourage subnets to produce consistent predictions. Characteristics **(1)** and **(2)** naturally give the model a tendency to output consistent predictions. And explicit consistency constraints provide a supervised signal for unlabeled data. Therefore, the above three features make model training simpler and accelerate model convergence.

However, there are also three limitations hidden in it: **(1)** Structural sharing among subnets reduces model variability. **(2)** Due to the parameter update method of EMA, the teacher network is a weighted mixture of the historical states of the student network. Therefore, the performance of the teacher network is constrained by the student network. **(3)** Consistency regularization can also be regarded as a labeling strategy that subnets to generate pseudo-labels for each other. The quality of the pseudo-labels greatly affects the performance of the model. Combining limitations **(1)** and **(2)**, since the pseudo-labels come from a mixture of historical states with the same architecture as the student network, the consistency-based pseudo-label generation methods are more prone to trap the network in cognitive biases and difficult to correct for mispredictions. In addition, these limitations make the model waste the potential of the multi-subnet architecture.

To further demonstrate the cognitive bias of the model, we test MT and our proposed MCF every 15 iterations and record the number of erroneous pixels, as shown in Fig. 1 (c). It can be seen that as the training progresses, although the number of bias/wrong pixels continues to decrease, the model overfits some bias predictions that are difficult to correct on its own. The visualization result is shown in Fig. 1 (d). The red mask shows the model's bias predictions, and the white number shows the number of bias pixels. These bias pixels are mainly located in the target edge region, thus reducing the biased prediction is beneficial to improve the accuracy of edge segmentation. Finally compared with MT, MCF reduces more biased predictions under the same training steps.

In summary, our goal is to find a mechanism for the network to be aware of cognitive biases and correct them. To this end, we propose a mutual correction framework in the semi-supervised medical image segmentation for the exploration of model bias correction. We think that while the network is highly capable of learning, it is difficult to correct biases on its own. Inspired by the idea of contrast, MCF consists of two distinct structural subnets with independent parameter updates, which learn to correct each other through a strong inter-subnet interaction. Specifically, MCF proposes contrastive difference review (CDR) and dynamic competitive pseudo-label generation (DCPLG) for labeled and unlabeled data training, respectively. The CDR takes prediction discrepancies of subnets as potential bias areas and guiding the subnets to correct them. Furthermore, we observe that one of the differences between the medical image segmentation databases and natural image databases is that all medical images are related to the target object. Therefore, it is reasonable to evaluate the performance of the subnets on a small amount of labeled data. Based on this, unlike MT-like methods [14, 23], MCF does not bind teacher or student roles to fixed subnets, but instead proposes DCPLG to dynamically evaluate and select pseudo-label generation networks for more reliable label propagation. The main contributions of this work are as follows:

- We explore the problem of model bias correction and propose a new framework MCF for semi-supervised medical image segmentation.
- A CDR module is proposed to guide the network to pay attention and correct its own potential bias.
- Combined with the characteristics of medical image segmentation databases, DCPLG is proposed to obtain more reliable pseudo-labels.

We evaluate the proposed MCF framework on semi-supervised medical image segmentation with both CT and MRI modalities. Experiments verify the effectiveness of this framework, showing that MCF outperforms the SOTA method, especially in edge segmentation accuracy.

2. Related Work

Pseudo-labeling method. Generating pseudo-labels for unlabeled data is a classic practice in semi-supervised learning. The key point of the pseudo-labeling method is how to generate reliable pseudo-labels. [8] is an early exploration of semi-supervised learning using pseudo-label. This work directly uses a fixed threshold to select high-confidence unlabeled samples for pseudo-labeling. Following Tri-training [35], Tri-Net [6] is proposed to utilize two subnets to generate pseudo-labels for the third one. In order to improve the quality of pseudo-labels, uncertainty estimation is used in [17] to select more reliable pseudo-labels to improve model performance. Inspired by extreme value theory, the authors of [3] propose to use increasing percentage scores to select pseudo-label samples by imitating curriculum learning. These methods lack designs to address bias.

