A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images

Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang

***Abstract*—Segmentation of pneumonia lesions from CT scans of COVID-19 patients is important for accurate diagnosis and follow-up. Deep learning has a potential to automate this task but requires a large set of high-quality annotations that are difficult to collect. Learning from noisy training labels that are easier to obtain has a potential to alleviate this problem. To this end, we propose a novel noise-robust framework to learn from noisy labels for the segmentation task. We first introduce a noise-robust Dice loss that is a generalization of Dice loss for segmentation and Mean Absolute Error (MAE) loss for robustness against noise, then propose a novel COVID-19 Pneumonia Lesion segmentation network (COPLE-Net) to better deal with the lesions with various scales and appearances. The noise-robust Dice loss and COPLE-Net are combined with an adaptive self-ensembling framework for training, where an Exponential Moving Average (EMA) of a student model is used as a teacher model that is adaptively updated by suppressing the contribution of the student to EMA when the student has a large training loss. The student model is also adaptive by learning from the teacher only when the teacher outperforms the student. Experimental results showed that: (1) our noise-robust Dice loss outperforms existing noise-robust loss functions, (2) the proposed COPLE-Net achieves higher performance than state-of-the-art image segmentation networks, and (3) our framework with adaptive self-ensembling significantly outperforms a standard training process and surpasses other noise-robust training approaches in the scenario of learning from noisy labels for COVID-19 pneumonia lesion segmentation.**

***Index Terms*—COVID-19, convolutional neural network, noisy label, segmentation, pneumonia.**

I. Introduction

THE coronavirus disease 2019 (COVID-19) has become a global pandemic since the beginning of 2020 [1]–[3]. The disease has been regarded as a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) and the end of January 2020. Up to April 10, 2020, there have been more than 1.5 million cases of COVID-19 reported globally, with more than 92 thousands deaths [4].

The most common symptoms of COVID-19 patients include fever, cough and shortness of breath [5], and the patients typically suffer from pneumonia [1], [6]. Computed Tomography (CT) imaging plays a critical role for detection of manifestations in the lung associated with COVID-19 [7], [8], where segmentation of the infection lesions from CT scans is important for quantitative measurement of the disease progression [9] in accurate diagnosis and follow-up assessment [3], [6], [10]. As manual segmentation of the lesions from 3D volumes is labor-intensive, time-consuming and suffers from inter- and intra-observer variabilities, automatic segmentation of the lesions is highly desirable in clinic practice [3].

Despite its importance for diagnosis and treatment decisions, automatic segmentation of COVID-19 pneumonia lesions from CT scans is challenging due to several reasons. First, the infection lesions have a variety of complex appearances such as Ground-Glass Opacity (GGO), reticulation, consolidation and others [5]. Second, the sizes and positions of the pneumonia lesions vary largely at different stages of the infection and among different patients. In addition,

Manuscript received April 28, 2020; revised May 9, 2020 and May 27, 2020; accepted May 29, 2020. Date of publication June 5, 2020; date of current version July 30, 2020. This work was supported by the National Natural Science Foundation of China under Grant 81771921 and Grant 61901084. *(Corresponding authors: Guotai Wang; Shaoting Zhang.)*

Guotai Wang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: guotai.wang@uestc.edu.cn).

Xinglong Liu and Ning Huang are with SenseTime Research, Beijing 100080, China (e-mail: liuxinglong@sensetime.com; huangning@sensetime.com).

Chaoping Li is with the Fengcheng People's Hospital, Fengcheng 331100, China (e-mail: 18625547@qq.com).

Zhiyong Xu is with the Huanggang Traditional Chinese Medicine Hospital, Huanggang 438000, China (e-mail: 1304113673@qq.com).

Jiugen Ruan is with the Xinyu City People's Hospital, Xinyu 338000, China (e-mail: 813832561@qq.com).

Haifeng Zhu is with the Civil Aviation General Hospital, Beijing 100123, China (e-mail: 155108954@qq.com).

Tao Meng is with RIMAG Medical Imaging Corporation, Beijing 100022, China (e-mail: eluaid@163.com).

Kang Li is with the West China Biomedical Big Data Center, Sichuan University, West China Hospital, Chengdu 610041, China (e-mail: likang@wchscu.cn).

Shaoting Zhang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with SenseTime Research, Shanghai 200233, China (e-mail: zhangshaoting@uestc.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3000314

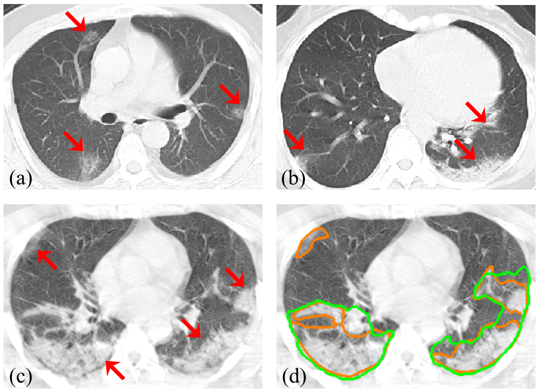

Fig. 1. Complex appearances of pneumonia lesions in CT scans of COVID-19 patients. (a-c) are from three different patients, where red arrows highlight some lesions. (d) shows annotations of (c) given by different observers.

the lesions have irregular shapes and ambiguous boundaries, and some lesion patterns such as GGO have a low contrast with surrounding regions, as shown in Fig. 1(a-c). These challenges not only make it difficult to automatically segment the lesions, but also bring obstacles for obtaining accurate manual annotations for training. In recent years, deep learning with Convolutional Neural Networks (CNNs) has achieved state-of-the-art performance for many medical image analysis tasks [3], [11]. For medical image segmentation [12], its success relies on accurate annotation of a large set of training images implemented by experts, which is expensive to acquire and limited by the availability of experts.

For the COVID-19 pneumonia lesion segmentation task, the pixel-level annotations are often noisy and clean annotations are extremely difficult to collect due to several reasons. First, different annotators may have different annotation standards that lead to inter-observer variability, and high intra-observer variability may also exist. These variabilities are very likely to cause noise in the annotations, as shown in Fig. 1(d) that demonstrates disagreement between two annotators. Second, to reduce the annotation efforts, some researchers use a human-in-the-loop strategy [9], where the annotator only provides refinements to algorithm-generated labels for annotating. In such cases, the annotations can be largely biased towards the results of the algorithm and thus contain noisy pixel-level labels. In addition, collecting less accurate annotations from non-experts is another promising solution to overcome the limited availability of experts [13], but these annotations are also noisy at pixel level and may limit the performance of the deep learning model [14], [15].

Despite several recent studies on automatic segmentation of COVID-19 pneumonia lesions from CT scans, existing works [6], [8], [10] mainly use off-the-shelf CNN models such as U-Net [16] for segmentation, and they use a standard training process that ignores the existence of noisy labels. In this work, we aim at developing a more advanced CNN model for the challenging segmentation task and try to overcome the noisy annotations to achieve better segmentation performance.

