Co-training with High-Confidence Pseudo Labels for Semi-supervised Medical Image Segmentation

Zhiqiang Shen\textsuperscript{1,2}, Peng Cao\textsuperscript{1,2},*, Hua Yang\textsuperscript{3}, Xiaoli Liu \textsuperscript{4}, Jinzhu Yang\textsuperscript{1,2}, Osmar R. Zaiane\textsuperscript{5}

\textsuperscript{1}College of Computer Science and Engineering, Northeastern University, Shenyang, China
\textsuperscript{2}Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang, China
\textsuperscript{3}College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, China
\textsuperscript{4}DAMO Academy, Alibaba Group, China
\textsuperscript{5}Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada

\texttt{xxszqyy@gmail.com, caopengneu@gmail.com, qsx20200801@student.fjnu.edu.cn, liuxiaoli.lxl@alibaba-inc.com, yangjinzhu@cse.neu.edu.cn, zaiane@cs.ualberta.ca}

Abstract

Consistency regularization and pseudo labeling-based semi-supervised methods perform co-training using the pseudo labels from multi-view inputs. However, such co-training models tend to converge early to a consensus, degenerating to the self-training ones, and produce low-confidence pseudo labels from the perturbed inputs during training. To address these issues, we propose an Uncertainty-guided Collaborative Mean-Teacher (UCMT) for semi-supervised semantic segmentation with the high-confidence pseudo labels. Concretely, UCMT consists of two main components: 1) collaborative mean-teacher (CMT) for encouraging model disagreement and performing co-training between the sub-networks, and 2) uncertainty-guided region mix (UMIX) for manipulating the input images according to the uncertainty maps of CMT and facilitating CMT to produce high-confidence pseudo labels. Combining the strengths of UMIX with CMT, UCMT can retain model disagreement and enhance the quality of pseudo labels for the co-training segmentation. Extensive experiments on four public medical image datasets including 2D and 3D modalities demonstrate the superiority of UCMT over the state-of-the-art. Code is available at: https://github.com/Senyh/UCMT.

1 Introduction

Semantic segmentation is critical for medical image analysis. Great progress has been made by deep learning-based segmentation models relying on a large amount of labeled data [Chen et al., 2018; Ronneberger et al., 2015]. However, labeling such pixel-level annotations is laborious and requires expert knowledge especially in medical images, resulting in that labeled data are expensive or simply unavailable. Unlabeled data, on the contrary, are cheap and relatively easy to obtain. Under this condition, semi-supervised learning (SSL) has been the dominant data-efficient strategy through exploiting information from a limited amount labeled data and an arbitrary amount of unlabeled data, so as to alleviate the label scarcity problem [Van Engelen and Hoos, 2020].

Consistency regularization [Tarvainen and Valpola, 2017] and pseudo labeling [Lee and others, 2013] are the two main methods for semi-supervised semantic segmentation. Currently, combining consistency regularization and pseudo labeling via cross supervision between the sub-networks, has shown promising performance for semi-supervised segmentation [Lee and others, 2013; Ouali et al., 2020; Ke et al., 2020; Chen et al., 2021; Liu et al., 2022]. One critical limitation of these approaches is that the sub-networks tend to converge early to a consensus situation causing the co-training model degenerating to the self-training [Yu et al., 2019b]. Disagreement between the sub-networks is crucial for manipulating the input images according to the disagreement variation between the co-training sub-networks in terms of dice loss (w.r.t. number of iterations) of two branches (Y\textsuperscript{1} and Y in MT; Y\textsuperscript{1} and Y\textsuperscript{2} in CPS; Y\textsuperscript{1} and Y\textsuperscript{2} in UCMT), the uncertainty variation of the pseudo labels in terms of entropy w.r.t. number of iterations, and (f) the performance of MT, CPS, and UCMT on semi-supervised skin lesion segmentation under different ratio of labeled data. EMA: exponential moving average.