Consistency regularization. In recent years, consistency regularization has become a popular method in semi-supervised learning. A representative method, MT [22],

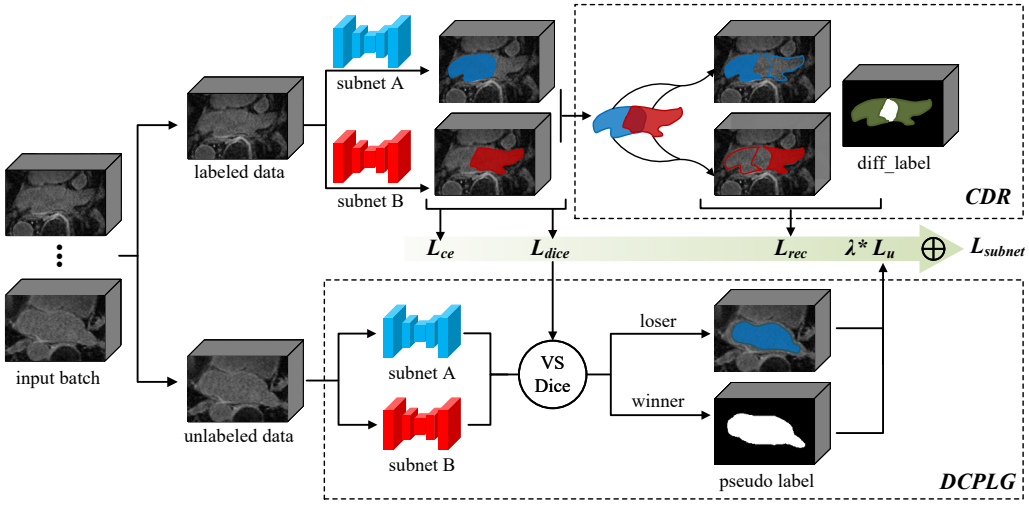


Figure 2. The process of our proposed MCF framework. The loss of the subnet L_{subnet} includes supervised loss (L_{dice}, L_{ce}, L_{rec}) on labeled dataset and unsupervised loss L_u on unlabeled dataset.

consists of a student network that updates parameters using gradient propagation and a teacher network that updates parameters using EMA. MT performs label propagation by forcing consistent predictions for the perturbed samples. After that, some work inspired by MT appeared. For example, the authors of [14] argues that previous consistency methods only consider the perturbations around each data point, while ignoring the connections between data points. Therefore, a graph-based SNTG method is proposed to encourage adjacent points on the teacher graph to maintain consistency against perturbations. In addition, [16] introduces adversarial perturbations into consistency learning, however, Verma *et al.* finds that adversarial perturbations might impair generalization performance, so an interpolation-consistent training method ICT [23] is proposed by them to avoid this problem. [24] introduces multi-task learning into the MT framework and develops triple uncertainty to guide the student model to learn more reliable predictions from the teacher model. [7] proposes a two-stage semi-supervised learning method for neuron segmentation by exploiting the pixel-level prediction consistency between unlabeled samples and their perturbed counterparts. These methods ignore the interactions between subnets and also cannot correct the biases of the network itself.

Semi-supervised medical image segmentation. The difficulty of labeled medical image data collection motivates the development of semi-supervised medical image segmentation research. Among these methods, [4, 19, 26, 32] mainly focuses on pseudo-labeling, while [5, 11, 13, 20, 28] explores the application of consistency regularization. In particular, [5, 11, 20, 28] strives to make the model invariant to samples with different perturbations, while [13] explores consistency between different tasks. In addition, Li *et al.* propose a self-ensembling semi-supervised co-training framework for COVID-19 CT segmentation [9].

Correction learning. There are several studies exploring error correction related to our method. [34] utilizes the multi-task complementary information between inpainting and segmentation to gradually optimize the segmentation results. Some researchers introduce an additional network in the segmentation model to learn the difference between prediction and GT. Specifically, [25] utilizes a complementary correction network to map the output of a base network to GT and guide the training of another base network. [15] uses the correction network to judge the matching degree between the segmentations and images, and generates supervised signals for the unlabeled data. These methods require complex loss functions and transformation rules. Different from the above methods, we propose a new framework that does not require additional correction network or complex optimization objectives, which utilizes the inter-network interactions to correct network biases directly.

3. Mutual Correction Framework

3.1. The Overall process of the MCF

The training process of MCF is shown in Fig. 2. As mentioned above, MCF consists of two subnets with comparable performance and different structures in our work, denoted by subnet A ($f_A(\cdot)$) and subnet B ($f_B(\cdot)$). In semi-supervised scenario, the training data contains a small amount of labeled data, denoted by $D_L = \{(x_i^L, y_i^L)\}_{i=1}^{N}$, and a large amount of unlabeled data, denoted by $D_u = \{x_i^U\}_{i=N+1}^{N+M}$, where $N \ll M$. $x_i \in \mathbb{R}^{H\times W\times D}$ is the medical volume and $y_i \in \{0,1\}^{H\times W\times D}$ is the ground-truth. A