Although learning from noisy labels has been increasing investigated, most existing works focus on image classification tasks [13]. Among various methods to deal with noisy labels, some novel noise-robust loss functions are attracting increasing attentions as they do not require a specific network structure and can be easily combined with typical training procedures. For example, the Mean Absolute Error (MAE) loss [14] and Generalized Cross Entropy (GCE) loss [17] have been shown to be noise-robust for image classification. However, simply applying them to segmentation tasks may lead to limited performance due to the imbalanced foreground and background pixel numbers [18]. To deal with noisy labels for segmentation of medical images, a CNN was designed in [15] to distinguish noisy labels from clean ones. In [19], meta-learning was used to assign lower weights to pixels whose loss gradient direction is further from those of clean data. Both methods require a set of clean labels for training. However, for COVID-19 pneumonia lesions, even expert annotations may contain some noise due to the above challenges, and it is difficult to obtain a set of absolutely clean labels. In this paper, we deal with the learning problem in the scenario where all the annotations of training images are assumed to be noisy.

A. Contributions

The contributions of this work are three-fold. First, to deal with noisy annotations for training CNNs to segment COVID-19 pneumonia lesions, we propose a novel noise-robust Dice loss function, which is a combination and generalization of MAE loss [14] that is robust against noisy labels and Dice loss [18] that is insensitive to foreground-background imbalance. Second, we propose a novel noise-robust learning framework based on self-ensembling of CNNs [20], [21], where an Exponential Moving Average (EMA, a.k.a. teacher) of a model is used to guide a standard model (a.k.a. student) to improve the robustness. Differently from previous self-ensembling methods for semi-supervised learning [20], [21] and domain adaptation [22], [23], we propose two adaptive mechanisms to better deal with noisy labels: adaptive teacher that suppresses the contribution of the student to EMA when the latter has a large training loss, and adaptive student that learns from the teacher only when the teacher outperforms the student. Thirdly, to better deal with the complex lesions, we propose a novel COVID-19 Pneumonia Lesion segmentation network (COPLE-Net) that uses a combination of max-pooling and average pooling to reduce information loss during down-sampling, and employs bridge layers to alleviate the semantic gap between features in the encoder and decoder. Experimental results with CT images of 558 COVID-19 patients showed the effectiveness of our proposed noise-robust Dice loss function, COPLE-Net and adaptive self-ensembling in learning from noisy labels for COVID-19 pneumonia lesion segmentation. Our method has a potential to reduce the annotation burden for large-scale 3D image datasets and alleviate the limited access to high-quality pixel-level labels given by experts.

B. Related Works

1) Segmentation of COVID-19 Pneumonia Lesions: Despite the important role of infection lesion segmentation for quantitative assessment in diagnosis and follow-up, there have been a few works on automatic segmentation of COVID-19 pneumonia lesions from medical images [3]. Li *et al.* [8] employed the U-Net [16] to segment the lungs from CT scans to distinguish COVID-19 pneumonia from community acquired pneumonia. Cao *et al.* [10] and Huang *et al.* [6] also employed U-Net to segment the lungs and pulmonary opacities for quantitative measurements such as lung opacification percentage that can be used for longitudinal assessment of the disease. The UNet++ [24] was also used for detection [25] and segmentation [26] of the infection lesions from CT scans. In [9], a VB-Net combining V-Net [18] and the bottleneck structure [27] was used to segment multiple structures including lung lobes, lung segments and infection regions from CT scans of COVID-19 patients, and a human-in-the-loop strategy was adopted for efficient annotation. However, these works employed a standard training procedure for the segmentation task without considering noise in the annotations.

2) Noise-Robust Deep Learning: Most existing methods to deal with noisy labels with deep learning were initially proposed for image classification tasks [13]. Ghosh *et al.* [14] showed that Mean Absolute Error (MAE) is tolerant to label noise and performs better than standard Cross Entropy (CE) loss. Zhang and Sabuncu [17] showed that MAE down-weights difficult samples with correct labels and proposed Generalized Cross Entropy (GCE) as a better noise-robust loss for image classification. In [28], a self-ensemble label filtering method was proposed to progressively filter out wrong labels during training. Sample re-weighting was used in [29], [30] for suppressing potential incorrectly labeled images. In [31], samples with large training loss were regarded as noisy and discarded during training. Other techniques including consistency-based regularization [32], model ensemble [33] and iterative training [31] have also been used for learning from noisy labels for classification. However, few of them have been validated with segmentation tasks.

To learn from noisy labels for medical image segmentation, Zhu *et al.* [15] designed a CNN to distinguish noisy labels from clean ones. Mirikharaji *et al.* [19] assigned lower weights to pixels whose loss gradient direction is further from those of clean data. However, both of them require a set of clean labels for training. In [34], an attention network was proposed for semi-supervised biomedical image segmentation with noisy labels. In [13], dual CNNs with iterative label update was used for fetal brain segmentation. However, these methods were validated with simulated noisy labels, and it is still challenging to deal with noisy labels for medical image segmentation tasks.

II. Method

The proposed noise-robust framework for learning from noisy labels for automatic segmentation of COVID-19 pneumonia lesions is illustrated in Fig. 2. We propose a novel COVID-19 Pneumonia Lesion segmentation network (COPLE-Net) to deal with the lesions at different scales. To make the training process robust against noisy labels, we propose a novel noise-robust Dice loss function and integrate it into a self-ensembling framework, where an adaptive teacher and an adaptive student are introduced to further improve the performance in dealing with noisy labels.

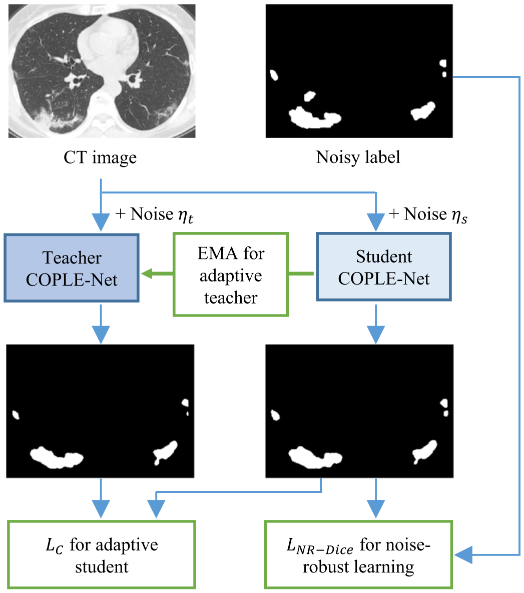

Fig. 2. The proposed noise-robust framework for COVID-19 pneumonia lesion segmentation with noisy training labels. It consists of three components: a noise-robust Dice loss $L_{\text{NR-Dice}}$, a novel network COPLE-Net and an adaptive self-ensembling, where the teacher model T is an exponential moving average of the student model S, and is adaptively updated according to the performance of S. The student S learns from the teacher T through a consistency loss L_C adaptively, i.e., only when T outperforms S.

A. COVID-19 Pneumonia Lesion Segmentation Network

Due to the large range of slice spacing of clinical CT scans of COVID-19 patients, we use 2D CNNs for the segmentation task in this work. Our proposed COPLE-Net is shown in Fig. 3. Inspired by the state-of-the-art performance of variants of U-Net [16], [36]–[38] with an encoder-decoder structure, we use a similar backbone but extend it with several important modules. First, differently from [16], [36]–[38] that only use max-pooling for down-sampling, we introduce a dual pooling, i.e., a concatenation of max-pooling and average-pooling as down-sampling, which has a lower information loss than a simple max-pooling. Second, we replace the typical skip connection between the encoder and the decoder with a bridge layer (i.e., 1 × 1 convolution) to map the low-level features from the encoder to a lower dimension (i.e., the channel number is reduced by half) before concatenating them with high-level features from the decoder, in order to alleviate the semantic gap between the low-level and high-level features [39]. Thirdly, to better segment lesions at different scales, we add an Atrous Spatial Pyramid Pooling (ASPP) module [35] at the bottleneck of the encoder-decoder structure, where the ASPP consists of four parallel layers of dilated convolution with different dilation rates, and their outputs are concatenated so that the network can better capture multi-scale features for segmentation of small and large lesions.