Figure 1: Illustration of the architectures and curves for co-training based semi-supervised semantic segmentation. (a) Mean-teacher [Tarvainen and Valpola, 2017], (b) Cross pseudo supervision [Chen et al., 2021], (c) Uncertainty-guided collaborative mean-teacher, (d) the disagreement variation between the co-training sub-networks in terms of dice loss (w.r.t. number of iterations) of two branches (Y\textsuperscript{1} and Y in MT; Y\textsuperscript{1} and Y\textsuperscript{2} in CPS; Y\textsuperscript{1} and Y\textsuperscript{2} in UCMT), (e) the uncertainty variation of the pseudo labels in terms of entropy w.r.t. number of iterations, and (f) the performance of MT, CPS, and UCMT on semi-supervised skin lesion segmentation under different ratio of labeled data. EMA: exponential moving average.
for co-training, where the sub-networks initialized with different parameters or trained with different views have different biases (i.e., disagreement) ensuring that the information they provide is complementary to each other. Another key factor affecting the performance of these approaches is the quality of pseudo labels. Intuitively, high quality pseudo labels should have low uncertainty [Grandvalet and Bengio, 2004]. More importantly, these two factors influence each other. Increasing the degree of the disagreement between the co-training sub-networks by different perturbations or augmentations could result in their opposite training directions, thus increasing the uncertainty of pseudo labels. To investigate the effect of the disagreement and the quality of pseudo labels for co-training based semi-supervised segmentation, which has not been studied in the literature, we conduct a pilot experiment to illustrate these correlations. As shown in Figure 1, compared with mean-teacher (MT) [Tarvainen and Valpola, 2017] [Figure 1 (a)], cross pseudo supervision (CPS) [Chen et al., 2021] [Figure 1 (b)] with the higher model disagreement ([d]) and the lower uncertainty [Figure 1 (e)] produces higher performance [Figure 1 (f)] on semi-supervised segmentation. Note that the dice loss of two branches are calculated to measure the disagreement. The question that comes to mind is: how to effectively improve the disagreement between the co-training sub-networks and the quality of pseudo labels jointly in a unified network for SSL.

In this paper, we focus on two major goals: maintaining model disagreement and the high-confidence pseudo labels at the same time. To this end, we propose the Uncertainty-guided Collaborative Mean Teacher (UCMT) framework that is capable of retaining higher disagreement between the co-training segmentation sub-networks [Figure 1 (d)] based on the higher confidence pseudo labels [Figure 1 (e)], thus achieving better semi-supervised segmentation performance under the same backbone network and task settings [Figure 1 (f)]. Specifically, UCMT involves two major components: 1) collaborative mean-teacher (CMT), and 2) uncertainty-guided region mix (UMIX), where UMIX operates the input images according to the uncertainty maps of CMT while CMT performs co-training under the supervision of the pseudo labels derived from the UMIX images. Inspired by the co-teaching [Han et al., 2018; Yu et al., 2019b; Chen et al., 2021] for struggling with early converging to a consensus situation and degrading into self-training, we introduce a third component, the teacher model, into the co-training framework as a regularizer to construct CMT for more effective SSL. The teacher model acts as self-ensemble by averaging the student models, serving as a third part to guide the training of the two student models. Further, we develop UMIX to construct high-confident pseudo labels and perform regional dropout for learning robust semi-supervised semantic segmentation models. Instead of random region erasing or swapping [DeVries and Taylor, 2017; Yun et al., 2019], UMIX manipulates the original image and its corresponding pseudo labels according to the epistemic uncertainty of the segmentation models, which not only reduces the uncertainty of the pseudo labels but also enlarges the training data distribution. Finally, by combining the strengths of UMIX with CMT, the proposed approach UCMT significantly improves the state-of-the-art (sota) results in semi-supervised segmentation on multiple benchmark datasets. For example, UCMT and UCMT(U-Net) achieve 88.22% and 82.14% Dice Similarity Coefficient (DSC) on ISIC dataset under 5% labeled data, outperforming our baseline model CPS [Chen et al., 2021] and the state-of-the-art UGCL [Wang et al., 2022] by 1.41% and 9.47%, respectively.

In a nutshell, our contributions mainly include:

- We pinpoint the problem in existing co-training based semi-supervised segmentation methods: the insufficient disagreement among the sub-networks and the lower-confidence pseudo labels. To address the problem, we design an uncertainty-guided collaborative mean-teacher to maintain co-training with high-confidence pseudo labels, where we incorporate CMT and UMIX into a holistic framework for semi-supervised medical image segmentation.
- To avoid introducing noise into the new samples, we propose an uncertainty-guided regional mix algorithm, UMIX, encouraging the segmentation model to yield high-confidence pseudo labels and enlarge the training data distribution.
- We conduct extensive experiments on four public medical image segmentation datasets including 2D and 3D scenarios. Comprehensive results demonstrate the effectiveness of each component of our method and the advantage of UCMT over the state-of-the-art.

2 Related Work
2.1 Semi-Supervised Learning

Semi-supervised learning aims to improve performance in supervised learning by utilizing information generally associated with unsupervised learning, and vice versa [Van Egenlen and Hoos, 2020]. A common form of SSL is introducing a regularization term into the objective function of supervised learning to leverage unlabeled data. From this perspective, SSL-based methods can be divided into two main lines, i.e., pseudo labeling and consistency regularization. Pseudo labeling attempts to generate pseudo labels similar to the ground truth, for which models are trained as in supervised learning [Lee and others, 2013]. Consistency regularization enforces the model’s outputs to be consistent for the inputs under different perturbations [Tarvainen and Valpola, 2017]. Current state-of-the-art approaches have incorporated these two strategies and shown convincing performance for semi-supervised image classification [Sohn et al., 2020; Zhang et al., 2021]. Based on this line of research, we explore more effective consistency learning algorithms for semi-supervised segmentation.

