Self-Ensembling Co-Training Framework for Semi-Supervised COVID-19 CT Segmentation

Caizi Li, Li Dong, Qi Dou, Fan Lin, Kebao Zhang, Zuxin Feng, Weixin Si, *Member, IEEE*, Xuesong Deng, Zhe Deng, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*—The coronavirus disease 2019 (COVID-19) has become a severe worldwide health emergency and is spreading at a rapid rate. Segmentation of COVID lesions from computed tomography (CT) scans is of great importance for supervising disease progression and further clinical treatment. As labeling COVID-19 CT scans is labor-intensive and time-consuming, it is essential to develop a segmentation method based on limited labeled data to conduct this task. In this paper, we propose a self-ensembled co-training framework, which is trained by limited labeled data and large-scale unlabeled data, to automatically extract COVID lesions from CT scans. Specifically, to enrich the diversity of unsupervised information, we build a co-training framework consisting of two collaborative models, in which the two models teach each other during training by using their respective predicted pseudo-labels of unlabeled data. Moreover, to alleviate the adverse impacts of noisy pseudo-labels for each model, we propose a self-ensembling strategy to perform consistency regularization for the up-to-date predictions of unlabeled data, in which the predictions of unlabeled data are gradually ensembled via moving average at the end of every training epoch. We evaluate our framework on a COVID-19 dataset containing 103 CT scans. Experimental results show that our proposed method achieves better performance in the case of only 4 labeled CT scans compared to the state-of-the-art semi-supervised segmentation networks.**

***Index Terms*—COVID-19 CT segmentation, semi-supervised image segmentation, self-ensembling model, co-training.**

I. Introduction

The COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has become an ongoing pandemic and caused a lot of deaths. Chest CT has been regarded as an effective tool to screen the patients with clinical and epidemiologic features for COVID-19 infection [1], [2]. In the course of COVID-19 treatment, segmentation is an essential step that can provide the delineation and quantification of infection regions which is the COVID lesions caused by SARS-CoV-2, and the results of segmentation can be applied into disease progression evaluation and further assessment. However, manually delineating the COVID infection regions from CT scans is very challenging. First of all, delineating a CT scan needs to annotate the infection regions slice by slice, while a CT scan usually contains dozens to hundreds of slices, making manual annotation labor-intensive and time-consuming. According to the statistics of Ma *et al.* [3], it takes about 400 minutes to delineate one CT scan with 250 slices. Another issue is that lesion can vary greatly in size and appearance. In infection areas, the lesions may present as ground glass opacities (GGO) with the density increases or consolidation with the accumulation of fluid progresses on CT scans. Moreover, the boundaries of infection areas are usually blurry and cannot be distinguished clearly, which illustrated in Fig. 1. Considering the challenges of delineating COVID lesions, automatic segmentation method with limited labeled data is in urgent need for practical clinical application.

With the application of artificial intelligence for treatment of COVID-19 [4], deep learning-based techniques have attracted widespread attention in imaging-based analysis of COVID-19 [5]. Popular effective segmentation networks, such as U shape models [6], [7] and attention-based model [8], [9], have been evaluated for COVID-19 lesion segmentation task. Besides,

Manuscript received December 7, 2020; revised July 4, 2021; accepted August 5, 2021. Date of publication August 10, 2021; date of current version November 5, 2021. This work was supported in part by Shenzhen Fundamental Research Program under Grants JCYJ20200109110208764 and JCYJ20200109110420626, in part by the National Natural Science Foundation of China under Grants U1813204 and 61802385, and in part by the Natural Science Foundation of Guangdong under Grant 2021A1515012604. *(Caizi Li and Li Dong contribute equally to this work.) (Corresponding authors: Weixin Si; Xuesong Deng; Zhe Deng.)*

Caizi Li and Weixin Si are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: cz.li@siat.ac.cn; wx.si@siat.ac.cn).

Li Dong is with the Department of Radiology, Zhijiang People's Hospital, Hubei 443200, China (e-mail: 895664464@qq.com).

Qi Dou and Pheng-Ann Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: qidou@cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

Fan Lin is with the Department of Radiology, the First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, Guangdong 518035, China (e-mail: foxetfoxet@gmail.com).

Kebao Zhang and Zhe Deng are with the Department of Emergency Medicine, the First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, Guangdong 518035, China (e-mail: 243905580@qq.com; dengz163@163.com).

Zuxin Feng is with the Department of Emergency Medicine, Peking University ShenZhen Hospital, Shenzhen, Guangdong 518036, China (e-mail: 438926214@qq.com).

Xuesong Deng is with the Department of Hepatobiliary Surgery, the First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, Guangdong 518035, China (e-mail: dengxuesong2017@126.com).

Digital Object Identifier 10.1109/JBHI.2021.3103646

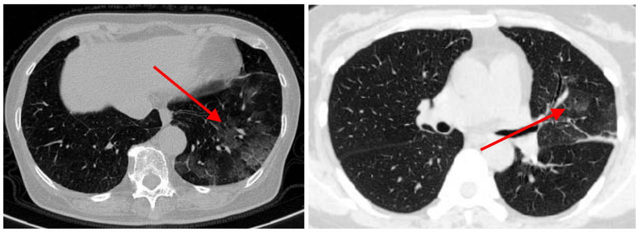

Fig. 1. Two cases of blurry COVID-19 lesions, showing the difficulties of labeling COVID-19 lesions from CT scans.

Zhou *et al.* [10] proposes a segmentation method for COVID-19 via decomposing 3D CT scans into 2D ones along three planes. Despite of these successful methods, training a robust segmentation model usually relies on sufficient labeled data. To mitigate the request of labeled data, several methods using a few labeled images or weak supervised data are proposed for COVID-19 lesion segmentation. For instance, Fan *et al.* [11] adopt a randomly selected propagation strategy which progressively generates pseudo-labels for unlabeled data to improve the performance of network. Laradji *et al.* [12] proposes a consistency based method by utilizing weak labeled data. However, without accurate manual labels, weak supervised methods often meet limitation in performance due to the variety of COVID-19 lesions. Besides, progressively generating pseudo-labels needs high computational cost due to multiple iterative training. Different from the previous works, in this paper, we focus on developing an end-to-end semi-supervised method for COVID-19 lesion segmentation from CT scans.

Recently, employing multiple networks have been widely used in semi-supervised settings, such as consistency based methods [13]–[15] and co-training methods [16]–[19]. Specifically, consistency-based methods usually follow the assumption that the same unlabeled inputs perturbed by different noises should be predicted into same outputs, thus a model can be trained with consistency regularization by given unlabeled data [20]. For example, TCSM proposed by Li *et al.* [13] performs transformation consistency to constrain the predictions of unlabeled data via mean teacher model [21]. However, the existed consistency-based medical segmentation methods are built with a single trainable model, which cannot provide multi views to enrich the unsupervised knowledge for unseen data. As for co-training methods, although it also has been widely used in semi-supervised learning, the existing works are not suitable for our COVID-19 lesion segmentation task. Since the networks are usually initialized differently in co-training framework, predictions from different networks inevitably suffer from unreliability which has negative impact on co-training. Thus, different strategies have been proposed to deal with this issue in the existing co-training works. For instance, in [18], the authors proposed a complementary correction network to minimize the prediction divergence between the mutual learning networks, however, the method in [18] is only designed for image classification task. Ke *et al.* [19] proposed a flaw detector to generate flaw probability map which can guide the networks to learn from unlabeled data collaboratively, however, even though it is designed for pixel-wise semi-supervised learning, the image processing pipeline in [19] estimating the flaw probability map, including dilation, blurring and normalization, is not suitable for our task. Accordingly, due to the characteristic of COVID lesion, the co-training framework needs to be further studied before being applied in our task.