In addition, in each convolutional block, we use residual connection [27] to facilitate the training and add a Squeeze-and-Excitation (SE) block [40] to calibrate different channels for better performance. We also implement each up-sampling layer in the decoder by a 1 × 1 convolution layer followed by

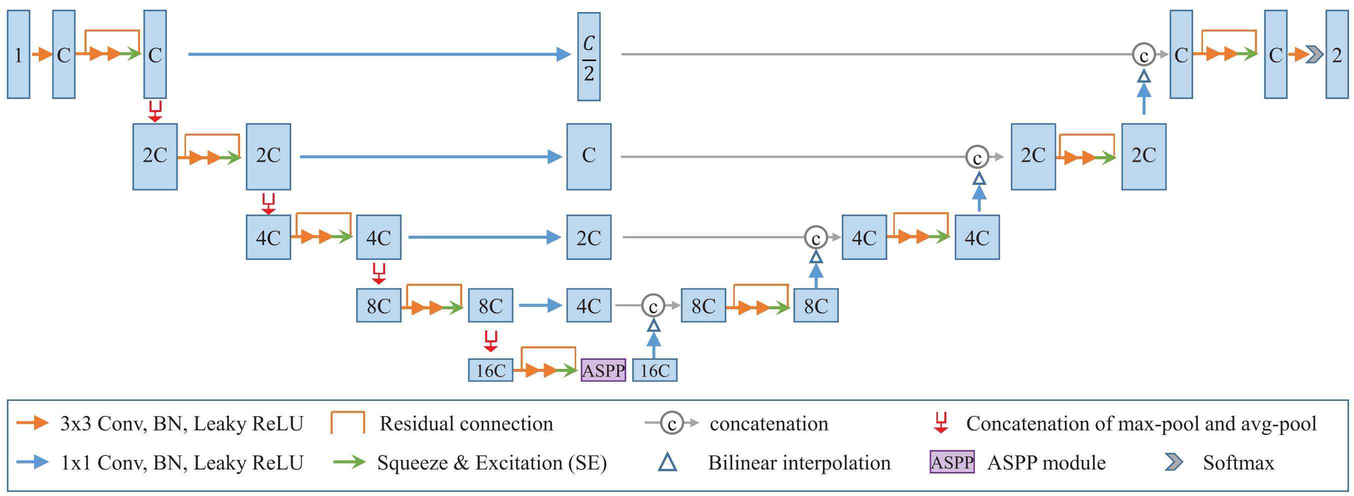

Fig. 3. The proposed COVID-19 Pneumonia Lesion segmentation network (COPLE-Net). Blue rectangles with number C represent feature maps with C channels. We employ a concatenation of max-pooling and average pooling to reduce information loss during down-sampling, and use bridge layers to alleviate the semantic gap between features from the encoder and the decoder. ASPP [35] block is used at the bottleneck to better deal with lesions at multiple scales.

bilinear interpolation, where the former reduces the number of feature channels to match that of the output of the bridge layer, and the latter improves the spatial resolution before the concatenation of the low-level and high-level features, as shown in Fig. 3. Compared with standard transposed convolution for up-sampling [16], our implementation leads to a reduced number of parameters with higher efficiency. The channel number in the first resolution level of COPLE-Net is C, and is doubled each time when using a down-sampling layer, as illustrated in Fig. 3.

B. Noise-Robust Dice Loss

Pneumonia lesions at an early stage often occupy a small region of the image, which could lead the CNN's prediction to be strongly biased towards the background when trained with a standard image classification loss function, i.e., cross entropy. The Dice loss function in Eq.(1) proposed by Milletari *et al.* [18] has been shown to be effective to overcome this problem by implicitly establishing a balance between foreground and background classes:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} = \frac{\sum_i^N (p_i - g_i)^2}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{1}$$

where N is the number of pixels in the image. p_i is the foreground probability of pixel i predicted by the CNN, and g_i is the corresponding value derived from the binary training label. L_{Dice} can be seen as a variant of Mean Square Error (MSE), with the numerator assigning higher weights to pixels with larger prediction errors. In [14], it was shown that MSE is not robust against noisy labels, and the authors theoretically demonstrated that MAE could achieve higher robustness than cross entropy and MSE under certain assumptions. The MAE loss can be reformulated for segmentation as:

$$L_{\text{MAE}} = \frac{\sum_i^N \mid p_i - g_i \mid}{N} \tag{2}$$

Despite its tolerance to noisy labels, L_{MAE} treats each pixel equally and can perform poorly with deep CNNs and challenging datasets, as demonstrated in [17]. It also cannot deal with the foreground-background imbalance in segmentation tasks. To make use of the advantages of L_{Dice} and L_{MAE} and overcome their limitations, we propose a novel noise-robust Dice loss $L_{\text{NR-Dice}}$ that is a generalization of L_{Dice} and L_{MAE}:

$$L_{\text{NR-Dice}} = \frac{\sum_i^N \mid p_i - g_i \mid^{\gamma}}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \epsilon} \tag{3}$$

where $\gamma \in [1.0,\ 2.0]$ is a hyper-parameter and $\epsilon = 10^{-5}$ is a small number for numerical stability. When $\gamma = 2.0$, $L_{\text{NR-Dice}}$ equals to the Dice loss L_{Dice}. When $\gamma = 1.0$, $L_{\text{NR-Dice}}$ becomes a weighted version of L_{MAE}.

The best value of γ was 1.5 based on our experiments as shown in Fig. 4. Note that $\gamma = 1.0$ (i.e., variant of L_{MAE} and not emphasizing hard samples) has a good robustness against noisy labels [14], but is not suitable for deep CNNs and challenging datasets [17]. In contrast, $\gamma = 2.0$ (i.e., L_{Dice}) is sensitive to noise [14] as it largely emphasizes hard samples, but more suitable for training with deep CNNs and challenging datasets [18]. Setting $\gamma = 1.5$ in our $L_{\text{NR-Dice}}$ is less sensitive to noisy labels than $\gamma = 2.0$ in L_{Dice}, and $L_{\text{NR-Dice}}$ handles foreground-background imbalance in the same way as L_{Dice}, i.e., weighting each class basically based on the inverse of its volume, as shown in Eq. (3). Therefore, our $L_{\text{NR-Dice}}$ is robust against noisy labels and foreground-background imbalance at the same time. It should be noticed that the proposed $L_{\text{NR-Dice}}$ is not equavalent to simply using $L_{\text{MAE}} + L_{\text{Dice}}$ (i.e., linear combination of MAE and MSE), where the L_{MAE} term is still not suitable for deep CNNs and biased towards the background class, and the MSE in L_{Dice} is still sensitive to noise. Therefore, our proposed $L_{\text{NR-Dice}}$ is is more suitable than $L_{\text{MAE}} + L_{\text{Dice}}$ for segmentation tasks with noisy labels and deep CNNs.