2.2 Semi-Supervised Semantic Segmentation

Compared with image classification, semantic segmentation requires much more intensively and costly labeling for pixel-level annotations. Semi-supervised semantic segmentation inherits the main ideas of semi-supervised image classification. The combination of consistency regularization and
pseudo labeling, mainly conducting cross supervision between sub-networks using pseudo labels, has become the mainstream strategy for semi-supervised semantic segmentation in both natural images [Ouali et al., 2020; Chen et al., 2021] and medical images [Yu et al., 2019a; Luo et al., 2021; Wu et al., 2021b; Wu et al., 2022]. Specifically, these combined approaches enforce the consistency of the predictions under different perturbations, such as input perturbations [Li et al., 2020b; Tu et al., 2022], feature perturbations [Ouali et al., 2020], and network perturbations [Tarvainen and Valpola, 2017; Chen et al., 2021; Wu et al., 2021b; Wu et al., 2022]. In addition, adversarial learning-based methods, rendering the distribution of model predictions from labeled data to be aligned with those from unlabeled data, can also be regarded as a special form of consistency regularization [Hung et al., 2018; Li et al., 2020a]. However, such cross supervision models may converge early to a consensus, thus degenerating to self-training ones. We hypothesize that enlarging the disagreement for the co-training models based on the high-confidence pseudo labels can improve the performance of SSL. Therefore, we propose a novel SSL framework, i.e., UCMT, to generate more accurate pseudo labels and maintain co-training for semi-supervised medical image segmentation.

2.3 Uncertainty-Guided Semi-Supervised Semantic Segmentation

Model uncertainty (epistemic uncertainty) can guide the SSL models to capture information from the pseudo labels. Two critical problems for leveraging model uncertainty are how to obtain and exploit model uncertainty. Recently, there are mainly two strategies to estimate model uncertainty: 1) using Monte Carlo dropout [Gal and Ghahramani, 2016], and 2) calculating the variance among different predictions [Zheng and Yang, 2021]. For semi-supervised semantic segmentation, previous works exploit model uncertainty to re-weight the training loss [Yu et al., 2019a] or selecting the contrastive samples [Wang et al., 2022]. However, these methods require manually setting a threshold to neglect the low-confidence pseudo labels, where the fixed threshold is hard to determine. In this paper, we obtain the epistemic uncertainty by the entropy of the predictions of CMT for the same input and exploit the uncertainty to guide the region mix for gradually exploring information from the unlabeled data.

3 Methodology

Before introducing our method, we first define the notations used in this work. The training set $D = \{D_L,D_U\}$ contains a labeled set $D_L = \{(X_i,Y_i)\}_{i=1}^N$ and an unlabeled set $D_U = \{(X_j)\}_{j=N+1}^M$, where $X_i,Y_i$ denotes the $i$th labeled image, and $N$ and $M-N$ are the number of labeled and unlabeled samples, respectively. Given the training data $D$, the goal of semi-supervised semantic segmentation is to learn a model $f(\cdot;\theta)$ performing well on unseen test sets.

3.1 Overview

To avoid the co-training degrading to the self-training, we propose to encourage model disagreement during training and ensure pseudo labels with low uncertainty. With this motivation, we propose uncertainty-guided collaborative mean-teacher for semi-supervised image segmentation, which includes 1) collaborative mean-teacher, and 2) uncertainty-guided region mix. As shown in Figure 1 (d), CMT and UCMT gradually enlarge the disagreement between the co-training sub-networks. Meanwhile, CMT equipped with UMIX guarantees low-uncertainty for the pseudo labels. With the help of these conditions, we can safely maintain the co-training status to improve the effectiveness of SSL for exploring unlabeled data. Figure 2 illustrates the schematic diagram of the proposed UCMT. Generally, there are two steps in the training phase of UCMT. In the first step, we train CMT using the original labeled and unlabeled data to obtain the uncertainty maps; Then, we perform UMIX to generate the new samples based on the uncertainty maps. In the second step, we re-train CMT using the UMIX samples. Details of the training process of UCMT are shown in Algorithm 1. Although UCMT includes three models, i.e., one teacher model and two student models, only the teacher model is required in the testing stage.

3.2 Collaborative Mean-Teacher

Current consistency learning-based SSL algorithms, e.g., Mean-teacher [Tarvainen and Valpola, 2017] and CPS [Chen et al., 2021], suggest to perform consistency regularization among the pseudo labels in a multi-model architecture rather than in a single model. However, during the training process, the two-network SSL framework may converge early to a consensus and the co-training degenerate to the self-training [Yu et al., 2019b]. To tackle this issue, we design the collaborative mean teacher (CMT) framework by introducing a “arbiter”, i.e., the teacher model, into the co-training architecture [Chen et al., 2021] to guide the training of the two student models. As shown in Figure 2, CMT consists of one teacher model $f(\cdot;\theta)$ and two student models $f(\cdot;\theta_1)$ and $f(\cdot;\theta_2)$,
Algorithm 1 UCMT algorithm