To address the above bottlenecks and develop an effective semi-supervised segmentation method for COVID-19, we propose a novel framework called *Self-Ensembling Co-Training* to accurately delineate COVID-19 lesions. Particularly, our framework contains two peer segmentation models, which are built with the same architecture and initialized by different parameters. With different initialization, the two models can generate different decision boundaries, so that they have abilities to learn different knowledge from same unlabeled data. Next, with a certain unlabeled data as input, we use the two models to infer the predictions and take the two predictions as pseudo-labels to supervise each other mutually. In this way, the two models can be trained in a collaborative way by the combination of labeled data and pseudo-labeled data. Moreover, since the pseudo-labels from the peer model may not be accurate enough, in order to avoid the negative impacts of imperfect pseudo-labels, we propose a self-ensembled mechanism for the two collaborative models to constrain the learning progress of unlabeled data. Inspired by the temporal ensembling proposed by Laine *et al.* [22], for a certain training epoch, we gradually ensemble the predictions of unlabeled data from the first epoch to the last epoch by moving average, the ensembled predictions are then regarded as the consistency targets to be used to perform consistency regularization for the up-to-date predictions. To our best knowledge, this is the first time to build consistency regularization in co-training framework via self-ensembling strategy for semi-supervised COVID-19 CT segmentation. We summarize the contributions of this paper as follows:

1) We present a co-training framework for semi-supervised COVID-19 CT segmentation, in which two models teach each other during training by their respective predicted pseudo-labels of unlabeled data.
2) We propose a self-ensembling consistency regularization to reduce the negative impacts of pseudo-labeled data during co-training progress. The consistency regularization can be directly integrated into our co-training framework and significantly improves the efficiency of utilizing unlabeled data.
3) We evaluate our proposed framework on a COVID-19 dataset which contains 103 CT scans. Experimental results demonstrate the considerable effectiveness of our method to reduce the requirement of pixel-wise labeled COVID-19 data.

II. Related Works

In this section, we only introduce the areas highly relevant to our work. We first present an brief overview for applications of deep learning approaches for COVID-19 lesion segmentation, and then review the development of semi-supervised medical image segmentation methods.

A. Deep Learning for COVID-19 CT Segmentation

Since delineating a CT scan is a labor-intensive and time-consuming job, deep learning methods have attracted people's attention for fast and automatic inferring segmentation results. Popular segmentation networks such as U-Net [23]–[25], U-Net++ [26], [27] and VB-Net [28] are first applied to analyze COVID-19 at the early stage of COVID-19 outbreak. With the release of several small scale labeled CT scans [3], [29], several segmentation methods are proposed to segment COVID-19 lesions. Fan *et al.* [11] propose a model called Inf-Net which uses aggregated high-level features and attention mechanism to extract COVID-19 infection areas, they also adopt a randomly selected propagation strategy to perform SSL to alleviate the shortage of labeled data. Because it is hard to obtain large scale well labeled dataset and noisy labeled dataset is easier to obtain, Wang *et al.* [30] propose a robust framework to against label noise for COVID-19 lesion segmentation. Since the difficulties of labeling CT scans, segmenting COVID-19 lesions from weak labeled data has drawn a lot of attentions [12], [31], [32], besides, active learning combined with weakly supervised learning is proposed to conduct COVID-19 segmentation in [31]. However, the performance of current data-efficient methods toward COVID-19 lesion segmentation still exists big gap compared with fully supervised methods.

B. Semi-Supervised Segmentation for Medical Images

Semi-supervised segmentation methods for medical image can be roughly divided into four categories: self-training [33], GAN based methods [34], [35], consistency based methods [13]–[15], [22], [36], [37], co-training methods [16], [17], [19]. In self-training such as [33], model first predicts the pseudo-labels for unlabeled data to extend training dataset, then the model is trained by the extended training dataset, this process will be iterated many times until the performance improvement becomes negligible. GAN based methods usually build a framework which contains a segmentation network and a discrimination network, among which the discrimination network is used to distinguish the quality of predictions and the segmentation network tries to predict accurate results to fool discrimination network. Consistency based strategy is widely used in medical image segmentation. The common practice follows the assumption that the inputs under different perturbations would be predicted the same result. For instance, in [13], based on mean teacher model [21], the authors apply perturbations like Gaussian noise, randomly rotation and scaling to the inputs and the outputs, then encourage the network to be transformation consistent for unlabeled data. Besides, Yu *et al.* [14] propose an uncertainty estimation strategy to improve performance of consistency based model by learning meaningful and reliable targets during training. Co-training methods also draw a lot of attention and show promising results for semi-supervised medical image segmentation. In [17], Xia *et al.* introduce a co-training method which adopt uncertainty estimation strategy to improve the performance of network. In this paper, we will investigate the performance gap of the different categories of semi-supervised methods for COVID-19 and further propose a more effective method to utilize unlabeled data.

III. Methods

An overview of our proposed framework is shown in Fig. 2. Our self-ensembling co-training framework contains two mutual learning models. Three training objectives including supervised loss, pseudo-supervised loss and self-ensembling consistency regularization are deployed for each model. In this section, we first introduce the co-training architecture that the two models mutually teach each other. Then we describe the details of consistency regularization for co-training framework and the way to construct the consistency targets by self-ensembling which is used to constrain the co-training process during training. Finally, we define the overall training objective of our framework.

We formulate the problem of our task as follows. Given a dataset $\mathcal{D}$ contains a labeled dataset with N CT scans denoted as $\mathcal{D}_L = \{(x_i^l, y_i)\}_{i=1}^N$ and a unlabeled dataset with M CT scans denoted as $\mathcal{D}_U = \{x_i^u\}_{i=1}^M$, where M $\gg$ N, x_i^l and x_i^u denote CT scans and y_i is the corresponding ground truth of labeled data, we aim at building a data-efficient deep learning model which is trained over the combination of $\mathcal{D}_L$ and $\mathcal{D}_U$ in a semi-supervised manner and aim to make the performance to be comparable to an optimal model trained over fully labeled $\mathcal{D}$ as much as possible.

A. Co-Training for Semi-Supervised Segmentation

Due to the unique feature of COVID-19 lesions, even experienced clinical experts may have different understandings towards a same CT scan [32], thus there must be two or more experts to participate in manually labeling the lesions for training an accurate data-driven model. Views from different experts can be regarded as the complementary knowledge for each other to rectify the labeling errors. Inspired by this mutual label process, we intend to solve the semi-supervised problem with a co-training framework, in which two independent segmentation models are trained by labeled data and mutually teach each other the knowledge which is learned from unlabeled data.

Specifically, in our co-training framework, the two independent models are denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$, as is shown in Fig. 2. We denote the outputs of models as $f^k(x_i; \theta^k)$, where $k \in \{1, 2\}$ indicates the index of models and x_i which including x_i^l and x_i^u denotes the input data. At the beginning of training, the model weights θ^k are initialized randomly and independently to guarantee the diversity of two models. Since the two models have different initialization, the representation for a same input should be different so that we can regard the two models as different views. In our work, the different views are utilized as complementary knowledge to enrich the representation of the whole framework towards the same inputs. In practice, we utilize the labeled data to perform supervised learning to make sure the two models can be trained normally. Meanwhile, given a certain unlabeled input, we utilize the predictions of two models to supervise each other, so that the two models can learn the complementary knowledge from each other. Consequently, the training objective of each model contains two parts: supervised loss and pseudo-supervised loss.