C. Noise-Robust Adaptive Self-Ensembling

To further improve the performance of dealing with noisy labels, we integrate our COPLE-Net and $L_{\text{NR-Dice}}$ into a self-ensembling framework that was originally proposed for semi-supervised learning [20], [21]. It consists of a teacher model T and a student model S with the same CNN structure, and T is updated as an Exponential Moving Average (EMA) of S:

$$\theta_t^* = \alpha\theta_{t-1}^* + (1-\alpha)\theta_t \tag{4}$$

where θ^* and θ are parameters of T and S, respectively, and both T and S are implemented by our COPLE-Net in this work. t is the training step, and $\alpha \in (0,\ 1)$ is a smoothing coefficient. α and $1-\alpha$ specify the contributions of θ_{t-1}^* and θ_t to θ_t^*, respectively. Due to EMA, the teacher model T is more stable than the student model S and can be used to supervise S through a consistency loss L_c [20], [21]. Let x and y represent a training image and its noisy annotation respectively, the overall loss function for training is:

$$L = L_{\text{seg}}\Big(S(x+\eta_s), y\Big) + \lambda L_c\Big(S(x+\eta_s), T(x+\eta_t)\Big) \tag{5}$$

where η_s and η_t are random Gaussian noises added to the input of S and T respectively for data augmentation. L_{seg} is a segmentation loss that can be implemented by L_{Dice} or $L_{\text{NR-Dice}}$. L_c is a consistency loss to encourage T and S to give close predictions, and is implemented by MAE due to its robustness. $\lambda \geq 0$ is the weight of L_c.

In previous self-ensembling methods [20], [21] for semi-supervised learning, values of α and λ were manually set as fixed numbers or changed gradually according to the training step t, without considering the performance of S and T and the existence of noisy labels. In our segmentation task, the student model S may have a poor performance at step t due to noisy labels, and updating the teacher model T with a predefined α (e.g., 0.99 in [21]) may lead T to be corrupted by the noisy labels as well. In addition, there is no guarantee that T always performs better than S during training, and applying L_c when T performs worse than S may decrease the performance of S. To address these problems, we propose adaptive self-ensembling where S and T are adaptively updated according to the performance of each other, as shown in Fig. 2.

First, we propose an adaptive teacher by suppressing the contribution of S to T (i.e., EMA) when S has a poor performance with large training loss values that may be caused by noisy labels [31]. This is implemented by making the value of α dependent on the training loss of S:

$$\alpha = \begin{cases} \alpha', & \text{if } L_{seg}\Big(S(x+\eta_s), y\Big) < \beta \\ 1.0, & \text{otherwise} \end{cases} \tag{6}$$

where α' is the typical EMA parameter (e.g., 0.99 or 0.999 in [20], [21]) when S has a relatively good performance with low training loss. β is a dynamic threshold value for the segmentation loss of S, and we adaptively set it as the p-th percentile of the student model's loss during the last K (i.e., 100) training steps. When the noisy labels lead the loss of S to exceed β at one certain training step, the teacher model T is not updated by S at that step. Therefore, the affect of noisy labels on the teacher model is suppressed. As large training loss values may be caused by either noisy labels or correctly labeled difficult samples, making both T and S ignore samples with large training loss values may reject hard samples at the same time and lead the framework to learn only from easy samples, which could result in decreased performance in challenging cases. To avoid this problem, we follow a standard strategy that does not ignore hard samples to train S.

Second, we propose an adaptive student model by suppressing T's supervision on S when T has a lower performance than S, which is implemented by making the value of λ dependent on the performance of S and T:

$$\lambda = \begin{cases} \lambda', & \text{if } L_{seg}\Big(T(x+\eta_t), y\Big) > L_{seg}\Big(S(x+\eta_s), y\Big) \\ \lambda'', & \text{otherwise} \end{cases} \tag{7}$$

where λ' is set in the same way as typical self-ensembling frameworks (e.g., 0.1) [20], [21] when T performs better than S. $\lambda'' = 0.1\lambda'$ is a small number to suppress the weight of the consistency loss L_c when T does not outperform S.

III. Experiments and Results

A. Experimental Setting

1) Data and Implementation Details: Clinical CT scans of 558 pneumonia patients with COVID-19 collected from 10 different hospitals were used for experiments. The images had a large range of slice thickness/inter-slice spacing (from 0.625 mm to 8.0 mm), and the pixel size ranged from 0.61 mm to 0.93 mm. We randomly split the images into 378, 50 and 130 for training, validation and testing, respectively. The slice number for training, validation and testing was 52489, 6556 and 17205, respectively. Annotations of the training images were obtained by a human-in-the-loop strategy similar to that in [9], where an initial model trained on a small dataset obtained pseudo labels for the training images, which were refined by non-expert researchers as the annotations. Due to the ambiguous lesion boundaries, inter- and intra-observer variabilities and potential bias towards the initial model, these annotations were inevitably noisy. The ground truth labels for validation and testing images were obtained by manual segmentation by two experts with consensus. As obtaining manual segmentation by experts for all the training samples is time-consuming and challenging, to estimate the noise level in the training set, we randomly selected 50 training images and asked the experts to provide manual annotations as the ground truth. The Dice score between the noisy annotations and the expert annotations of these images was 0.88±0.06.

As the images had a large range of slice thickness, we used 2D CNNs for slice-by-slice segmentation, and implemented our COPLE-Net, [1] $L_{\text{NR-Dice}}$ and the adaptive self-ensembling framework in Pytorch [2] with the PyMIC [3] library on a Ubuntu desktop with an NVIDIA GTX 1080 Ti GPU. The basic channel number C in our COPLE-Net was set as 32. The

[1] Code available at: https://github.com/HiLab-git/COPLE-Net

[2] https://pytorch.org

[3] https://github.com/HiLab-git/PyMIC

dilation rates in the four parallel convolution layers of the ASPP were 1, 2, 4 and 6, respectively. The negative slope of leaky ReLU in our network was set as 0.01. For training, our proposed noise-robust Dice loss $L_{\text{NR-Dice}}$ was combined with the Adam optimizer, weight decay 10^{-5} and mini-batch of 30 slices. The learning rate was initialized as 10^{-4} and halved every 10k iterations, and the training was ended when performance on the validation set stopped to increase for 10k iterations. For our adaptive self-ensembling framework, we followed [20] to set α' to 0.99 and 0.999 at the early and late stage of training, respectively. We set $\lambda' = 0.1$ and β to the 90-th percentile of S's segmentation loss during the last 100 steps according to grid search based on the validation set. The training time of our proposed framework was 5.3 hours, and the inference time for one 3D image was 4.24±2.83s. After training, the model was deployed at SenseCare platform to support clinic research [42].

2) Evaluation Metrics: To better understand the performance of different models when dealing with lesions at different scales, we split the 130 testing images into three groups: Group A containing 42 images with pneumonia lesions smaller than 50 Cubic Centimeters (CC), Group B containing 52 images with pneumonia lesions between 50 and 200 CC, and group C containing 36 images with pneumonia lesions larger than 200 CC.