Input: \( \mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^N \), \( \mathcal{D}_U = \ {(X_j)_{j=N+1}^M} \)
Parameter: \( \theta, \theta_1, \theta_2 \)
Output: \( f(\cdot; \theta) \)

1: for \( T \in [1, \text{numepochs}] \) do
2: for each minibatch \( B \) do
3: \( \text{// \# } \)
4: \( \text{step 1: uncertainty estimation} \)
5: \( \hat{Y}_i^0 \leftarrow f(X_i; \theta) \)
6: \( \hat{Y}_i^1 \leftarrow f(X_i; \theta_1) \)
7: \( \hat{Y}_i^2 \leftarrow f(X_i; \theta_2) \)
8: Update \( f(\cdot; \theta) \), \( f(\cdot; \theta_1) \), \( f(\cdot; \theta_2) \) using optimizer
9: \( U_i^{[1/2]} \leftarrow \text{Uncertain}(f(X_i; \theta_{[1/2]}), f(X_i; \theta)) \)
10: \( U_i^{[1]} \leftarrow \text{Uncertain}(f(X_i; \theta_{[1/2]}), f(X_i; \theta)) \)
11: \( \text{step 2: training with UMIX} \)
12: \( X_i^0/Y_i^0 \leftarrow \text{UMIX}(X_i/Y_i; U_i^0; \alpha; k, r) \)
13: \( X_i^0/Y_i^0 \leftarrow \text{UMIX}(X_i/Y_i; U_i^0; \alpha; k, r) \)
14: Repeat 5-8 using \( X_i^0, X_i^0, Y_i^0, \) and \( Y_i^0 \)
15: end for
16: end for
17: return \( f(\cdot; \theta) \)

where the teacher model is the self-ensemble of the average of the student models. These models have the same architecture but initialized with different weights for network perturbations. For labeled data, these models are all optimized by supervised learning. For unlabeled data, there are two critical factors: 1) co-training between the two student models, and 2) direct supervision from the teacher to the student models.

To explore both the labeled and unlabeled data, the total loss \( L \) for training UCMT involves two parts, i.e., the supervised loss \( L_s \) and the unsupervised loss \( L_u \).

\[
L = L_s + \lambda L_u. \tag{1}
\]

where \( \lambda \) is a regularization parameter to balance the supervised and unsupervised learning losses. We adopt a Gaussian ramp-up function to gradually increase the coefficient, i.e., \( \lambda(t) = \lambda_m \times \exp[-5(1 - \frac{t}{t_m})^2] \), where \( \lambda_m \) scales the maximum value of the weighted function, \( t \) denotes the current iteration, and \( t_m \) is the maximum iteration in training.

Supervised Learning Path. For the labeled data, the supervised loss is formulated as,

\[
L_s = \frac{1}{N} \sum_{i=1}^N \{ L_{seg}(f(X_i; \theta_1), Y_i) + L_{seg}(f(X_i; \theta_2), Y_i) \}, \tag{2}
\]

where \( L_{seg} \) can be any supervised semantic segmentation loss, such as cross entropy loss and dice loss. Note that we choose dice loss in our experiments as its compelling performance in medical image segmentation.

Unsupervised Learning Path. The unsupervised loss \( L_u \) acts as a regularization term to explore potential knowledge for the labeled and unlabeled data. \( L_u \) includes the cross-supervised \( L_{csp} \) between the two student models and the mean-teacher supervision \( L_{mts} \) for guiding the student models from the teacher, as follow:

\[
L_u = L_{cps} + L_{mts}. \tag{3}
\]

1) Cross Pseudo Supervision. The aim of \( L_{cps} \) is to promote two students to learn from each other and to enforce the consistency between them. Let \( L_{cps} = L_{cps1} + L_{cps2} \) for encouraging bidirectional interaction for the two student sub-networks \( f(\cdot; \theta_1) \) and \( f(\cdot; \theta_2) \). The losses of \( L_{cps1} \) and \( L_{cps1} \) are defined as:

\[
L_{cps1/2} = \frac{1}{M-N} \sum_{j=1}^{M-N} L_{seg}(f(X_j; \theta_{[1/2]}), \hat{Y}_j^{[1/2]}), \tag{4}
\]

where \( \hat{Y}_j^1 \) and \( \hat{Y}_j^2 \) are the pseudo segmentation maps for \( X_j \) predicted by \( f(\cdot; \theta_1) \) and \( f(\cdot; \theta_1) \), respectively.

2) Mean-Teacher Supervision. To avoid the two students co-training in the wrong direction, we introduce a teacher model to guide the optimization of the student models. Specifically, the teacher model is updated by the exponential moving average (EMA) of the average of the student models: \( \theta^t = \alpha \theta^{t-1} + (1 - \alpha)\beta \theta_t^1 + (1 - \beta) \theta_t^2 \), where \( t \) represents the current training iteration and \( \alpha \) is the EMA decay that controls the parameters' updating rate.