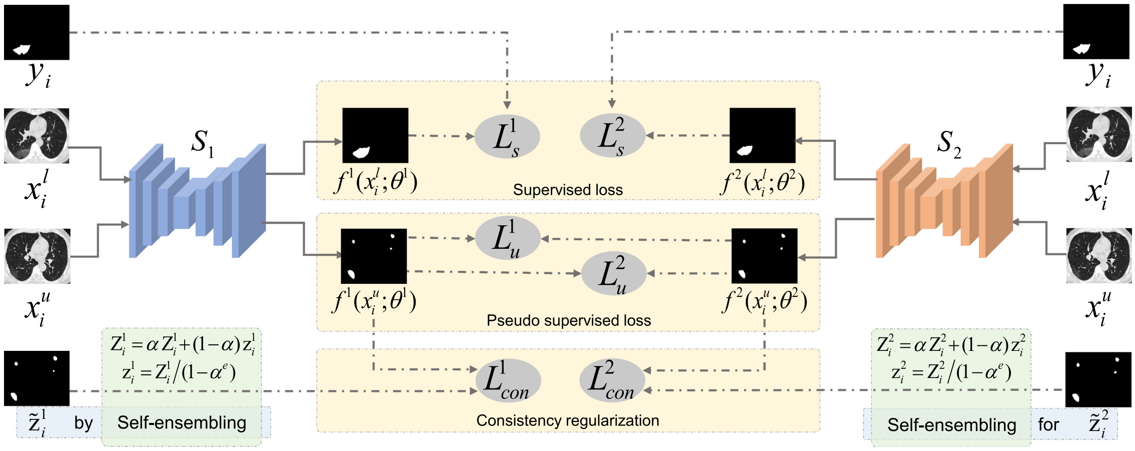
y_i
y_i
x_i^l
x_i^u
x_i^l
x_i^u
S_1
S_2
L_s^1
L_s^2
$f^1(x_i^l;\theta^1)$
Supervised loss
$f^2(x_i^l;\theta^2)$
L_u^1
L_u^2
$f^1(x_i^u;\theta^1)$
Pseudo supervised loss
$f^2(x_i^u;\theta^2)$
$Z_i^1=\alpha Z_i^1+(1-\alpha)z_i^1$
$z_i^1=Z_i^1/(1-\alpha^e)$
L_{con}^1
L_{con}^2
$Z_i^2=\alpha Z_i^2+(1-\alpha)z_i^2$
$z_i^2=Z_i^2/(1-\alpha^e)$
$\tilde{z}_i^1$ by Self-ensembling
Consistency regularization
Self-ensembling for $\tilde{z}_i^2$

Fig. 2. Illustration of our proposed framework for semi-supervised COVID-19 CT segmentation. The framework contains two collaborative training models $\mathcal{S}_1$ and $\mathcal{S}_2$ which take the CT slices as inputs. Each model is trained by three loss functions, including supervised loss L_s^k, pseudo-supervised loss L_u^k and consistency regularization L_{con}^k. In our implements, L_s^k and L_u^k share the same equation, the difference between them is that L_u^k takes pseudo-labels predicted by the peer model in the framework as inputs. The regularization item L_{con}^k is used to alleviate the adverse impacts of noisy pseudo-labels generated from the peer model. To construct consistency targets $\tilde{z}_i^k$ of each model, we gradually ensemble the predictions of unlabeled data from the first epoch to the last epoch through moving average, which is called self-ensembling. The total loss for each model is the weighted sum of the three loss functions.

During training, each model takes a batch of labeled data $\{(x_i^l, y_i)\}_{i=1}^B \in \mathcal{D}_L$ and a batch of unlabeled data $\{(x_i^u)\}_{i=1}^B \in \mathcal{D}_U$ as inputs, where B indicates the training batch size. We denote the two losses as L_s^k and L_u^k for each model, respectively. Dice coefficient is employed to measure the loss between model outputs and the ground truths, which is proposed in [38]. The supervised loss function L_s^k is formulated as follows,

$$L_s^k(f^k(x_i^l;\theta^k), y_i) = 1 - \frac{2 * \sum f^k(x_i^l;\theta^k) * y_i}{\sum f^k(x_i^l;\theta^k) + \sum y_i}. \tag{1}$$

In our COVID-19 lesion segmentation task, since the foreground regions usually occupy a small part of the whole image, there exists inevitable imbalance between the foreground and background regions. This often leads to the learning process trapped in local minima of the loss function, thus the predictions of a network are strongly biased towards the background, making the foreground region missing or partially detected [38]. In this case, balancing the foreground and background regions during training is of great importance for our task. Since the Dice coefficient can be regarded as the ratio of intersection and union between two sets, the balance between foreground and background can be achieved [38]. Therefore, to mitigate the negative impact of imbalance between COVID-19 lesions and the background, we also adopt Dice loss as the pseudo-supervised loss. Accordingly, the form of L_u^k is the same as L_s^k except the input data is changed into x_i^u. Consequently, the optimization objective of one model in our co-training framework is summarized as follows,

$$\begin{aligned} L^k = &\sum_{i=1}^{N} L_s^k(f^k(x_i^l;\theta^k), y_i) \\ &+ \lambda \sum_{i=1}^{M} L_u^k(f^k(x_i^u;\theta^k), \tilde{y}_i^k), \end{aligned} \tag{2}$$

where k is the index of two models, $\tilde{y}_i^k$ denotes the pseudo-labels of unlabeled data. Considering that the predictions of unlabeled data may be unreliable, we weight the supervised loss and the pseudo-supervised loss to avoid the training process dominated by pseudo-supervised learning, which is indicated by the trade-off coefficient λ. Note that the loss function $L_s^k(\cdot)$ and $L_u^k(\cdot)$ share the same equation except the different input data.

B. Consistency Regularization for Co-Training

In SSL, the performance improvement benefits from learning unsupervised knowledge from unlabeled data, thus the quality and reliability of knowledge learned from unlabeled data play an important role in semi-supervised framework. However, in co-training framework, the pseudo-labels of each model for unlabeled data directly come from the up-to-date peer model without extra guidance, these pseudo-labels are expected to be noisy, leading to limitations of performance. To alleviate the adverse impacts of noisy pseudo-labels, we intend to utilize consistency regularization to constrain the learning progress of each model in co-training. Specifically, for each model, we construct consistency targets for the unlabeled data and encourage the model to generate consistent predictions between the raw unlabeled data and the corresponding consistency targets, as is illustrated in Fig. 2.

Formally, given an unlabeled data $x_i^u \in \mathcal{D}_U$, the inferred prediction of x_i^u can be denoted as $f^k(x_i;\theta^k)$. We denote the consistency targets as $\tilde{z}_i^k$, following the rule of consistency regularization, the distance between $f^k(x_i^u;\theta^k)$ and $\tilde{z}_i^k$ should be as small as possible, thus the training objective of consistency regularization can be described as follows,

$$L_{con}^k(f^k(x_i^u;\theta^k), \tilde{z}_i^k) = \left\| f^k(x_i^u;\theta^k) - \tilde{z}_i^k \right\|^2. \tag{3}$$

In consistency regularization, a key issue is to create a proper consistency targets. Current works [13], [14] realize it via applying different perturbations on the inputs such as Gaussian noise, rotation and scaling. However, as the appearance and size of COVID-19 vary greatly, inappropriate perturbations may degrade the model performance, which will be demonstrated by our experiments in Section IV. In this paper, we propose to construct the consistency targets by ensembling the predictions from previous training epochs, which called self-ensembling.

C. Self-Ensembling for Consistency Targets

The self-ensembling strategy to build consistency target in our framework is computed via exponential moving average (EMA), which is an extension of temporal ensembling model developed for image classification [22]. To compute the consistency target, we first define three variables: accumulated target Z_i^k, up-to-date prediction z_i^k and bias corrected target $\tilde{z}_i^k$. The Z_i^k is initialized as zero matrix. After each training epoch, we update Z_i^k via EMA which is the weighted sum of the up-to-date prediction of unlabeled data z_i^k and the Z_i^k from the previous epoch. To reflect the relationship between the update process and training epoch, we reformulate the Eq. (4) as follows,

$$Z_i^k[e] = \alpha Z_i^k[e-1] + (1-\alpha) z_i^k[e], \tag{4}$$

where e denotes the training epoch and α is a momentum term that controls how far the ensemble reaches into accumulated target.

Considering the accumulated target Z_i^k is initialized as zero matrix, we can observe from Eq. (4) that the value of Z_i^k is smaller than true value at early training stages, which is not suitable to be used to perform consistency regularization. Therefore, similar with the bias correction introduced in [22], the Z_i^k is magnified to approximate the true value via multiplying a factor $1/(1-\alpha^e)$ after performing Eq. (4), the magnified Z_i^k is denoted as the bias corrected target $\tilde{z}_i^k$ which is the final consistency target to be used in Eq. (3) for each epoch e. The formulation is presented as follows,

$$\tilde{z}_i^k[e] = Z_i^k[e]/(1-\alpha^e), \tag{5}$$

According to Eq. (5), the value of the factor decreases dynamically as the epoch e increases until it approaches the constant 1. Thus, the impact of bias correction dynamically decreases as training process continues so that the consistency regularization is performed correctly.