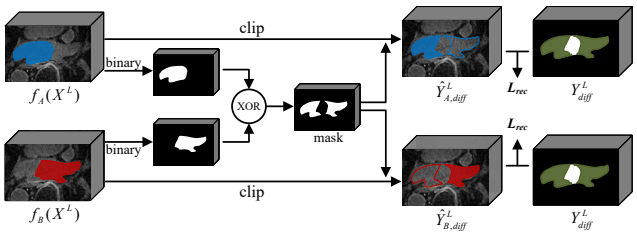


Figure 3. CDR is designed to obtain potential discrepant regions from the discrepancy mask and guide the network to review these regions.

batch of input data X contains equal labeled data (X^L, Y^L) and unlabeled data X^U, and these volumes are sent to subnet A and B :

$$\hat{Y}_A = f_A(X) \tag{2}$$

$$\hat{Y}_B = f_B(X) \tag{3}$$

The outputs include labeled and unlabeled volumes predictions: $\hat{Y} = \hat{Y}^L \cup \hat{Y}^U$. Note that the subnet index subscripts are omitted here for brevity. The loss function of the subnets L_{subnet} includes supervised loss and unsupervised loss. Specifically, for labeled data prediction $\hat{Y}^L$, excepting the conventional segmentation loss (i.e., L_{dice} and L_{ce}), a rectification loss L_{rec} is also introduced through CDR for potential mispredictions correction. For unlabeled data prediction $\hat{Y}^U$, DCPLG is used to dynamically generate pseudo-labels to supervise the non-pseudo-label generation subnet by unsupervised loss L_u. The heterogeneous subnet structure and independent parameter update introduce more diversity to the MCF, fully unlocking the potential of the multi-subnet architecture.

3.2. Contrastive discrepancy review

The process of CDR is shown in Fig. 3. CDR is inspired by a simple truth: in a binary classification scenario, if two classifiers have different predictions for the same voxel, then one of them must be wrong. Therefore, we treat the inconsistent predictions as possible areas of misprediction and let the network review these areas. The misprediction area masks can be obtained by the XOR with binarization between the softmax output $\hat{Y}^U$ of the two subnets, which is formalized as follows:

$$M_{\text{diff}} = \text{BINA}(\hat{Y}_A^L) \oplus \text{BINA}(\hat{Y}_B^L) \tag{4}$$

And then the masks can be used to find potential wrong prediction regions:

$$\hat{Y}_{diff}^L = \text{CLIP}(M_{\text{diff}}, \hat{Y}^L) \tag{5}$$

Here, $\text{CLIP}(\cdot)$ represents an operation to obtain the predictions corresponding to the masked areas.

Algorithm 1 DCPLG

Require:
The set of labelled volumes for a batch: $\{X^L, Y^L\}$
The set of unlabelled samples for a batch: $\{X^U\}$
The subnet A: $f_A(\cdot)$
The subnet B: $f_B(\cdot)$
The Dice loss function: $L_{Dice}(\cdot)$
The batch size: bs

Ensure:
The pseudo labels of unlabeled samples for a batch, Y_p^U

$X \leftarrow \{X^L, X^U\}$
$T \leftarrow 0.1$
$dice_A \leftarrow 0$; $dice_B \leftarrow 0$
$\hat{Y}_A^L, \hat{Y}_A^U \leftarrow \text{softmax}(f_A(X))$
$\hat{Y}_B^L, \hat{Y}_B^U \leftarrow \text{softmax}(f_B(X))$
for each $y_i^L \in Y^L$ and $\hat{y}_{A,i}^L \in \hat{Y}_A^L$ and $\hat{y}_{B,i}^L \in \hat{Y}_B^L$ **do**
  $dice_A \leftarrow dice_A + \frac{2}{bs} L_{Dice}(\hat{y}_{A,i}^L, y_i^L)$
  $dice_B \leftarrow dice_B + \frac{2}{bs} L_{Dice}(\hat{y}_{B,i}^L, y_i^L)$
end for
if $dice_A < dice_B$ **then**
  $P \leftarrow \hat{Y}_A^U$
  $Y_p^U \leftarrow \frac{P^{1/T}}{P^{1/T}+(1-P)^{1/T}}$ # sharpening function
else
  $P \leftarrow \hat{Y}_B^U$
  $Y_p^U \leftarrow \frac{P^{1/T}}{P^{1/T}+(1-P)^{1/T}}$ # sharpening function
end if
return Y_p^U

We design a rectification loss to guide the model to review these potentially mispredicted areas, the rectification loss is as follows:

$$L_{rec} = MSE(\hat{Y}_{diff}^L, Y_{diff}^L) \tag{6}$$

Here $MSE(\cdot)$ represents the Mean Squared Error (MSE) loss function, Y_{diff}^L is the ground truth corresponding to these areas.