For quantitative evaluation, we measured the Dice similarity, Relative Volume Error (RVE), and the 95-th percentile of Hausdoff Distance (HD_{95}) between segmentation results and the ground truth in 3D space.

$$Dice(\mathcal{R}_a, \mathcal{R}_b) = \frac{2 \mid \mathcal{R}_a \cap \mathcal{R}_b \mid}{\mid \mathcal{R}_a \mid + \mid \mathcal{R}_b \mid} \tag{8}$$

$$RVE(\mathcal{R}_a, \mathcal{R}_b) = \frac{abs(\mid \mathcal{R}_a \mid - \mid \mathcal{R}_b \mid)}{\mid \mathcal{R}_b \mid} \tag{9}$$

where $\mathcal{R}_a$ and $\mathcal{R}_b$ represent the region segmented by a CNN and the ground truth, respectively.

$$HD'(\mathcal{S}_a, \mathcal{S}_b) = \max_{i \in \mathcal{S}_a} \min_{j \in \mathcal{S}_b} \mid\mid i - j \mid\mid_2 \tag{10}$$

$$HD(\mathcal{S}_a, \mathcal{S}_b) = \max\Big(HD'(\mathcal{S}_a, \mathcal{S}_b), HD'(\mathcal{S}_b, \mathcal{S}_a)\Big) \tag{11}$$

where $\mathcal{S}_a$ and $\mathcal{S}_b$ represent the set of surface points of the target segmented by a CNN and the ground truth respectively. HD_{95} is similar to HD, and it uses the 95-th percentile instead of the maximal value in Eq. (10).

B. Effectiveness of the Noise-Robust Dice Loss

We first investigated the optimal value of hyper-parameter $\gamma \in [1.0, 2.0]$ for our proposed $L_{\text{NR-Dice}}$ in Eq. (3) based on the validation set. We used our training set with noisy annotations to train a 2D version of nnU-Net [41] with different γ values for $L_{\text{NR-Dice}}$. Fig. 4 shows evolution of the segmentation performance on the validation set when γ changes from 1.0 to 2.0. Note that $\gamma = 1.0$ corresponds to a weighted MAE loss and $\gamma = 2.0$ corresponds to the Dice loss. It can be observed that increasing γ from 1.0 to 1.5 leads to improved performance, and when γ is larger than 1.5, the segmentation performance decreases gradually. This shows that our $L_{\text{NR-Dice}}$, which is a generalization of MAE loss and Dice loss, performs better than both of them in dealing with the noisy training labels. According to Fig. 4 showing that $L_{\text{NR-Dice}}$ has the best performance when γ is around 1.5, we set $\gamma = 1.5$ for our $L_{\text{NR-Dice}}$ in the following experiments.

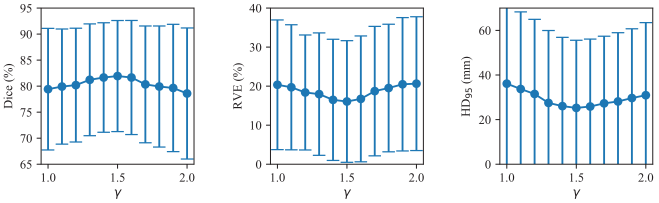

Fig. 4. Segmentation performance on the validation set with different γ values for our noise-robust Dice loss $L_{\text{NR-Dice}}$. The results are based on 2D nnU-Net [41] without using self-ensembling.

TABLE I

QUANTITATIVE EVALUATION OF A SEGMENTATION MODEL TRAINED WITH DIFFERENT LOSS FUNCTIONS. THE RESULTS ARE BASED ON 2D NNU-NET [41] WITHOUT USING SELF-ENSEMBLING

	Dice (%)	RVE (%)	HD_{95} (mm)
CE	76.55±13.30	21.92±17.64	25.80±29.98
Focal CE (γ=2.0) [43]	74.75±15.34	26.14±19.10	29.76±41.30
Focal CE (γ=0.5) [43]	76.62±14.09	21.57±18.41	25.19±30.35
MAE [14]	76.48±14.93	20.46±19.09	28.00±34.65
GCE [17]	76.96±14.54	20.87±10.85	26.79±31.57
Dice [18]	77.89±11.74	20.02±17.14	26.93±30.81
Sørensen-Dice [44]	77.74±12.44	20.48±18.06	27.08±30.93
Focal Dice (β=0.5) [45]	77.28±12.66	19.94±17.55	26.28±32.79
CE + Dice	78.14±12.34	19.89±17.70	23.10±30.34
MAE + Dice	78.34±12.12	19.84±19.03	23.27±28.11
NR-Dice	**79.07±12.88**	**18.37±17.43**	**20.11±25.37**

We further compared the proposed $L_{\text{NR-Dice}}$ with two existing noise-robust loss functions: MAE loss L_{MAE} [14] and Generalized Cross Entropy (GCE) loss L_q [17] where the hyper-parameter q was set to 0.7 as suggested in [17]. They were also compared with: standard CE loss L_{CE}, Dice loss L_{Dice} proposed by Milletari *et al.* [18], $L_{\text{CE}} + L_{\text{Dice}}$, $L_{\text{MAE}} + L_{\text{Dice}}$, Sørensen-Dice [44], focal CE loss [43] and focal Dice loss [45]. The hyper parameter γ for the original focal CE [43] was 2.0, which corresponds to assigning higher weights to pixels with larger prediction error. However, in our task, this tends to overfit noisy labels with large prediction error. Therefore, it would be better to set γ to a small value (e.g., 0.5) to down-weight such pixels. These two variants are referred to as focal CE ($\gamma = 2.0$) and focal CE ($\gamma = 0.5$) respectively. Similarly, the hyper-parameter β of focal Dice [45] was set as 0.5 to down-weight samples with large prediction error during training that maybe caused by noisy labels.

These loss functions were used to train the 2D version of nnU-Net [41] respectively without using self-ensembling. The hyper-parameters for training were the same except the learning rate that was respectively determined for each loss function based on the performance on the validation set. Quantitative comparison listed in Table I shows that focal CE ($\gamma = 2.0$) achieved a lower performance than CE, indicating the ineffectiveness of assigning higher weights to misclassified pixels in training images with noisy labels. Focal CE ($\gamma = 0.5$), MAE loss and GCE loss performed better than CE

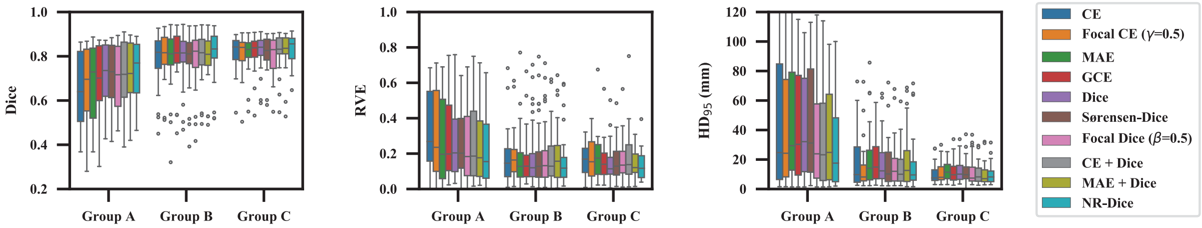

Fig. 5. Group-wise quantitative comparison of a segmentation model trained with different loss functions. Group A: lesions smaller than 50 CC. Group B: lesions between 50 and 200 CC. Group C: lesions larger than 200 CC. The results are based on 2D nnU-Net [41] without using self-ensembling.