D. Overall Training Objective and Implementation Details

The overall training objective of our framework is the weighted sum of supervised loss, pseudo-supervised loss and consistency regularization item, it can be summarized as follows,

$$\begin{aligned} L_{total}^k = & \sum_{i=1}^{N} L_s^k(f^k(x_i^l;\theta^k), y_i) \\ & + \lambda \sum_{i=1}^{M} L_u^k(f^k(x_i^u;\theta^k), \tilde{y}_i^k) + \mu \sum_{i=1}^{M} L_{con}^k(f^k(x_i^u;\theta^k), \tilde{z}_i^k) \end{aligned} \tag{6}$$

where μ is a trade-off coefficient like λ. Since no predictions at epoch 1, the consistency regularization will join in training from the second epoch. To reduce the negative impact of unreliable predictions of unlabeled data at early training stage, we gradually increase the values of λ and μ from 0 to their maximum values λ_{max} and μ_{max} within e_{max} epochs by multiply their maximum values by a ramp-up weight

$$w(e) = \mathrm{e}^{(-5.0*(1-e/e_{\max})^2)}. \tag{7}$$

We employ 2D U-Net [6] which is commonly used in medical image segmentation as our backbone with CT slices as its inputs. Kaiming Initialization [42] is adopted to initialize models in our framework. The kernel size is set to 3 for all convolution layers except the last one. The number of feature maps in convolution layers starts from 32 and is doubled after each maxpooling, a total of four maxpooling are employed in the backbone. Each convolution layer is followed by an Instance Normalization layer [40] and a LeakyReLU function [41]. Sigmoid function is employed as activation layer for the last convolution layer. We implement our framework with PyTorch library [39] using an NVIDIA RTX 2080Ti. Two models of our framework are trained from scratch with Adam optimizer and share the same hyper-parameters. The learning rate and batch size are set to 1e-4 and 12 during training, we totally train 40 epochs with 250 iterations per epoch. The maximum value of trade-off coefficients λ_{max}, μ_{max} and maximum ramp-up value e_{max} in training objective are set to 0.1, 0.1 and 20. The momentum term α for self-ensembling is fixed as 0.6. Random data augmentation such as scaling with a factor of (0.85, 1.25), rotation, flipping are adopted during training.

IV. Experiments

A. Datasets and Evaluation Metrics

The dataset we adopt in this paper is an integration of three different COVID-19 datasets: self-collected dataset and another two online datasets (UESTC-COVID-19 Dataset [30], COVID-19 CT Lung and Infection Segmentation Dataset [43]). Our self-collected dataset contains 33 CT scans collected from three different hospitals: Shenzhen 2nd People's Hospital(10 CT scans), Peking University Shenzhen Hospital (6 CT scans) and Zhijiang People's Hospital (17 CT scans). In our self-collected dataset, each CT scan is collected from a different patient infected by SARS-CoV-2, and is annotated and confirmed by two experts under fixed window level -450HU and window width 1000HU. Another two datasets contain 50 and 20 CT scans, respectively. CT scans in [30] have been cropped based on the bounding box of the lung region and the intensity has been normalized into [0,1] using window width/level of 1500/-650. In [43], 10 of the CT scans have been already normalized into [0255]. Consequently, there are in total of 103 CT scans used for our semi-supervised segmentation task.

Because these CT scans come from different hospitals, there inevitably exist data discrepancy between different data sources. Considering we are not targeting data discrepancy problem, clipping and normalization for HU values of CT are adopted to mitigate the adverse impacts of data discrepancy. Specifically,

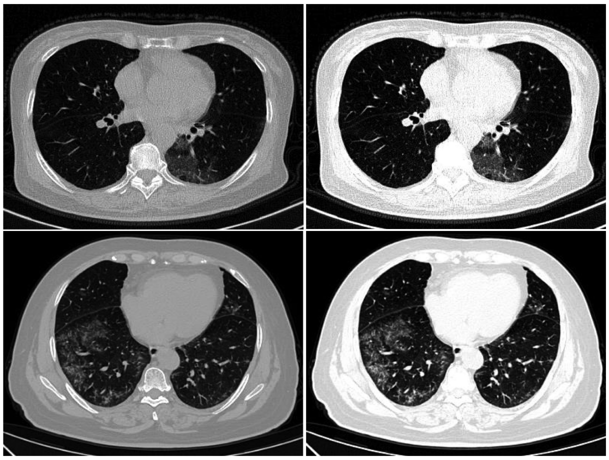

Fig. 3. The results before (left column) and after (right column) preprocessing of two cases from self-collected dataset (first row) and [43] (second row) are presented.

the HU values of all CT scans in self-collected dataset and 10 CT scans from [43] which have not been normalized are clipped into range (-950, 50) according to the window level and window width (-450HU, 1000HU), then normalized into $[0., 1.]$ by the formulation of $(x - x_{\min})\ (x_{\max} - x_{\min})$, where x denotes HU value of CT, $x_{\max}$ and $x_{\min}$ represent the upper and low bound which are -950 and 50 respectively. HU values of normalized CT scans in [43] are scaled into $[0., 1.]$ through dividing by 255. As CT scans in [30] has already been normalized into $[0., 1.]$, we just use it without any further normalization. We present two examples belong to different data sources before and after data preprocessing in Fig. 3, showing that the data discrepancy has been mitigated as much as possible. During training, we randomly select and crop a batch patches of size (256256) out from CT scans as the inputs of models.

Similar with [30], we also randomly split the whole dataset into three independent parts: training set, validation set and testing set, which contains 75 CT scans, 8 CT scans and 20 CT scans, respectively. We use five evaluation metrics to measure the segmentation performance for all experiments of our work, including Dice coefficient (DI), Jaccard index (JA), Sensitivity (SEN), Specificity (SP), AUC and the 95-th percentile of Hausdoff Distance (HD95). All results reported in tables are calculated based on the testing set. The means and variances are computed by averaging the metrics over the samples of testing set on volume level. Note that all experiments are based on patient level so that there is no samples from same patients for training and testing simultaneously.

B. Comparison With Other Semi-Supervised Methods

We implement the current state-of-the-art semi-supervised medical image segmentation methods, including GAN based methods DAN [35] and ASDNet [34], consistency based methods (MT) [21] and TCSM [13], self-training method [33] and uncertainty aware method (UA-MT) [14], to compared with our method. For GAN based methods, the trade-off coefficients for adversarial loss in DAN and ASDNet are set to $1e-3$. In particular, we start SSL in ASDNet from 15th epoch for reliable training and the threshold of confidence is set to 0.3. For consistency based methods, we set the decay of exponential moving average as 0.99, the metric of consistency between student and teacher model is set to Mean Square Error (MSE), the weight of consistency regularization is ramped up from 0 to 0.1 in 20 epochs following sigmoid function. Transformations such as Gaussian noise, randomly rotation and scaling are adopted in the implementation of TCSM. In self-training, we train 40 and 20 epochs for the initialization and SSL and perform alternate optimization for 3 iterations. The implementation of UA-MT follows the settings of the authors. Note that since the Noise-Robust [30] is developed for noisy labeled data, we replace the noisy labeled data with unlabeled data to perform semi-supervised learning. In addition, since the semi-supervised learning in Self-correcting [48] is based on weak-labeled dataset which the bounding box of foreground has been provided, we discard this strong prior in our implementation for fair comparison. The backbones and the training protocols of all implementations including our method are same with each other to ensure the fair comparison. Performance of different approaches are listed in Table I.

We train U-Net with 100% and 5% of training data which contain 75 and 4 CT scans respectively as upper bound and the baseline. We can observed from Table. I that the upper bound of performance can only exceed the baseline by Dice score of 8%. We argue that the reason can be summarized as follows. First, due to the blurry boundaries of COVID-19 lesions, it is challenging to distinguish the target accurately, therefore, the upper bound of performance is limited. In addition, albeit the CT scans are collected from different data source, in the process of data preprocessing described in the second paragraph of Section IV.A, we clip and normalize all raw CT scans with uniform window level and window width which are suggested by clinical experts to mitigate the data discrepancy. In this case, we can train a fairly satisfactory baseline model with only 5% labeled data.