3.3. Dynamic competitive pseudo label generation

Algorithm 1 describes the pseudo code of our DCPLG module. We subtly use Dice loss to measure segmentation performance of subnets in real-time and select a better performing subnet as a pseudo-label generator for another subnet. Following the entropy minimization, we utilize the sharpening function [27] to convert the network predictions into soft pseudo-labels. We use the dice loss as the evaluation criterion because the loss can directly reflect the dice coefficient and does not introduce additional computation, it is computed on labeled data. Therefore, DCPLG is almost free lunch that does not introduce additional network structures, and is easy to integrate into other semi-supervised

methods. DCPLG can adopt other metrics for subnet performance evaluation, such as Hausdorff distance, average surface distance, etc.

3.4. Overall loss function and model details

The overall optimization loss function of one subnet can be formalized as follows:

$$L_{subnet} = L_s + \lambda L_u \mathbb{I}(subnet \neq label\_generator) \quad (7)$$

Where L_s represents supervised loss, L_u represents unsupervised loss and $\mathbb{I}(\cdot)$ is an indicator of whether the subnet is a pseudo-label generator. λ is a weight that balances supervised and unsupervised losses.

Supervised loss includes common segmentation loss and rectification loss L_{rec} derived from CDR. It can be expressed as follows:

$$\begin{aligned} L_s = &Dice(\hat{Y}^L, Y^L) + CE(\hat{Y}^L, Y^L) \\ &+ \beta L_{rec}(\hat{Y}^L_{diff}, Y^L_{diff}) \end{aligned} \quad (8)$$

Here, hyperparameters β is used to balance rectification loss and other losses.

The unsupervised loss can be formulated as follows:

$$L_u = MSE(\hat{Y}^U, Y^U_p) \quad (9)$$

Here, Y^U_p is pseudo label.

We adopt VNet as subnet A which is a popular choice in medical image segmentation. To implement inter-subnet mutual correcting, the performance gap of heterogeneous subnets should be slight. Therefore, the encoder of the VNet is replaced with a 3D convolutional ResNet34 as subnet B, called 3D-ResVNet. During inference, we use the average of the outputs of the two subnets as the final prediction result. This framework is implemented by PyTorch with an NVIDIA V100 GPU. And most of the parameter settings are consistent with the comparison methods. Specifically, the SGD optimizer is used to update the network parameters with weight decay 0.0001, and momentum 0.9. The initial learning rate is 0.01 and is divided by 10 after every 2500 iterations for a total of 6000 iterations. The batch size is 4, which includes 2 labeled data volumes and 2 unlabeled volumes. Following [10, 13, 26, 33] Gaussian warming up function is used to control the weight λ: $\lambda(t) = 0.1 * e^{-5(1-t/t_{\max})^2}$. Where t represents the current number of iterations, and $t_{\max}$ represents the total number of training iterations. β is empirically set to 0.5.

4. Experiments

Following the practice in the comparative literature, all methods (including MCF) are trained for 6K fixed iterations to obtain the final model. To exclude the effect of dataset partitioning, we perform K-fold cross-validation and report the mean and standard deviation on two different modal medical datasets.

4.1. Datasets and Implementation Details

The Left Atrial Dataset (LA). The LA dataset [29] includes 100 3D gadolinium-enhanced MR imaging volumes with an isotropic resolution of $0.625 \times 0.625 \times 0.625 \mathrm{mm}^3$ and the corresponding ground truth labels. We divide this dataset into 5 folds of 20 volumes each. For pre-processing, we first normalize all volumes to zero mean and unit variance, then crop each 3D MRI volume with enlarged margins according to the targets. During training, the training volumes are randomly cropped to $112 \times 112 \times 80$ as the model input. During inference, a sliding window of the same size is used to obtain segmentation results with a stride of $18 \times 18 \times 4$.

The NIH pancreas dataset. A publicly available NIH Pancreas Dataset [18] provides 82 contrast-enhanced abdominal 3D CT volumes with manual annotation. The size of each CT volume is $512 \times 512 \times \mathrm{D}$, where $\mathrm{D} \in [181, 466]$. We divide the NIH pancreas dataset into four folds, and the number of each fold is 20, 20, 21, 21, respectively. In pre-processing, like [13] we use the soft tissue CT window of $[-120, 240]$ HU, and we crop the CT scans centering at the pancreas region, and enlarge margins with 25 voxels. The training volumes are randomly cropped to $96 \times 96 \times 96$ and the stride is $16 \times 16 \times 16$ at inference.