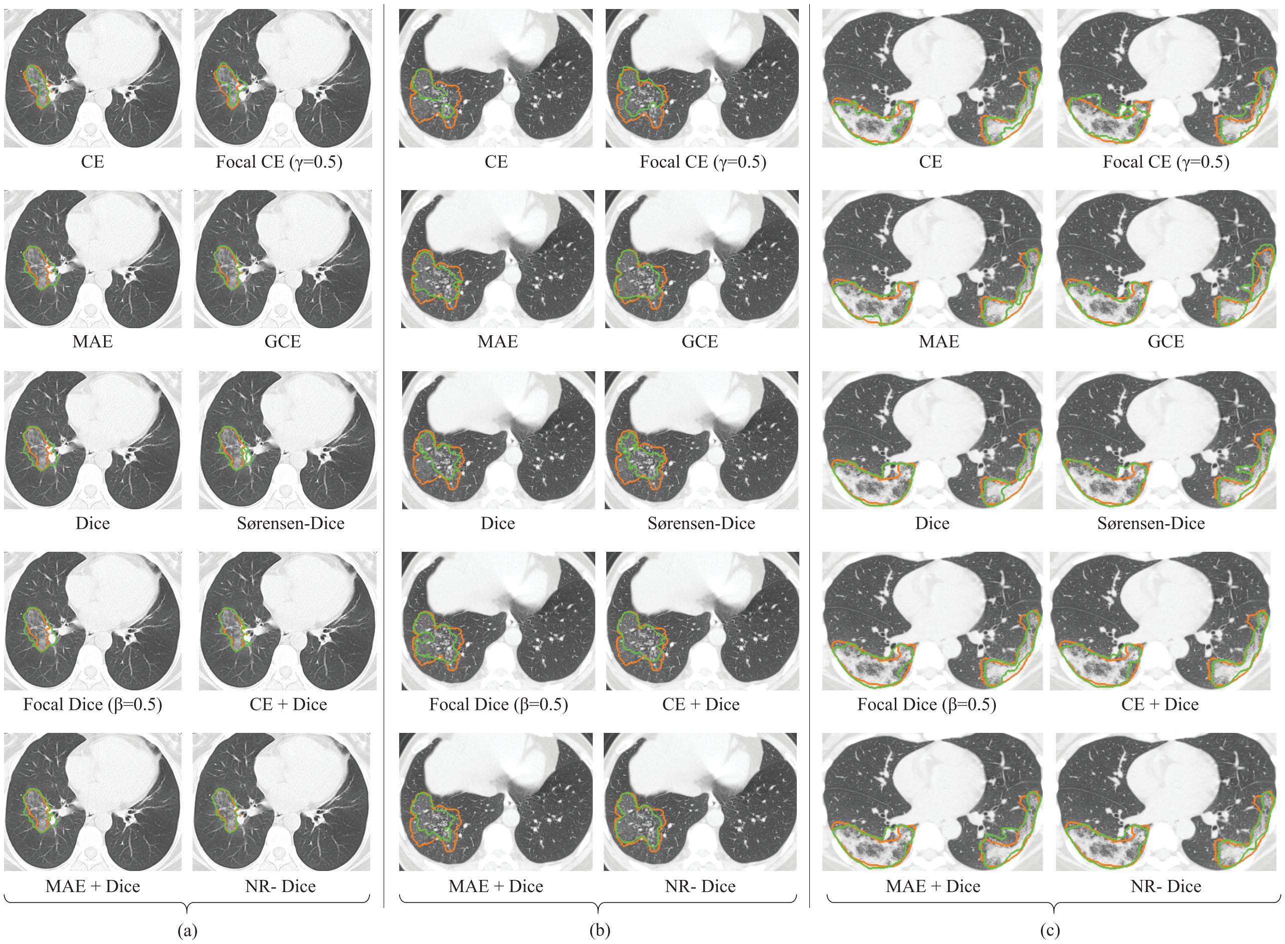

Fig. 6. Visual comparison of a segmentation model trained with different loss functions. The three rows are from group A, B and C, respectively. Green and orange curves show the segmentation results and the ground truth, respectively. The results are based on 2D nnU-Net [41] without using self-ensembling.

in dealing with noisy labels, and their performance was close to that of Dice loss L_{Dice} proposed by Milletari *et al.* [18]. Sørensen-Dice [44] and focal Dice loss [45] also achieved similar performance in average. $L_{\text{CE}}+L_{\text{Dice}}$ and $L_{\text{MAE}}+L_{\text{Dice}}$ achieved higher performance than L_{Dice}. Our $L_{\text{NR-Dice}}$ outperformed the others with average Dice, RVE and HD$_{95}$ values of 79.07%, 18.37% and 20.11 mm, respectively. Though $L_{\text{CE}} + L_{\text{Dice}}$ is comparable to $L_{\text{NR-Dice}}$ in terms of Dice and RVE scores, Table I shows their large difference in terms of HD$_{95}$ (23.10 mm VS 20.11 mm), which demonstrates that our $L_{\text{NR-Dice}}$ has more advantages in reducing the boundary error than $L_{\text{CE}} + L_{\text{Dice}}$.

Fig. 5 shows the performance of these loss functions on different groups of the test images. It can be observed that all the loss functions achieved the worst segmentation performance on group A (small lesions), and the best on group C (large lesions), indicating the difficulty of segmenting smaller lesions. Our $L_{\text{NR-Dice}}$ outperformed the other loss functions for all the three lesion groups. Fig. 6 shows a visual comparison of results obtained by these loss functions, where the three sub-figures are from group A, B and C, respectively. It can be observed that our $L_{\text{NR-Dice}}$ has a better performance than the others, which demonstrates the superiority of $L_{\text{NR-Dice}}$ for learning from noisy labels.

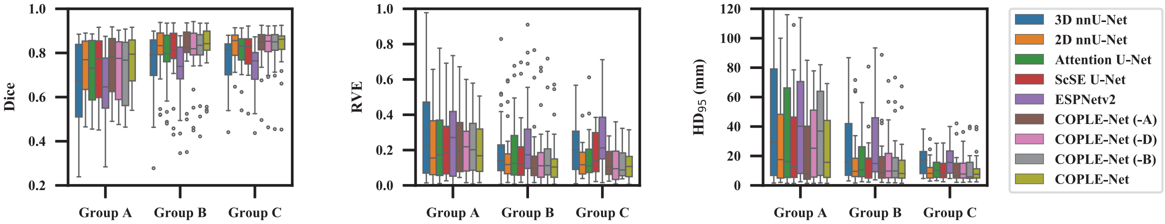

Fig. 7. Quantitative comparison of different networks for segmentation of COVID-19 pneumonia lesions at different scales. Our noise-robust Dice loss $L_{\text{NR-Dice}}$ was used for training.