The loss of mean-teacher supervision \( L_{mts} = L_{mts1} + L_{mts2} \) is calculated from two branches:

\[
L_{mts1/2} = \frac{1}{M-N} \sum_{j=1}^{M-N} L_{seg}(f(X_j; \theta_{[1/2]}), \hat{Y}_j^0), \tag{5}
\]

where \( \hat{Y}_j^0 \) is the segmentation map derived from \( f(X_j; \theta) \).

3.3 Uncertainty-Guided Mix

Although CMT can promote model disagreement for co-training, it also slightly increases the uncertainty of the pseudo labels as depicted in Figure 1. On the other hand, random regional dropout can expand the training distribution and improve the generalization capability of models [DeVries and Taylor, 2017; Yun et al., 2019]. However, such random perturbations to the input images inevitably introduce noise into the new samples, thus deteriorating the quality of pseudo labels for SSL. One sub-network may provide some incorrect pseudo labels to the other sub-networks, degrading their performance. To overcome these limitations, we propose UMIX to manipulate image patches under the guidance of the uncertainty maps produced by CMT. The main idea of UMIX is constructing a new sample by replacing the top \( k \) most uncertain (low-confidence) regions with the top \( k \) most certain (high-confidence) regions in the input image. As illustrated in Figure 2, UMIX constructs a new sample \( X' = \text{UMIX}(X, U^1, U^2, k, r) \) by replacing the top \( k \) most uncertain regions (red grids in \( V^1 \) and \( V^2 \)) with the top \( k \) most certain regions (green grids in \( V^1 \) and \( V^2 \)) in \( X \), where each region has size \( 1/r \) to the image size. To ensure the reliability of the uncertainty evaluation, we obtain the uncertain maps by integrating the outputs of the teacher and the student model instead of performing \( T \) stochastic forward passess designed by Monte Carlo.
Carlo Dropout estimate model [Gal and Ghahramani, 2016; Yu et al., 2019a], which is equivalent to sampling predictions from the previous and current iterations. This process can be formulated as:

\[
U^m = \text{Uncertain}(f(X; \theta_m), f(X; \theta)) = -\sum_c P_c \log(P_c),
\]

\[
P_c = \frac{1}{2}(\text{Softmax}(f(X; \theta_m)) + \text{Softmax}(f(X; \theta)));
\]

where \( m = 1, 2 \) denotes the index of the student models and \( c \) refers to the class index.

4 Experiments and Results

4.1 Experiments Settings

Datasets. We conduct extensive experiments on different medical image segmentation tasks to evaluate the proposed method, including skin lesion segmentation from dermoscopy images, polyp segmentation from colonoscopy images, and the 3D left atrium segmentation from cardiac MRI images.

Dermoscopy. We validate our method on the ISIC dataset [Codella et al., 2018] including 2594 dermoscopy images and corresponding annotations. Following [Wang et al., 2022], we adopt 1815 images for training and 779 images for validation.

Colonoscopy. We evaluate the proposed method on the two public colonoscopy datasets, including Kvasir-SEG [Jha et al., 2020] and CVC-ClinicDB [Bernal et al., 2015]. Kvasir-SEG and CVC-ClinicDB contain 1000 and 612 colonoscopy images with corresponding annotations, respectively.

Cardiac MRI. We evaluate our method on the 3D left atrial (LA) segmentation challenge dataset, which consists of 100 3D gadolinium-enhanced magnetic resonance images and LA segmentation masks for training and validation. Following [Yu et al., 2019a], we split the 100 scans into 80 samples for training and 20 samples for evaluation.

Implementation Details

We use DeepLabv3+ [Chen et al., 2018] equipped with ResNet50 as the baseline architecture for 2D image segmentation, and adopt VNet [Milletari et al., 2016] as the baseline in the 3D scenario. All images are resized to \( 256 \times 256 \) for inference, while the outputs are recovered to the original size for evaluation, in the 2D scenario. For 3D image segmentation, we randomly crop \( 80 \times 112 \times 112 (\text{Depth} \times \text{Height} \times \text{Width}) \) patches for training and iteratively crop patches using a sliding window strategy to obtain the final segmentation mask for testing. We empirically set \( \lambda_m = 1, k = 2, r = 16, \alpha = 0.99 \) and \( \beta = 0.99 \) for our method in the experiments. We implement our method using PyTorch framework on a NVIDIA Quadro RTX 6000 GPU. We adopt AdamW as an optimizer with the fixed learning rate of \( \text{le-4} \). The batchsize is set to 16, including 8 labeled samples and 8 unlabeled samples. All 2D models are trained for 50 epochs, while the 3D models are trained for 1000 epochs \(^1\).

\(^1\)Since UCMT performs the two-step training within one iteration, it is trained for half of the epochs.