Metrics. We use four metrics to evaluate model performance, including regional sensitive metrics: Dice similarity coefficient (Dice), Jaccard similarity coefficient (Jaccard), and edge sensitive metrics: 95% Hausdorff Distance (95HD) and Average Surface Distance (ASD).

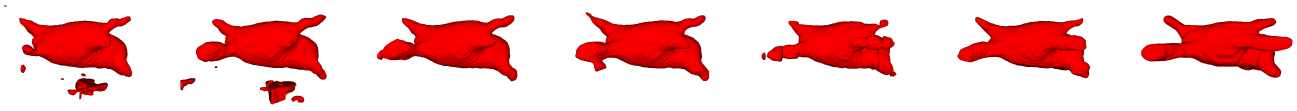
4.2. Comparison on the LA dataset

We first evaluate our proposed method on the left atrium segmentation task. The comparing methods include UA-MT [33] utilizing uncertainty-guided segmentation models, SASSNet [10] that incorporates geometric constraints into the network, DTC [13] that proposes multi-task consistency for medical image segmentation, and MC-Net [26] for mutual consistency learning with cycle pseudo-labels. In addition, we also implement the MT [22] based UA-MT for a more comprehensive comparison. The five-fold cross-validation results under 20% labeled data training are presented in Tab. 1. In addition, the metrics of VNet and 3D-ResVNet at 100% and 20% labeled data are reported as reference performance upper bound and baselines. As can be seen from the Tab. 1, all methods benefit from unlabeled data, but MT has the least gain and the worst stable performance with the largest standard deviation. UA-MT outperforms the MT method, illustrating that the uncertainty map can improve the performance of the student model. Compared with other existing methods, MC-Net achieves

Table 1. 5-fold cross-validation comparison results on the LA MRI dataset (average ± standard deviation)

Method	Volumes used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
VNet	80(100%)	0	91.28±0.008	84.07±0.012	5.00±0.757	1.61±0.291
3D-ResVNet	80(100%)	0	91.09±0.013	83.90±0.017	4.77±1.641	1.75±0.195
VNet	16(20%)	0	83.34±0.023	72.49±0.029	14.77±1.169	3.87±0.337
3D-ResVNet	16(20%)	0	84.09±0.022	73.56±0.025	17.36±2.748	4.96±1.008
MT	16(20%)	64	85.89±0.024	76.58±0.027	12.63±5.741	3.44±1.382
UA-MT	16(20%)	64	85.98±0.014	76.65±0.017	9.86±2.707	2.68±0.776
SASSNet	16(20%)	64	86.21±0.023	77.15±0.024	9.80±1.842	2.68±0.416
DTC	16(20%)	64	86.36±0.023	77.25±0.020	9.02±1.015	2.40±0.223
*MC-Net	16(20%)	64	87.65±0.011	78.63±0.013	9.70±2.361	3.01±0.700
MCF(Ours)	16(20%)	64	**88.71±0.018**	**80.41±0.022**	**6.32±0.800**	**1.90±0.187**

* means we report our reproduced results here because MC-Net does not release source code.



(a) MT (b) UA-MT (c) SASSNet (d) DTC (e) MCNet (f) Ours (g) GT

Figure 4. 3D segmentation visualization of different semi-supervised methods under 20% labeled on the LA dataset. (Best viewed in color)

the best results on Dice and Jaccard metrics with stable performance. While SASSNet outperforms other comparison methods on 95HD and ASD, showing that shape priors improve edge segmentation.

Notably, the MCF framework outperforms the SOTA method on all metrics, especially on the edge-sensitive metrics 95HD and ASD. Compared with SSANet, 95HD drops from 9.02 to 6.32, ASD drops from 2.40 to 1.90, and the performance is more stable. Fig. 4 shows the visualization results of MCF and comparison methods on left atrial segmentation. Compared with other methods, MCF has a higher overlap rate with labels and produces fewer false segmentations with more details.