The GAN based methods all perform better compared to the baseline, showing effectiveness of GAN based methods for semi-supervised segmentation. However, we can notice that the performance of ASDNet is lower than DAN, indicating that the application of confidence map is still challenging for our task. The consistency based methods TCSM and MT achieve comparable performance compared to GAN based methods, demonstrating the effectiveness of consistency regularization for effectively utilizing unlabeled data. Meanwhile, we can observe that TCSM which incorporates transformation consistency into MT does not show advantages over MT, revealing that the transformation consistency does not have a positive effect on our task. In addition, similar with ASDNet, uncertainty strategy implemented by Monte Carlo sampling in UA-MT does not achieve higher performance compared with MT, showing that the application of uncertainty strategy on our task remains challenging. Self-training obtains 74.27% of Dice score, which is higher than the baseline and the above semi-supervised methods, highlighting the superiority of self-training for semi-supervised segmentation task. It is worth noting that the performance of

TABLE I
COMPARISON WITH OTHER SEMI-SUPERVISED METHODS

Method	L/U[2]	Metrics						P-Value
		DI[%]	JA[%]	SEN[%]	SP[%]	AUC[%]	HD95(mm)	
Upper bound [6]	75/0	80.20±7.16	67.54±9.89	85.00±9.87	99.73±0.28	92.37±4.88	48.28±73.37	-
Baseline [6]	4/0	72.04±10.55	57.35±12.75	79.02±16.39	99.67±0.33	89.35±8.12	68.34±62.71	9e-4
Baseline-h[1] [6]	4/0	73.16±11.94	59.00±14.00	76.87±14.71	99.70±0.34	88.29±7.28	47.76±53.90	0.03
DAN [35]	4/71	73.11±9.65	58.51±11.84	79.02±10.92	99.62±0.4	89.32±5.35	63.86±59.38	0.031
ASDNet [34]	4/71	72.63±11.44	58.26±13.87	77.12±14.72	99.70±0.3	88.41±7.29	55.33±60.91	0.006
TCSM [13]	4/71	72.56±10.09	57.88±12.04	77.54±12.89	99.68±0.35	88.61±6.35	65.68±60.42	1e-4
MT [21]	4/71	72.94±11.06	58.57±13.39	78.30±12.58	99.66±0.39	88.98±6.19	56.34±60.79	0.005
UA-MT [14]	4/71	72.65±10.36	58.05±12.41	77.09±13.25	99.68±0.35	88.38±6.52	61.30±62.27	5e-5
Self-training [33]	4/71	74.27±8.28	59.75±10.28	77.71±10.85	99.67±0.36	88.69±5.34	69.18±64.15	0.039
SemiInfNet [11]	4/71	73.32±10.24	58.89±12.44	77.76±12.65	99.70±0.30	88.73±6.25	45.41±52.85	0.004
Noise-Robust [30]	4/71	70.17±11.67	55.28±13.79	74.82±15.67	99.66±0.35	87.24±7.76	71.44±64.30	1e-4
Self-paced [46]	4/71	72.37±8.97	57.47±10.84	74.02±14.28	**99.76±0.24**	86.89±7.08	**44.22±50.87**	0.002
CCT [47]	4/71	70.55±13.05	55.99±14.83	76.50±16.04	99.63±0.42	88.07±7.91	69.02±70.23	0.004
Self-Correcting [48]	4/71	71.18±11.15	56.33±12.51	72.15±15.70	99.74±0.25	85.95±7.81	55.02±55.61	0.020
GCT [19]	4/71	63.85±14.05	48.41±14.72	73.76±19.30	99.55±0.49	86.65±9.50	86.34±55.98	0.002
Ours	4/71	**75.92±8.95**	**62.01±11.42**	**79.38±11.68**	99.75±0.28	**89.57±5.77**	53.00±66.16	-

[1] '-h' means the number of convolution layers is doubled.
[2] 'L/U' means the number of Labeled/Unlabeled CT scans.

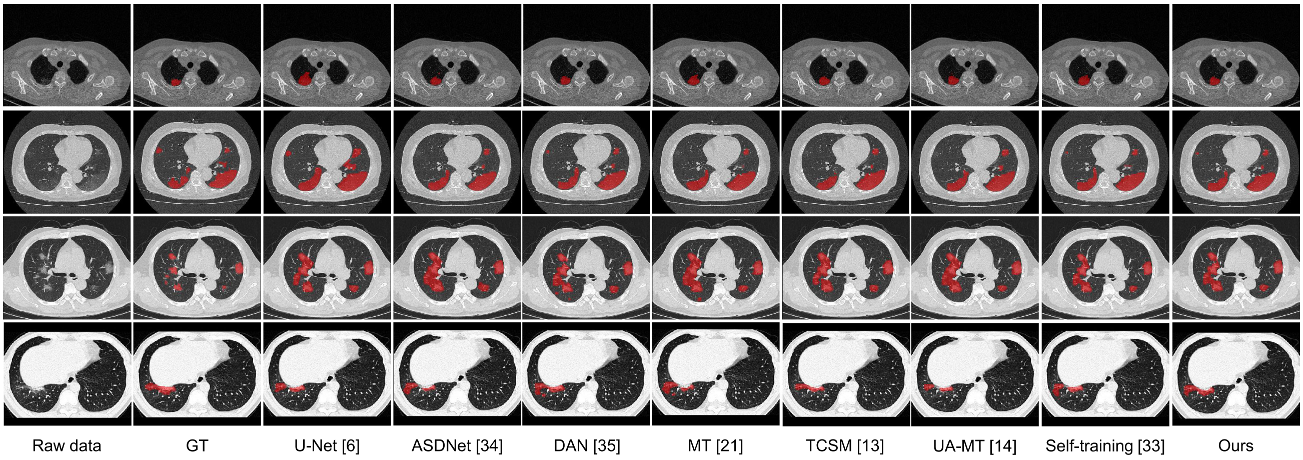

Fig. 4. Visualization of results predicted by different semi-supervised methods. The case of every row comes from different raw CT scans. It can be observed that our method can predict better than the other methods for the blurry lesions.

Noise-Robust, CCT, Self-Correcting and GCT is worse than the Baseline. For Noise-Robust, it can be demonstrated that the proposed noise-robust Dice loss is not suitable for our semi-supervised task. In CCT, the authors proposed to use several decoders to perform consistency regularization via taking the perturbed data as inputs. Considering that it is designed for natural images, the multiple decoders in CCT is inappropriate for our task, leading to poor performance of CCT. For Self-Correcting, due to without strong prior provided by the weak labels, the self-correcting network may have negative impact on our task. In GCT, the performance of framework degrades sharply, demonstrating that the proposed flaw detector cannot provide a satisfactory guidance with an unsuitable image processing pipeline for our task. As for [46], the authors proposed a self-paced strategy with an uncertainty regularizer to force the networks focus on the targets from easy to hard, similar with UA-MT, the uncertainty strategy remains unsatisfactory for our task. While the performance of SemiInfNet is improved compared with the baseline, demonstrating the effectiveness of iterative training, which is similar with Self-training. Different from other methods, our framework achieves the state-of-the-art performance, showing the effectiveness of our semi-supervised method.