<table>
<thead>
<tr>
<th>Method</th>
<th>15% ( D_L )</th>
<th>30% ( D_L )</th>
<th>5% ( D_L )</th>
<th>10% ( D_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT [Tarvainen and Valpola, 2017]</td>
<td>86.67</td>
<td>87.42</td>
<td>88.22</td>
<td>88.46</td>
</tr>
<tr>
<td>CCT [Ouali et al., 2020]</td>
<td>83.97</td>
<td>86.43</td>
<td>88.08</td>
<td>89.06</td>
</tr>
<tr>
<td>CPS [Chen et al., 2021]</td>
<td>86.81</td>
<td>87.70</td>
<td>88.68</td>
<td>89.46</td>
</tr>
<tr>
<td>UGCL(U-Net) [Wang et al., 2022]</td>
<td>72.67</td>
<td>79.48</td>
<td>72.67</td>
<td>79.48</td>
</tr>
<tr>
<td>UCMT(U-Net) (ours)</td>
<td>82.14</td>
<td>83.33</td>
<td>88.22</td>
<td>88.46</td>
</tr>
<tr>
<td>CMT (ours)</td>
<td>87.86</td>
<td>88.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCMT (ours)</td>
<td><strong>88.22</strong></td>
<td><strong>88.46</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison with state-of-the-art methods on ISIC dataset. 5% \( D_L \) and 10% \( D_L \) of the labeled data are used for training, respectively. Results are measured by DSC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Kvasir-SEG</th>
<th>CVC-ClinicDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15% ( D_L )</td>
<td>30% ( D_L )</td>
</tr>
<tr>
<td>AdvSemSeg</td>
<td>56.88</td>
<td>76.09</td>
</tr>
<tr>
<td>ColAdv</td>
<td>76.76</td>
<td>80.95</td>
</tr>
<tr>
<td>MT</td>
<td>87.44</td>
<td>88.72</td>
</tr>
<tr>
<td>CCT</td>
<td>81.14</td>
<td>84.67</td>
</tr>
<tr>
<td>CPS</td>
<td>86.44</td>
<td>88.71</td>
</tr>
<tr>
<td>CMT (ours)</td>
<td>88.08</td>
<td>88.61</td>
</tr>
<tr>
<td>UCMT (ours)</td>
<td><strong>88.68</strong></td>
<td><strong>89.06</strong></td>
</tr>
</tbody>
</table>

Table 2: Comparison with state-of-the-art methods on Kvasir-SEG and CVC-ClinicDB datasets in terms of DSC. 15% \( D_L \) and 30% \( D_L \) of the labeled data are individually used for training.

4.2 Comparison with State of the Arts

We compare the proposed method with state-of-the-art on the four public medical image segmentation datasets. We reimplement MT [Tarvainen and Valpola, 2017], CCT [Ouali et al., 2020], and CPS [Chen et al., 2021] by adopting implementations from [Chen et al., 2021]. For other approaches, we directly use the results reported in their original papers.

Results on Dermoscopy. In Table 1, we report the results of our methods on ISIC and compare them with other state-of-the-art approaches. UCMT substantially outperforms all previous methods and sets new state-of-the-art of 88.22% DSC and 88.46 DSC under 5% and 10% labeled data. For fair comparison with UGCL [Wang et al., 2022], replace the backbone of UCMT with U-Net. The results indicate that our UCMT(U-Net) exceeds UGCL by a large margin. Moreover, our CMT version also outperforms other approaches under the two labeled data rates. For example, CMT surpasses MT and CPS by 1.19% and 1.08% on 5% labeled data, showing the superiority of collaborative mean-teacher against the current consistency learning framework. By introducing UMIX, UCMT consistently increases the performance under different labeled data rates, which implies that promoting model disagreement and guaranteeing high-confident pseudo labels are beneficial for semi-supervised segmentation.

Results on Colonoscopy. We further conduct a comparative experiment on the polyp segmentation task from colonoscopy images. Table 2 reports the quantitative results on both Kvasir-SEG and CVC-ClinicDB datasets. Compared with the adversarial learning-based [Hung et al., 2018; Wu et al., 2021a] and consistency learning-based [Tarvainen and Valpola, 2017; Ouali et al., 2020; Chen et al., 2021] al-
Table 3: Comparison with state-of-the-art methods on LA dataset. 10% $D_L$ and 20% $D_L$ of the labeled data are used for training.