4.3. Comparison on the pancreas dataset

The pancreas is located deep in the abdomen and varies considerably in size, location, and shape. In addition, pancreatic CT volumess have a more complex background compared to left atrial MRI volumes. Therefore, pancreas segmentation is more challenging than left atrial segmentation. Based on this, we conduct experiments on the pancreas dataset to further evaluate our proposed method. As shown in Tab. 2, we report the 4-fold cross-validation results under 12 labeled data and 50 unlabeled data. It can be seen that the segmentation metrics of all methods are worse than the left atrium segmentation, demonstrating the challenge of semi-supervised pancreas segmentation. Like left atrium segmentation, VNet outperforms 3D-ResNet in fully supervised settings, but 3D-ResVNet performs better overall with fewer labeled training samples. MT outperforms other comparison methods in Dice and Jaccard metrics, while DTC and SASSNet win at 95HD and ASD, respectively. What is more interesting is that the performance of these comparison methods on the two databases is not consistent, it seems that the method that performs poorly on LA segmentation shows an advantage on pancreas segmentation, such MT. Overall, the performance gap between the compared methods is not large. This shows the complexity of pancreas segmentation and the variability between different medical image segmentation tasks. We think this may be related to data processing, dataset size, volumetric properties, and method applicability. However, this also shows that our method has stronger robustness across datasets.

Notably, the MCF framework outperforms the SOTA method on all metrics, especially the 95HD drop of 1.61. Fig. 5 shows the visual segmentation results of these methods. Compared with other comparison methods, MCF obtains more accurate segmentation with a smoother and clearer edge. The other comparison methods are prone to discrete mispredictions and small protrusions or depressions in the edges.

4.4. Analysis and Ablation

CDR and bias correction. The purpose of CDR is to use prediction discrepancies to guide the network to correct its own mistakes. A simple and effective experiment is used to verify the effectiveness of CDR: We train two baseline networks (i.e., VNet and 3D-ResVNet) with and without

Table 2. 4-fold cross-validation comparison results on the Pancreas CT dataset (average ± standard deviation)

Method	Volumes used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
VNet	62(100%)	0	80.75±0.010	68.49±0.014	7.23±0.564	1.69±0.363
3D-ResVNet	62(100%)	0	79.78±0.021	67.29±0.025	7.30±1.632	1.60±0.074
VNet	12(20%)	0	64.18±0.073	49.26±0.077	17.74±3.572	4.69±0.935
3D-ResVNet	12(20%)	0	66.53±0.043	51.25±0.047	19.01±4.129	5.64±1.467
MT	12(20%)	50	74.43±0.024	60.53±0.030	14.93±2.000	4.61±0.929
UA-MT	12(20%)	50	74.01±0.029	60.00±3.031	17.00±3.031	5.19±1.267
SASSNet	12(20%)	50	73.57±0.017	59.71±0.020	13.87±1.079	3.53±1.416
DTC	12(20%)	50	73.23±0.024	59.18±0.027	13.20±2.241	3.81±0.953
*MC-Net	12(20%)	50	73.73±0.019	59.19±0.021	13.65±3.902	3.92±1.055
MCF(Ours)	12(20%)	50	**75.00±0.026**	**61.27±0.030**	**11.59±1.611**	**3.27±0.919**

* means we report our reproduced results here because MC-Net does not release source code.

(a) MT (b) UA-MT (c) SASSNet (d) DTC (e) MCNet (f) Ours (g) GT

Figure 5. 3D segmentation visualization of different semi-supervised methods under 20% labeled on the LA dataset. (Best viewed in color)

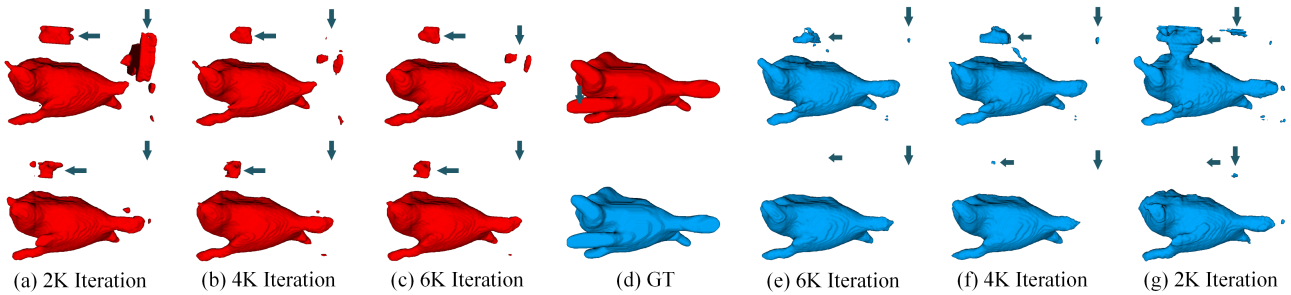
CDR on the LA dataset and show the single network performance in Tab. 3. Note that V and R in this table represent VNet and 3D-ResVNet, respectively. We can see that the performances of the two baselines without CDR are similar. CDR brings significant and consistent improvement to the baseline networks in all metrics, especially 95HD: VNet drops 5.83 points and 3D-ResVNet drops 4.13 points. This is because the bias pixels are mainly located on the edge of the object, like shown in Fig. 1 (d). Regarding the region-sensitive metrics (Dice and Jaccard), both baselines improved by about 2 points after adding CDR. Although VNet with CDR has the smallest improvement in ASD, it is also close to 1 point. In general, VNets benefit more from CDR. The segmentation visualization results are shown in Fig. 6. Red and blue represent the results of VNet and 3D-ResVNet, respectively. First and second rows represent w/ or w/o CDR, respectively. It can be clearly found that there are some wrong predictions that always exist during the training process of these baseline networks, and CDR helps the network correct these errors. This confirms our hypothesis that although the network has a strong learning ability, they easily fall into their own biases. Moreover, CDR reduces out-of-body discrete errors and makes edges clearer.