To analyze the significance of the improvements between our method and the other semi-supervised methods, we take Dice scores of testing samples as input to perform statistical test via paired t-test. From the results, we can observe that all the p-values are smaller than 0.05, indicating the statistically significant improvements of our framework over other semi-supervised methods. Visualization of results predicted by different methods are shown in Fig. 4. We can see that methods often perform well for the lesions which are easy to distinguish from the

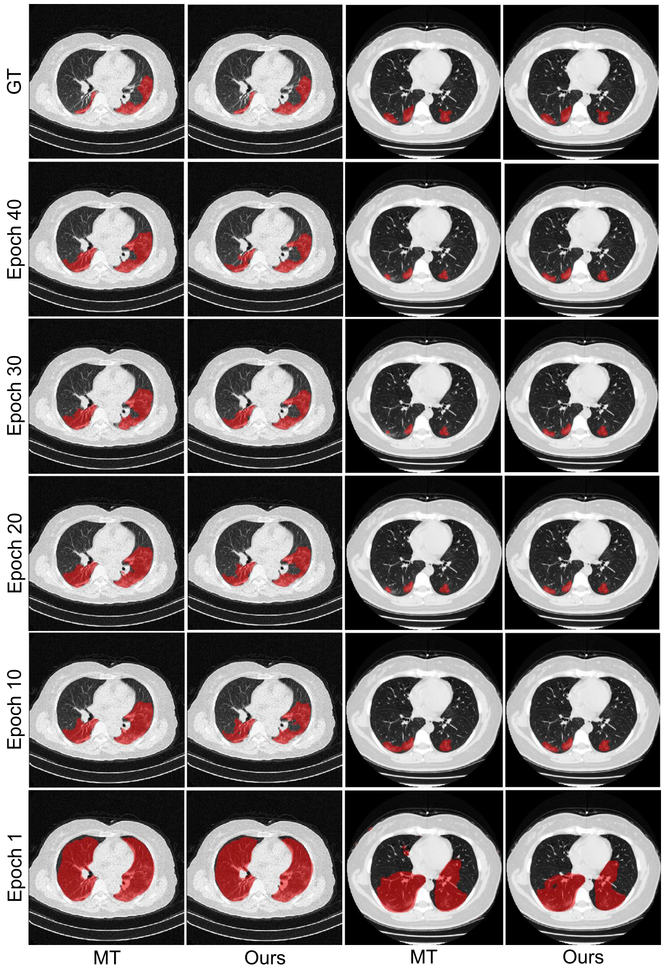

Fig. 5. Comparison of ensembled target between MT and our method.

background. However, for the blurry lesions, our method can predict better than the other methods. In addition, to show the advantage of our self-ensembled target, we take the predictions of the teacher model in MT framework as the comparative item. The visualization of predictions in different epochs is presented in Fig. 5 for the ensembled targets of MT and our method. It can be observed from the predictions of different epochs that the quality of our ensembled target is better than the ensembled target of MT especially in the mid to late training stages, demonstrating the superiority of our self-ensembling strategy.

C. Performance Using Different Percentages of Labeled Data

We study the impact of different percentages of labeled data on performance for four methods including our method, MT(the representative exponential moving average model), Self-training(the iterative-training method) and the Baseline. Except for the setting of 4 labeled CT scans in the previous section, we add another five settings of 8, 15, 23, 30 and 38 CT scans which are 10%, 20%, 30%, 40% and 50% of labeled training data. The results are presented in Fig. 6.

From the figure, it can be observed that our method outperforms other methods consistently with different percentages of labeled data, demonstrating the superiority of our method. Compared with the baseline, our method obtains a large improvement of for 3.88% of Dice score by using only 5% labeled CT scans, furthermore, our method can achieve comparable performance with only a small gap of 0.43% with 5% labeled CT scans compared with the baseline with 20% labeled CT scans, demonstrating the significant advantage of our method under small-scale labeled dataset. Compared with the baseline trained with 100% labeled CT scans reported in Table. I, our method achieves comparable performance with 0.14% gap by using only 30% labeled data, indicating that our method can significantly reduce the need of labeled data. It also can be noticed that the performance of all method increases slowly with the increase of labeled data, which illustrates that the performance of models tend to converge with labeled data increases.

D. Ablation Study

Our framework contains two main components: co-training and self-ensembling consistency regularization. To investigate the effectiveness of each component, we perform an ablation study by adding the two components on the baseline one by one. The experiments are conducted on the setting of 4 labeled CT scans. Results of different settings are presented in Table II. Without consistency regularization, the performance of the plain co-training can achieve 74.50% of Dice score, surpassing the baseline for 2.46%, the other measurements except sensitivity are all better than the baseline, showing the effectiveness of co-training framework. Without co-training strategy, the framework degenerates into a single view model. From the results, we can observe that the self-ensembling strategy can also promote the performance of our baseline by 1.48%. However, only training with a single view limits the effectiveness of utilizing unlabeled data. Finally, with the joint learning of co-training and self-ensembling consistency regularization, the performance of our framework is further promoted to the state-of-the-art. We also conduct paired t-test to verify the significance of the improvements between our method and the ablation studies under different settings. Results show that the p-values are all smaller than 0.05, demonstrating the significant improvements of our method.

E. Analysis of Loss Functions

1) Comparison Between Cross-Entropy and Dice Loss: In order to investigate the effect of Dice loss for our task, we conduct experiments to compare the performance between the well-known Cross-Entropy (CE) loss and Dice loss. The results are listed in Table III. Baseline(Dice) and Ours(Dice) represent the Baseline and the proposed method, respectively. Baseline(CE) and Ours(CE) are the comparison items whose supervised loss and pseudo-supervised loss are replaced with Cross-Entropy loss. It can be observed that both Baseline and our method suffer from performance degradation by replacing the Dice loss with CE loss. For CE loss, the foreground and background are treated equally during training, leading to the

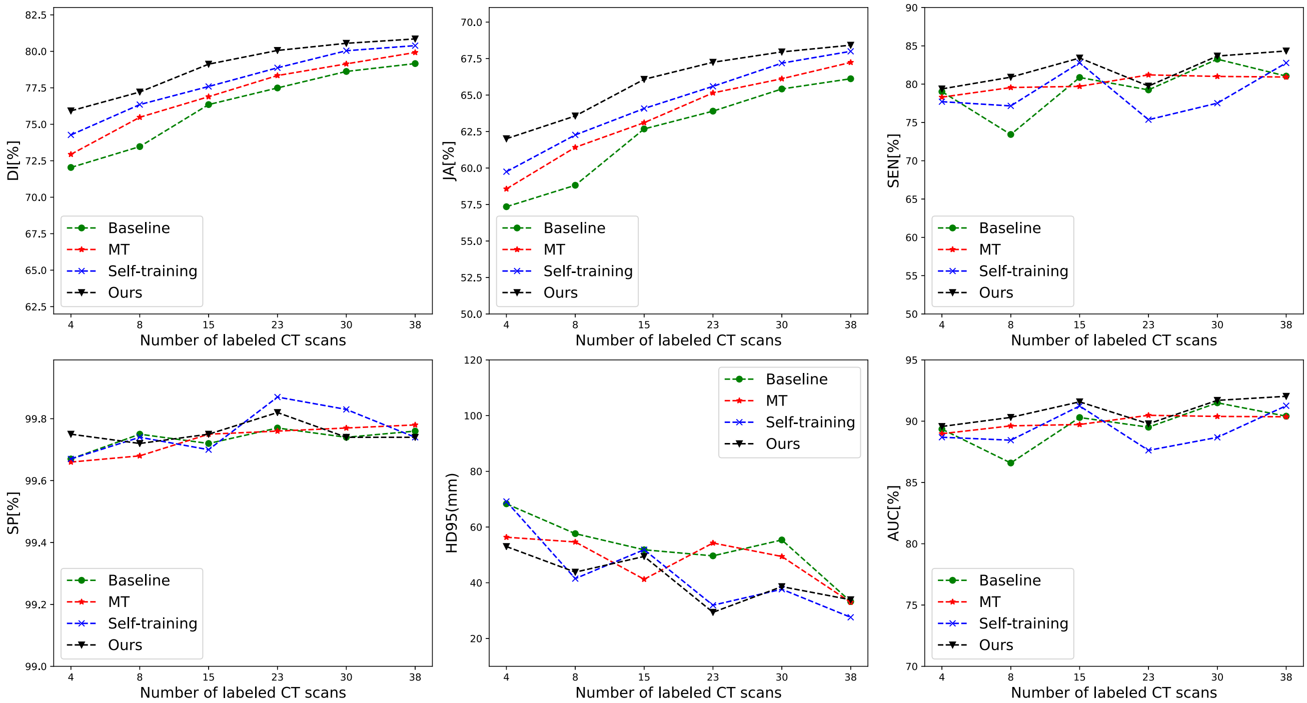

Fig. 6. Comparison between multiple methods using different percentages of labeled CT scans.