TABLE II

QUANTITATIVE EVALUATION OF DIFFERENT NETWORKS FOR SEGMENTATION. THE PROPOSED $L_{\text{NR-DICE}}$ WAS USED FOR TRAINING

	Dice (%)	RVE (%)	HD_{95} (mm)
3D nnU-Net [41]	70.35±18.69	25.41±24.73	32.25±40.73
2D nnU-Net [41]	79.07±12.88	18.37±17.43	20.11±25.37
Attention U-Net [36]	77.23±12.33	19.77±18.41	24.62±32.45
ScSE U-Net [46]	78.04±12.48	18.85±16.69	21.83±33.24
ESPNetv2 [47]	69.82±14.77	23.69±20.26	31.45±32.61
COPLE-Net (-A)	79.70±12.53	18.94±21.77	19.55±28.50
COPLE-Net (-D)	78.98±12.55	18.73±25.31	25.57±31.44
COPLE-Net (-B)	79.03±12.68	19.18±27.31	22.14±30.69
COPLE-Net	**80.29±11.14**	**17.72±23.40**	**18.72±27.26**

C. Comparison of Different Networks

We compared our proposed COPLE-Net with four state-of-the-art networks for semantic or medical image segmentation: 1) 3D nnU-Net [41] that is extended from 3D U-Net [48] with minor modifications and achieved top performance in several medical image segmentation tasks [37], [41]; 2) 2D version of nnU-Net [41]; 2) Attention U-Net [36] that uses a spatial attention gate to enable the network to focus more on the target region; 3) ScSE U-Net [46] that combines concurrent Spatial and Channel SE (ScSE) blocks with U-Net [16]; 4) ESPNetv2 [47] that is a light-weight, power efficient and general purpose CNN and achieved state-of-the-art performance for semantic segmentation. We used 2D versions of Attention U-Net and ScSE U-Net, and set their channel numbers in the same way as our COPLE-Net, i.e., 32 in the first block and doubled after each down-sampling. In addition, COPLE-Net was compared with three variants: COPLE-Net (-A), COPLE-Net (-D) and COPLE-Net (-B) that refer to COPLE-Net without using ASPP module, dual pooling and bridge layers, respectively. We trained these networks with our proposed noise-robust Dice loss $L_{\text{NR-Dice}}$ respectively.

Table II shows a quantitative comparison of these networks on the entire testing set. It can be observed that ESPNetv2 achieved the lowest Dice score in our COVID-19 pneumonia lesion segmentation task. This is mainly because that ESPNetv2 was designed for learning from large-scale RGB images, which have different features and intensity distributions from our CT images. The 3D nnU-Net achieved the worst performance in terms of RVE and HD_{95}, which is mainly because our images had a large range of slice thickness. Our COPLE-Net achieved the best performance among the compared networks. Attention U-Net and ScSE U-Net achieved lower HD_{95} values than 2D nnU-Net, but their Dice and RVE values were worse than those of 2D nnU-Net, demonstrating that it is difficult to beat the nnU-Net structure, as also shown by previous works [37], [41]. However, compared with 2D nnU-Net, our COPLE-Net improved the average Dice from 79.07% to 80.29% and reduced the average HD_{95} from 20.11 mm to 18.72 mm, respectively.

Fig 7 presents a quantitative comparison of these networks for each group of test images, which shows that our proposed COPLE-Net achieved the best performance for lesions at different scales. Fig. 8 shows a visual comparison of results obtained by these networks, where (a), (b) and (c) are from group A (small lesions), B (medium lesions) and C (large lesions), respectively. It demonstrates that COPLE-Net outperforms the others when dealing with lesions at different scales.

D. Results of Noise-Robust Adaptive Self-Ensembling

We further trained our COPLE-Net with the proposed adaptive self-ensembling framework. For ablation study, we started with the baseline of training our COPLE-Net using Dice loss [18], and then added self-ensembling with standard mean teacher [20], our adaptive teacher and adaptive student gradually, and finally combined our COPLE-Net, $L_{\text{NR-Dice}}$ and the adaptive self-ensembling together. The quantitative evaluation results are shown in Table III. It can be observed that compared with the baseline, using self-ensembling with standard mean teacher improved the average Dice score from 78.61% to 79.36%. The use of adaptive teacher and adaptive student led to further improvement of accuracy, respectively. Our adaptive self-ensembling combining adaptive teacher and adaptive student outperformed the above variants when trained with the same Dice loss, achieving an average Dice score of 80.09%. Finally, when combined with our $L_{\text{NR-Dice}}$, the entire proposed framework achieved average Dice of 80.72%, RVE of 15.96% and HD_{95} of 17.12 mm respectively, and the performance is significantly higher than that of the baseline (p-value < 0.05 according to paired t-test).

E. Comparison With Other Noise-Robust Methods

We also compared our proposed framework with two other methods to deal with noisy labels: 1) data re-weighting that treats samples with large training loss values as noisy labels to suppress, as suggested by [31]. Similarly to our adaptive teacher, we used the 90-percentile of the training loss in the last 100 steps as a threshold and ignored samples with loss values larger than that. 2) Label update [33] that uses an ensemble of five models trained with the initial annotations