DCPLG and consistency. To show the performance changes during training for both consistency regularization and DCPLG, we replace DCPLG in MCF with consistency regularization i.e., encourage subnets to have the same predictions, and remove other redundant components to avoid superfluous influence. Fig. 7 shows the dynamic changes in the performance of the model on the LA dataset with 20% labeled data under these two settings. An interesting finding is that all metrics of consistency regularization decrease to varying degrees with increasing training steps. Specifically, before 2K iterations, the consistency regularization outperforms DCPLG, but after 2K iterations, the situation reverses. In subsequent training, the gap gradually widens, especially in Dice, Jaccard, and 95HD. And after 5K iterations, both methods show performance degradation, but DCPLG is slight.

Table 3. Ablation results of different components of MCF. We conducted extensive experiments to test the impact of CDR and DCPLG on model performance. Here "V" refers to VNet and "R" refers to 3D-ResVNet.

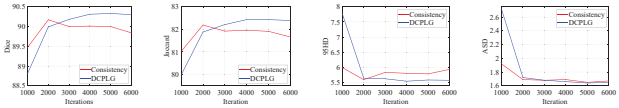
Method	Metrics			
	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
V	85.63	75.67	14.40	3.69
V+CDR	87.69	78.34	8.57	2.31
V+DCPLG	89.81	81.62	7.32	2.36
R	85.67	75.49	13.16	3.25
R+CDR	87.57	78.14	9.03	2.34
R+DCPLG	89.27	80.69	6.86	2.10
V-m-R	86.53	76.89	11.30	2.70
V+R+CDR	88.21	79.16	7.89	1.98
V+R+DCPLG	90.13	81.92	6.73	1.86
MCF	90.49	82.70	5.62	1.61

Effects of different components. Ablation experiments are performed to demonstrate the effectiveness of MCF, and the results are presented in Tab. 3. First, since the predicted results of MCF are obtained by averaging the results of the



(a) 2K Iteration (b) 4K Iteration (c) 6K Iteration (d) GT (e) 6K Iteration (f) 4K Iteration (g) 2K Iteration

Figure 6. Segmentation visualization of VNet and 3D-ResVNet w/ or w/o CDR. Note: Red and blue represent the results of VNet and 3D-ResVNet, respectively. First and second rows represent segmentation results w/o and w/ CDR, respectively. (Best viewed in color)



(a) Dice (b) Jaccard (c) 95HD (d) ASD

Figure 7. Performance comparison of consistency regularization training and DCPLG pseudo label training.

Table 4. Comparison when different loss functions are used as rectification loss.

Loss function	Metrics			
	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
CE loss	90.28	82.34	5.60	1.68
MSE loss	90.49	82.70	5.62	1.61

Table 5. Ablation results for different values of β on LA

β	Metrics			
	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
0.3	90.23	82.32	5.89	1.69
0.4	90.27	82.41	5.85	1.64
0.5	90.49	82.70	5.62	1.61
0.6	90.26	82.43	5.86	1.65
0.7	90.24	82.38	5.88	1.67

two subnets, we verify the performance when integrating VNet and 3D-ResNet with the averaging operation, i.e. V-m-R in the Tab. 3. The performance of the simple average model is improved, but it is weaker than the V/R+CDR. V/R+CDR means that the model is trained by V+R+CDR and then tested subnets separately, Setting V/R+DCPLG is similar. Compared with V-m-R, V+R+CDR achieves gains of 1.68, 2.27, 3.41, and 0.72 on Dice, Jaccard, 95HD, and ASD, respectively. While the gains of V+R+DCPLG are 3.6, 5.02, 4.57 and 0.84, indicating that DCPLG helps the model learn from unlabeled data. In the end, the complete MCF model achieved the best results.