TABLE II
DIFFERENT ABLATION SETTINGS OF OUR METHOD TRAINED WITH 4 CT SCANS

Co-training	Consistency Regularization	Metrics						P-Value
		DI[%]	JA[%]	SEN[%]	SP[%]	AUC[%]	HD95(mm)	
✗	✗	72.04±10.55	57.35±12.75	79.02±16.39	99.67±0.33	89.35±8.12	68.34±62.71	9e-4
✓	✗	74.50±10.11	60.36±12.42	78.70±11.96	99.71±0.31	89.20±5.91	62.69±61.96	0.027
✗	✓	73.52±10.95	59.28±13.36	**80.76±13.30**	99.68±0.33	**90.22±6.58**	67.93±62.01	0.006
✓	✓	**75.92±8.95**	**62.01±11.42**	79.38±11.68	**99.75±0.28**	89.57±5.77	**53.00±66.16**	-

TABLE III
COMPARISON BETWEEN CROSS-ENTROPY LOSS AND DICE LOSS

Method	Metrics					
	DI[%]	JA[%]	SEN[%]	SP[%]	AUC[%]	HD95(mm)
Baseline(CE)	69.46±10.08	54.06±10.95	60.05±15.20	99.91±0.10	79.98±7.57	35.02±52.85
Baseline(Dice)	72.04±10.55	57.35±12.75	79.02±16.39	99.67±0.33	89.35±8.12	68.34±62.71
Ours(CE)	71.66±8.51	56.51±10.07	69.95±15.49	99.82±0.21	84.88±7.69	57.02±60.84
Ours(Dice)	75.92±8.95	62.01±11.42	79.38±11.68	99.75±0.28	89.57±5.77	53.00±66.16

TABLE IV
COMPARISON BETWEEN DIFFERENT CONSISTENCY LOSSES

Loss	Metrics					
	DI[%]	JA[%]	SEN[%]	SP[%]	AUC[%]	HD95(mm)
CE	74.17±9.19	59.78±11.52	79.03±9.87	99.69±0.29	89.36±4.88	73.78±69.53
L1 norm	75.08±8.47	60.80±10.38	77.28±12.37	99.77±0.27	88.52±6.11	67.77±64.68
Dice	75.50±7.69	61.26±10.00	81.37±9.42	99.69±0.28	90.53±4.64	71.54±65.73
MSE	75.92±8.95	62.01±11.42	79.38±11.68	99.75±0.28	89.57±5.77	53.00±66.16

strong bias towards the background. On the contrary, the Dice loss can encourage the model to focus more on the foreground so that the performance can be promoted in our task.

2) Comparison Between Different Consistency Losses: We have conducted the experiments for different consistency loss, and the results are presented in Table IV. It can be noticed that the performance of framework with CE loss is significantly lower than other losses, demonstrating that the CE loss is not applicable to perform consistency regularization in our task. Compared with MSE, even though the performance of L1 norm and Dice loss is close to MSE, the HD95 is much lower than L1 norm and Dice loss with MSE as consistency loss, demonstrating the superiority of MSE loss on the boundary delineation. Consequently, MSE is adopted as the loss function to perform consistency regularization.

F. Analysis of Complexity

We summarize the FLOPs, params and runtime memory of methods including the ablation studies in Table V. It can be observed from Table V that the values of FLOPs are similar to the frameworks containing same number of models (*e.g.* MT and Ours). Since there are two trainable models, the needs of params and runtime memory of our framework are more than other frameworks with two models. To further investigate the efficiency of our framework, we provide the baseline with

TABLE V
COMPLEXITY ANALYSIS

Method	FLOPs(G)	Params(M)	Runtime Memory(GB)
Baseline	13.68	7.77	3.44
Baseline-h	54.6	31.04	6.04
DAN	13.79	7.94	3.45
ASDNet	14.55	8.25	3.61
TCSM	27.36	7.77	5.28
MT	27.36	7.77	6.48
UA-MT	27.36	7.77	7.20
Self-training	13.68	7.77	3.47
SemiInfNet	13.68	7.77	3.68
Noise-Robust	27.36	7.77	6.48
Self-paced	54.72	15.54	10.14
CCT	25.2	7.78	6.50
Self-Correcting	27.65	15.52	6.37
GCT	27.64	16.06	9.45
Only Co-training	27.36	15.54	9.42
Only Consistency	13.68	7.77	3.47
Ours	27.36	15.54	9.42

the version (baseline-h) of doubled number of convolutional layers. Combining Table V and Table I, it can be observed that our framework outperforms the baseline-h with much fewer params, revealing that it is more significant for our task to adopt effective learning strategy rather than simply increasing the complexity of network. Furthermore, compared with those methods with similar parameters (Self-paced, Self-Correcting, GCT), our method achieves the state-of-the-art performance which also demonstrates the effectiveness of our framework.

The complexity indices of ablation studies are also presented in Table V. Intuitively, compared with the Baseline model, more parameters lead to the performance improvement of co-training framework. However, it can be observed that the co-training framework outperforms the Baseline-h model with a large improvement even if the amount of parameters is only half of the Baseline-h, indicating that the contribution for performance improvement heavily depends on the co-training rather than the increase of parameters. In addition, the performance of model only with consistency regularization can outperform the Baseline and Baseline-h, demonstrating the effectiveness of consistency regularization. Consequently, our proposed method can achieve the state-of-the-art performance only with limited complexity.

V. DISCUSSIONS

Accurately quantifying the COVID-19 lesions is important for severity evaluation of COVID-19. Several methods have been proposed to segment COVID-19 lesions from CT scans [10], [11]. However, a CT scan usually contains a lot of slices, in addition, the appearances of COVID-19 lesions in CT scans vary greatly, thus labeling CT scans is labor-intensive and time-consuming. With another outbreak of COVID-19 in this summer, few experts will have enough time to do labeling job, which is another difficulty for labeling COVID-19 lesions. Under such circumstances, developing data-efficient methods for COVID-19 lesion segmentation is in urgent need. In this paper, we aim to build a semi-supervised segmentation method to deal with this problem.

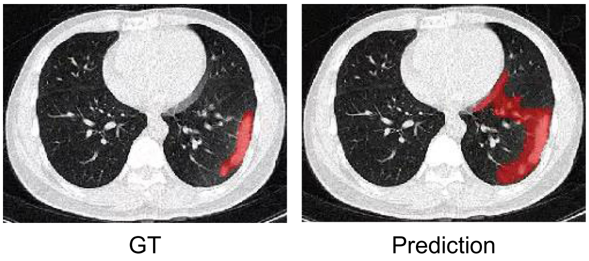

Fig. 7. Visualization of failure case of our method.

We evaluate the state-of-the-art semi-supervised methods for medical image segmentation, the results are reported in Table I. We observe that methods including GAN based methods and consistency based methods only obtain limited improvements compared to the baseline. To filter the unreliable knowledge during training, [14], [34] incorporate confidence map into training progress. However, the two methods do not achieve expectant results, which remind us that the complexity and variability of COVID-19 lesions may have negative impacts for uncertainty estimation. Self-training [33] obtains significant improvement compared to the baseline, which inspires us that models taught by itself can learn useful knowledge from unlabeled data. The defect of self-training is the requirement of several iterations, leading to high computational cost. Different from these methods, we proposed a co-training framework for COVID-19 segmentation which encourages two models to mutually learn unsupervised knowledge. In this framework, we initialize the two models by different parameters so that they can generate different views towards same unlabeled data. During training, the two views complement each other to improve each other's performance.

Without any extra guidance, the predictions of unlabeled data may be unreliable, so that two models suffer from noisy supervision from each other during training. Intuitive idea is to filter these noisy predictions. However, according to the results of [14], [34], estimating the uncertainty map for complicated COVID-19 lesions only guided by limited labeled data may not be a reasonable solution. Based on this observation, to alleviate the adverse impact of noisy pseudo-labels, we propose a self-ensembling consistency regularization to constrain the training progress, in which we encourage the current prediction and the ensembled predictions from previous epochs to be consistent. Our experimental results demonstrate the effectiveness of self-ensembling consistency targets in improving the performance of co-training framework.