<table>
<thead>
<tr>
<th>Method</th>
<th>10% $D_L$</th>
<th></th>
<th></th>
<th></th>
<th>20% $D_L$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC</td>
<td>Jaccard</td>
<td>95HD</td>
<td>ASD</td>
<td>DSC</td>
<td>Jaccard</td>
<td>95HD</td>
<td>ASD</td>
</tr>
<tr>
<td>UA-MT [Yu et al., 2019a]</td>
<td>84.25</td>
<td>73.48</td>
<td>13.84</td>
<td>3.36</td>
<td>88.88</td>
<td>80.21</td>
<td>7.32</td>
<td>2.26</td>
</tr>
<tr>
<td>SASSNet [Li et al., 2020a]</td>
<td>87.32</td>
<td>77.72</td>
<td>9.62</td>
<td>2.55</td>
<td>89.54</td>
<td>81.24</td>
<td>8.24</td>
<td>2.20</td>
</tr>
<tr>
<td>LG-ER-MT [Hang et al., 2020]</td>
<td>85.54</td>
<td>75.12</td>
<td>13.29</td>
<td>3.77</td>
<td>89.62</td>
<td>81.31</td>
<td>7.16</td>
<td>2.06</td>
</tr>
<tr>
<td>DUWM [Wang et al., 2020]</td>
<td>85.91</td>
<td>75.75</td>
<td>12.67</td>
<td>3.31</td>
<td>89.65</td>
<td>81.35</td>
<td>7.04</td>
<td>2.03</td>
</tr>
<tr>
<td>DTC [Luo et al., 2021]</td>
<td>86.57</td>
<td>76.55</td>
<td>14.47</td>
<td>3.74</td>
<td>89.42</td>
<td>80.98</td>
<td>7.32</td>
<td>2.10</td>
</tr>
<tr>
<td>MC-Net [Wu et al., 2021b]</td>
<td>87.71</td>
<td>78.31</td>
<td>9.36</td>
<td>2.18</td>
<td>90.34</td>
<td>82.48</td>
<td>6.00</td>
<td>1.77</td>
</tr>
<tr>
<td>MT [Tarvainen and Valpola, 2017]</td>
<td>86.15</td>
<td>76.16</td>
<td>11.37</td>
<td>3.60</td>
<td>89.81</td>
<td>81.85</td>
<td>6.08</td>
<td>1.96</td>
</tr>
<tr>
<td>CPS [Chen et al., 2021]</td>
<td>86.23</td>
<td>76.22</td>
<td>11.68</td>
<td>3.65</td>
<td>88.72</td>
<td>80.01</td>
<td>7.49</td>
<td>1.91</td>
</tr>
<tr>
<td>CMT (ours)</td>
<td>87.23</td>
<td>77.83</td>
<td>7.83</td>
<td>2.23</td>
<td>89.88</td>
<td>81.74</td>
<td>6.07</td>
<td>1.94</td>
</tr>
<tr>
<td>UCMT (ours)</td>
<td>88.13</td>
<td>79.18</td>
<td>9.14</td>
<td>3.06</td>
<td>90.41</td>
<td>82.54</td>
<td>6.31</td>
<td>1.70</td>
</tr>
</tbody>
</table>

Table 4: Ablation study of the different component combinations with only 5% labeled data on ISIC dataset. TS: teacher-student; SS: student-student; STS: student-teacher-student; $L_{cps}$: cross pseudo supervision; $L_{mts}$: mean-teacher supervision; U: UMIX.

<table>
<thead>
<tr>
<th>Method</th>
<th>TS</th>
<th>SS</th>
<th>STS</th>
<th>$L_{cps}$</th>
<th>$L_{mts}$</th>
<th>U</th>
<th>DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.31</td>
</tr>
<tr>
<td>MT</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>86.67</td>
</tr>
<tr>
<td>CPS</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>86.41</td>
</tr>
<tr>
<td>CMTv1</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>87.48</td>
</tr>
<tr>
<td>CMTv2</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>87.66</td>
</tr>
<tr>
<td>CMTv3</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>87.86</td>
</tr>
<tr>
<td>UCMT</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>U</td>
<td>88.22</td>
</tr>
</tbody>
</table>

4.3 Ablation Study

We conduct an ablation study in terms of network architectures, loss functions and region mix to investigate the effectiveness of each component and analyze the hyperparameters of the proposed method. There are three types of network architectures: 1) teacher-student (TS), 2) student-student (SS), and 3) student-teacher-student in the proposed CMT.

Effectiveness of Each Component. Table 4 reports the performance improvements over the baseline. It shows a trend that the segmentation performance improves when the components, including the STS (student-teacher-student), $L_{cps}$, $L_{mts}$, and UMIX are introduced into the baseline, and again confirms the necessity of encouraging model disagreement and enhancing the quality of pseudo labels for semi-supervised segmentation. The semi-supervised segmentation model is boosted for two reasons: 1) $L_{cps}$, $L_{mts}$ and the STS architecture that force the model disagreement in CMT for co-training, and 2) UMIX facilitating the model to produce high-confidence pseudo labels. All the components contribute to UCMT to achieve 88.22% DSC. These results demonstrate their effectiveness and complementarity for semi-supervised medical image segmentation. On the other hand, the two groups of comparisons between "TS (teacher-student) + $L_{mts}$" (i.e., MT) vs. STS + $L_{mts}$ (i.e., CMTv1), and between "SS (student-student) + $L_{cps}$" (i.e., CPS) vs. "STS + $L_{cps}$" (i.e., CMTv2) show that the STS-based approaches yield the improvements of 0.17% and 0.64%, indicating the effectiveness of the STS component. However, the performance gaps are not significant because the STS architecture increases the co-training disagreement but decreases the confidences of pseudo labels. It can be easily found that the results are improved to 87.86% by "STS + $L_{cps} + L_{mts}$" (i.e., CMTv3) and the relative improvements of 1.55% and 1.41% DSC have been obtained by "STS + $L_{cps} + L_{mts} + UMIX" (i.e., UCMT) compared with MT and CPS. The results demonstrate our hypothesis that maintaining co-training with high-confidence pseudo labels can improve the performance of semi-supervised learning.