Ablation for rectification loss. The discrepancy prediction regions between subnets are very small and scattered, which are not suitable for implementing Dice loss, these predictions are obtained under the guidance of CE loss, so a common other loss i.e. MSE is adopted. Tab. 4 reports the segmentation performance of the model when CE and MSE are used as rectification losses, respectively. Compared with CE, MSE achieves better segmentation results, so we adopt MSE as the correction loss.

Ablation for hyperparamerters β. we find that the model is not very sensitive to the hyperparameter β, and Tab. 5 shows the metrics when β takes 0.3-0.7. According to this empirical result, β is set to 0.5 in this work.

5. Discussions and Conclusion

In this paper, we propose a new framework called MCF for semi-supervised medical image segmentation, which enables the network to be aware of its own mistakes and perform bias correction through inter-subnet comparisons. To unleash the potential of the dual-subnet architecture, MCF introduces two different subnets that update parameters independently. The CDR takes the difference predictions of the subnets as potential bias areas and guides the network to review and correct them. This is almost a free lunch and can be easily integrated into other semi-supervised or fully-supervised methods. The DCPLG is essentially like a mock exam that students take before their final exams. Combined with the characteristics of medical image segmentation datasets (i.e., all samples are related to the target), DCPLG is used to dynamically select pseudo-label generators to improve the quality of pseudo-labels. Inheriting the idea of DCPLG to modify and apply it to natural image processing is our further work. Finally, segmentation results on two public benchmark datasets with different modalities demonstrate the potential of MCF on semi-supervised medical images, which achieves state-of-the-art performance.

6. Acknowledgements

This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0110800, in part by the National Natural Science Foundation of China under Grant 62172067 and 61976031, in part by the National Major Scientific Research Instrument Development Project of China under Grant 62027827, in part by Chongqing University of Posts and Telecommunications Ph.D. Innovative Talents Project under Grant BYJS202216.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 1

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[3] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021. 1, 2

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *arXiv preprint arXiv:2112.09645*, 2021. 3

[5] Jun Chen, Heye Zhang, Raad Mohiaddin, Tom Wong, David Firmin, Jennifer Keegan, and Guang Yang. Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data. *IEEE Transactions on Medical Imaging*, 41(2):420–433, 2022. 3

[6] WeiWang Dong-DongChen and Zhi-HuaZhou WeiGao. Tri-net for semi-supervised deep learning. In *Proceedings of twenty-seventh international joint conference on artificial intelligence*, pages 2014–2020, 2018. 2

[7] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022. 3

[8] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[9] Caizi Li, Li Dong, Qi Dou, Fan Lin, Kebao Zhang, Zuxin Feng, Weixin Si, Xuesong Deng, Zhe Deng, and Pheng-Ann Heng. Self-ensembling co-training framework for semi-supervised covid-19 ct segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(11):4140–4151, 2021. 3

[10] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 5

[11] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 3

[12] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20706, 2022. 1

[13] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021. 1, 3, 5

[14] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8896–8905, 2018. 1, 2, 3

[15] Robert Mendel, Luis Antonio de Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020. 3

[16] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 3

[17] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020. 1, 2

[18] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 5

[19] Constantin Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. *arXiv preprint arXiv:2112.00735*, 2021. 3

[20] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE transactions on medical imaging*, 41(3):608–620, 2021. 3

[21] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 1

[22] 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. 1, 2, 5

[23] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019. 1, 2, 3

[24] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Tripled-uncertainty guided mean

teacher model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 450–460. Springer, 2021. 3

[25] Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, and Hau-San Wong. Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6500–6509, 2019. 3

[26] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–306. Springer, 2021. 3, 5

[27] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 4

[28] Yutong Xie, Jianpeng Zhang, Zhibin Liao, Johan Verjans, Chunhua Shen, and Yong Xia. Intra-and inter-pair consistency for semi-supervised gland segmentation. *IEEE Transactions on Image Processing*, 2021. 3

[29] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67:101832, 2021. 5

[30] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, June 2022. 1

[31] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. 1

[32] Huifeng Yao, Xiaowei Hu, and Xiaomeng Li. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3099–3107, 2022. 3

[33] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 1, 5

[34] Ruifei Zhang, Sishuo Liu, Yizhou Yu, and Guanbin Li. Self-supervised correction learning for semi-supervised biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 134–144. Springer, 2021. 3

[35] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005. 2