Despite the performance improvement of our method, the area between COVID-19 lesions and normal issue may exists slight transition or have similarities. In this case, the prediction for this kind of area could be wrong, as is illustrated in Fig. 7. We argue that this is due to lack of prior knowledge to constrain models. Under the setting of semi-supervised learning, induced by the scarcity of labeled data, models cannot obtain enough reliable guidance to distinguish the ambiguity area, which is an inevitable defect for predicting unseen data.

The incorporation of prior knowledge in deep learning methods is attracting more and more attention for improving performance of deep learning models. Some prior knowledge which act as constrained item have been proposed to integrate into loss functions [44], [45]. The participation of proper prior knowledge often bring performance improvement for models. For our semi-supervised method, the predictions of unlabeled data may be wrong due to the lack of prior knowledge, whereas the wrong predictions still join in training. In the future work, we will focus on developing an prior knowledge constraint to optimize the predictions of unlabeled data, especially patient-specific prior knowledge for better utilizing unlabeled data.

VI. Conclusion

In this paper, we present a novel semi-supervised segmentation method trained by limited labeled data for segmenting COVID-19 lesions. Specifically, we build a co-training framework in which there are two models mutually teaching each other with their own predicted results on unlabeled data, in addition, we propose a self-ensembling consistency regularization for co-training framework to alleviate the negative impacts of unreliable exchanged knowledge between the two collaborative models. We evaluate our framework on a dataset contains 103 CT scans, experimental results show the significant performance of our method over the state-of-the-art semi-supervised methods for reducing the requirement of labeled CT scans. In the future work, we focus on distinguishing the unreliable pseudo-labels in co-training framework for further improving the performance of SSL.

References

[1] J. P. Kanne, “Chest CT findings in 2019 Novel Coronavirus (2019-nCoV) infections from Wuhan, China: Key points for the radiologist,” *Radiology*, vol. 295, no. 1, pp. 16–17, 2020.

[2] Y. Fang *et al.*, “Sensitivity of chest CT for COVID-19: Comparison to RT-PCR,” *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020.

[3] J. Ma *et al.*, “Towards data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation,” *Medical Physics*, vol. 48, no. 3, pp. 1197–1210, 2021.

[4] J. Bullock *et al.*, “Mapping the landscape of artificial intelligence applications against COVID-19,” *Journal of Artificial Intelligence Research*, vol. 69, pp. 807–845, 2020.

[5] F. Shi *et al.*, “Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19,” *IEEE Rev. Biomed. Eng.*, vol. 14, no. 4, pp. 4–15, Apr. 2021.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Cham, Switzerland: Springer, 2015, pp. 234–241.

[7] Z. Zhou *et al.*, “Unet++ : A nested u-net architecture for medical image segmentation,” *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[8] O. Oktay *et al.*, “Attention U-Net: Learning where to look for the pancreas,” *Int. Conf. Med. Imag. Deep Learn.*, 2018.

[9] X. Chen, L. Yao, and Y. Zhang, “Residual attention U-Net for automated multi-class segmentation of COVID-19 chest CT images,” 2020, *arXiv:2004.05645*.

[10] L. Zhou *et al.*, “A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2638–2652, Aug. 2020.

[11] D. P. Fan *et al.*, “Inf-Net: Automatic COVID-19 lung infection segmentation from CT images,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[12] I. Laradji *et al.*, “A weakly supervised consistency-based learning method for COVID-19 segmentation in CT images,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.* 2021, pp. 2453–2462.

[13] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, “Transformation-consistent self-ensembling model for semisupervised medical image segmentation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[14] L. Yu *et al.*, “Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.* Cham, Switzerland: Springer, Cham, 2019, pp. 605–613.

[15] C. S. Perone and J. Cohen-Adad, “Deep semi-supervised segmentation with weight-averaged consistency targets,” *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Cham, Switzerland: Springer, 2018, pp. 12–19.

[16] J. Peng *et al.*, “Deep co-training for semi-supervised image segmentation,” *Pattern Recognit.*, vol. 107, 2020, Art. no. 107269.

[17] Y. Xia *et al.*, “Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation,” *Med. Image Anal.*, vol. 65, 2020, Art. no. 101766.

[18] S. Wu, J. Li, C. Liu, Z. Yu, and H.-S. Wong, “Mutual learning of complementary networks via residual correction for improving semi-supervised classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6500–6509.

[19] Z. Ke *et al.*, “Guided collaborative training for pixel-wise semi-supervised learning,” in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2020, pp. 429–445.

[20] Y. Ouali, C. Hudelot, and M. Tami, “An overview of deep semi-supervised learning,” 2020, *arXiv:2006.05278*.

[21] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[22] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” 2016, *arXiv:1610.02242*.

[23] X. Wang *et al.*, “A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2615–2625, 2020.

[24] Y. Cao *et al.*, “Longitudinal assessment of COVID-19 using a deep learning-based quantitative CT pipeline: Illustration of two cases,” *Radiol., Cardiothoracic Imag.*, vol. 2, no. 2, 2020, Art. no. e200082.

[25] L. Li *et al.*, “Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy,” *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[26] J. Chen, L. Wu, J. Zhang *et al.*, “Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[27] S. Jin *et al.*, “AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system,” *Appl. Soft Comput.*, vol. 98, no. 106897, 2021.

[28] F. Shan *et al.*, “Lung infection quantification of COVID-19 in CT images with deep learning,” 2020, *arXiv:2003.04655*.

[29] “COVID-19 CT segmentation dataset,” Apr. 2020. [Online]. Available: https://medicalsegmentation.com/covid19/

[30] G. Wang *et al.*, “A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2653–2663, Aug. 2020.

[31] I. Laradji *et al.*, “A weakly supervised region-based active learning method for COVID-19 segmentation in CT images,” 2020, *arXiv:2007.07012*.

[32] Z. Xu *et al.*, “GASNet: Weakly-supervised framework for COVID-19 lesion segmentation,” 2020, *arXiv:2010.09456*.

[33] W. Bai *et al.*, “Semi-supervised learning for network-based cardiac MR image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv*., Cham, Switzerland: Springer, 2017, pp. 253–260.

[34] D. Nie *et al.*, “ASDNet: Attention based semi-supervised deep networks for medical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv*., Cham, Switzerland: Springer, 2018, pp. 370–378.

[35] Y. Zhang *et al.*, “Deep adversarial networks for biomedical image segmentation utilizing unannotated images,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv*., Cham, Switzerland: Springer, 2017, pp. 408–416.

[36] Q. Dou *et al.*, “Unpaired multi-modal segmentation via knowledge distillation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2415–2425, 2020.

[37] W. Cui *et al.*, “Semi-supervised brain lesion segmentation with an adapted mean teacher model,” in *Proc. Int. Conf. Inf. Process. Med. Imag*., Cham, Switzerland: Springer, 2019, pp. 554–565.

[38] F. Milletari, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[39] A. Paszke *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst*., 2019, pp. 8026–8037.

[40] D. Ulyanov, A. Vedaldi, V. Lempitsky, “Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis,” in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2017, pp. 6924–6932.

[41] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” *Proc. Int. Conf. Mach. Learn.*, vol. 30, no. 1, p. 3, 2013.

[42] K. He *et al.*, “Delving deep into rectifiers: Surpassing human-level performance on imagenet,” in *Proc. IEEE Int. Conf. Comput. Vis*., 2015, pp. 1026–1034.

[43] J. Ma *et al.*, “COVID-19 CT lung and infection segmentation dataset (version 1.0) [data set] Zenodo,” 2020. [Online]. Available: http://doi.org/10.5281/zenodo.3757476

[44] H. Kervadec *et al.*, “Constrained-CNN losses for weakly supervised segmentation,” *Med. Image Anal.*, vol. 54, pp. 88–99, 2019.

[45] J. Peng *et al.*, “Discretely-constrained deep network for weakly supervised segmentation,” *Neural Netw.*, vol. 130, pp. 297–308, 2020.

[46] P. Wang *et al.*, “Self-paced and self-consistent co-training for semi-supervised image segmentation,” *Med. Image Anal.*, vol. 73, Art. no. 102146, 2021.

[47] Y. Ouali, C. Hudelot, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 12674–12684.

[48] M. S. Ibrahim *et al.*, “Semi-supervised semantic image segmentation with self-correcting networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 12715–12725.