Comparison of Different Data Augmentation. We fur-
Figure 3: Comparison of UMIX and CutMix on ISIC (a) and LA (b) dataset under 2%, 5%, 10%, and 20% $D_L$.

<table>
<thead>
<tr>
<th>$1/r$</th>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/16</td>
<td>1</td>
<td>87.95</td>
<td>88.22</td>
<td>88.12</td>
<td>87.96</td>
<td>88.08</td>
</tr>
<tr>
<td>1/4</td>
<td>2</td>
<td>87.65</td>
<td>88.15</td>
<td>87.80</td>
<td>87.54</td>
<td>87.90</td>
</tr>
<tr>
<td>1/8</td>
<td>3</td>
<td>87.86</td>
<td>87.92</td>
<td>88.03</td>
<td>88.01</td>
<td>87.87</td>
</tr>
</tbody>
</table>

Table 5: Investigation on how the top $k$ and patch size affect the capacity of UMIX on ISIC dataset with 5% labeled data.

ther compare the proposed UMIX, component of our UCMT, with CutMix [Yun et al., 2019] on ISIC and LA datasets with different labeled data to investigate their effects in semi-supervised segmentation. As illustrates in Figure 3, UMIX outperforms CutMix, especial in the low-data regime, i.e., 2% labeled data. The reason for this phenomenon is that CutMix performs random region mix that inevitably introduces noise into the new samples, which reduces the quality of the pseudo labels, while UMIX processes the image regions according to the uncertainty of the model, which facilitates the model to generate more confident pseudo labels.

Parameter Sensitivity Analysis. UMIX has two hyper-parameters, i.e., the top $k$ regions for mix and the size of the regions (patches) defined as the ratio of $1/r$ to the image size. We study the influence of these factors to UCMT on ISIC dataset with 5% $D_L$. It can be observed in Table 5 that reducing the patch size leads to a slight increase in performance. Moreover, varying the number of $k$ does not bring us any improvement, suggesting that we can choose any value of $K$ to eliminate outliers, thus bringing high-confidence pseudo labels for semi-supervised learning, which indicates the robustness of UMIX.

4.4 Qualitative Results

Figure 4 visualizes some example results of polyp segmentation, skin lesion segmentation, and left atrial segmentation. As shown in Figure 4 (a), the supervised baseline insufficiently segments some lesion regions, mainly due to the limited number of labeled data. Moreover, MT [Figure 4 (b)] and CPS [Figure 4 (c)] typically under-segment certain objects, which can be attributed to the limited generalization capability. On the contrary, our CMT [Figure 4 (e)] corrects these errors and produces smoother segment boundaries by gaining more effective supervision from unlabeled data. Besides, our complete method UCMT [Figure 4 (f)] further generates more accurate results by recovering finer segmentation details through more efficient training. These examples qualitatively verify the robustness of the proposed UCMT. In addition, to clearly give an insight into the procedure of the pseudo label generation and utilization in the co-training SSL method, we illustrate the uncertainty maps for two samples during the training in Figure 5. As shown, UCMT generates the uncertainty maps with high uncertainty [Figure 5 (a)/(c)] in the early training stage whereas our model produces relative higher confidence maps [Figure 5 (b)/(d)] from the UMIX images. During training, UCMT gradually improves the confidence for the input images. These results prove that UMIX can facilitate SSL models to generate high-confidence pseudo labels during training, guaranteeing that UCMT is able to maintain co-training in a more proper way.

5 Conclusion

We present an uncertainty-guided collaborative mean-teacher for semi-supervised medical image segmentation. Our main ideas lies in maintaining co-training with high-confidence pseudo labels to improve the capability of the SSL models to explore information from unlabeled data. Extensive experiments on four public datasets demonstrate the effectiveness of this idea and show that the proposed UCMT can achieve state-of-the-art performance. In the future, we will investigate more deeply the underlying mechanisms of co-training for more effective semi-supervised image segmentation.
Ethical Statement
There are no ethical issues.

Acknowledgments
This work was supported by the National Natural Science Foundation of China under Grant 62076059 and the Natural Science Foundation of Liaoning Province under Grant 2021-MS-105.

References


[Ouali et al., 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation

4206