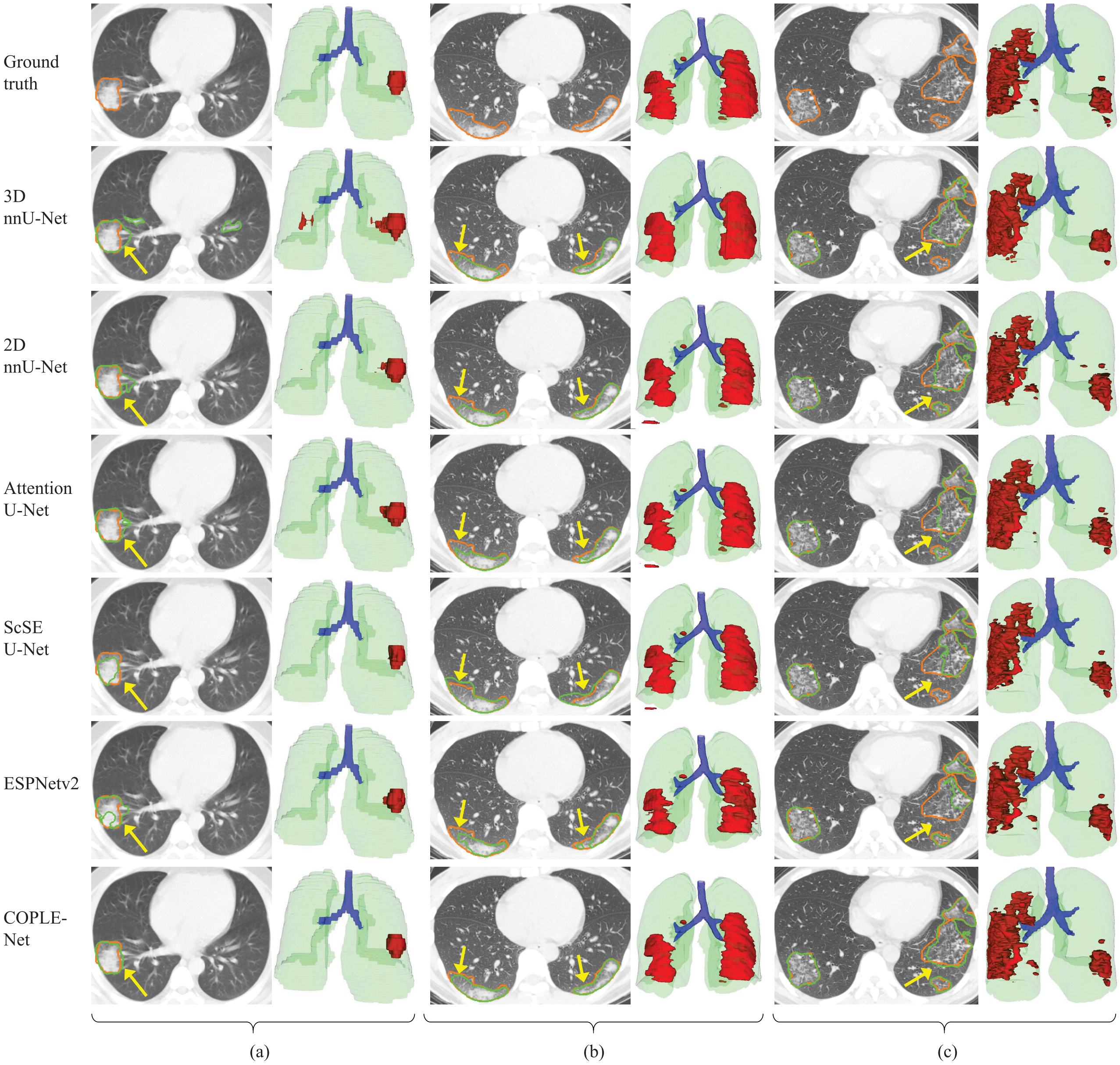

Fig. 8. Visual comparison of segmentation performance of different networks trained with our nosie-robust Dice loss $L_{\text{NR-Dice}}$, where (a), (b) and (c) are from group A, B and C respectively. For the 2D visualizations in odd columns, green and orange curves are segmentation results and the ground truth, respectively. Yellow arrows highlight the difference of local segmentation results. For the 3D visualizations in even columns, manual segmentation results of the lungs and the airway are shown for better visualization.

TABLE III
QUANTITATIVE EVALUATION OF OUR ADAPTIVE SELF-ENSEMBLING FOR LEARNING FROM NOISY ANNOTATIONS FOR COVID-19 PNEUMONIA LESION SEGMENTATION. THE FIRST METHOD IS A BASELINE OF TRAINING OUR COPLE-NET USING DICE LOSS WITHOUT SELF-ENSEMBLING. MT: SELF-ENSEMBLING WITH STANDARD MEAN TEACHER [20]. ADAT AND ADAS ARE OUR PROPOSED ADAPTIVE TEACHER AND ADAPTIVE STUDENT, RESPECTIVELY. * DENOTES SIGNIFICANT IMPROVEMENT FROM THE BASELINE (p-VALUE $<$ 0.05 BASED ON PAIRED T-TEST)

MT	AdaT	AdaS	$L_{\text{NR-Dice}}$	Dice (%)	RVE (%)	HD_{95} (mm)
				78.61±12.11	19.57±26.25	23.26±30.38
√				79.36±11.85	18.73±24.01	21.37±31.57
√	√			79.51±11.67	18.93±28.74	20.84±32.28
√		√		79.80±11.28*	18.42±27.51	20.07±30.64
√	√	√		80.09±10.97*	17.77±26.23*	19.32±31.79
√	√	√	√	**80.72±9.96***	**15.96±17.06***	**17.12±29.35**

to predict new labels for training images, then leverages the new labels to re-train the network. Both the data re-weighting and label update methods were implemented by COPLE-Net with Dice loss, and we refer to COPLE-Net trained with Dice loss as the baseline. They were compared with our framework of adaptive self-ensembling with COPLE-Net and $L_{\text{ND-Dice}}$.

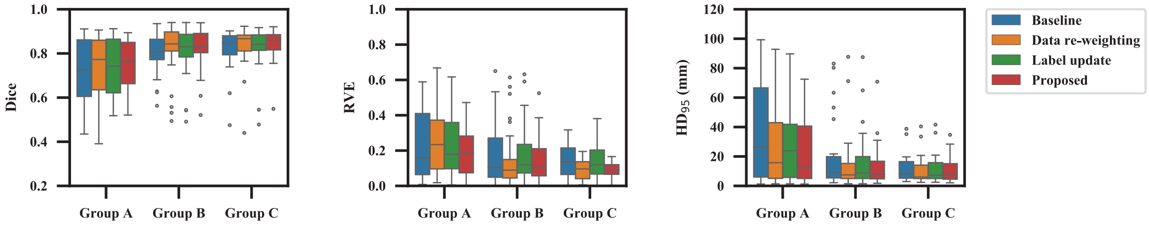

Fig. 9. Quantitative comparison of different training methods for segmentation of COVID-19 pneumonia lesions at different scales. The baseline is COPLE-Net trained with Dice loss.

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT TRAINING METHODS FOR COVID-19 PNEUMONIA LESION SEGMENTATION WITH NOISY LABELS. THE BASELINE IS COPLE-NET TRAINED WITH DICE LOSS. * DENOTES SIGNIFICANT IMPROVEMENT FROM THE BASELINE (p-VALUE < 0.05 BASED ON PAIRED T-TEST).

	Dice (%)	RVE (%)	HD_{95} (mm)
Baseline	78.61±12.11	19.57±26.25	23.26±30.38
Data re-weighting	79.97±12.34*	18.30±27.84	20.58±34.74
Label update	79.35±11.78	19.85±22.67	21.87±31.11
Proposed	**80.72±9.96***	**15.96±17.06***	**17.12±29.35**

Quantitative evaluation results of the different training methods are shown in Table IV. It can be observed that both data re-weighting and label update help to obtain better segmentation performance than the baseline, and our proposed framework outperforms these methods with the highest average Dice and the lowest average RVE and HD_{95}, respectively. Fig. 9 shows the performance of these methods on different lesion groups, which demonstrates the superiority of our method over the others when dealing with lesions at different scales.

IV. DISCUSSION AND CONCLUSION

Noisy labels exist widely for segmentation of large-scale 3D medical images. This can be either due to challenges for accurate annotation, such as low contrast, ambiguous boundaries and complex appearances of the target, or caused by low-cost inaccurate annotations such as annotations provided by non-experts, human-in-the-loop strategies [9] and some algorithms generating pseudo labels. Considering the difficulties of collecting absolutely clean labels for 3D segmentation tasks, learning from noisy labels can be a practical solution [13].

In contrast with existing noise-robust loss functions for classification tasks [14], [17], our noise-robust Dice loss function is specifically designed for segmentation tasks, dealing with the imbalance between foreground and background pixels and noisy labels at the same time. One advantage of our noise-robust Dice loss function is that it does not depend on a specific CNN and can be combined with different training strategies, such as a standard training process and the self-ensembling framework in our method. Note that the focal loss [43] uses a modulating factor $(1 - p_i)^\gamma$ (e.g., $\gamma = 2$) to weight the cross entropy loss and emphasize training samples with incorrect predictions. However, in our senario of training with noisy labels, incorrect predictions tend to be related to noisy labels. Assigning higher weights to harder samples in focal loss would lead the model to be corrupted by noisy labels, as shown in Table I. In fact, the Dice loss in Eq. (1) proposed by Milletari *et al.* [18] also tends to highlight hard samples due to the MSE term. In contrast, our proposed $\mathcal{L}_{\text{NR-Dice}}$ with $\gamma = 1.5$ neither ignores hard samples nor assigns too high weights to these samples, which could lead to higher robustness against noisy labels.

Our adaptive self-ensembling is extended from previous self-ensembling for semi-supervised learning [20], [21], and our adaptive mechanisms make it more suitable for dealing with noisy labels. Differently from previous works using simulated noisy labels for experiments [13], [34], we used noisy labels in real practice to validate the effectiveness of our proposed method. However, it would be of interest to investigate the performance of our framework under a larger range of noise levels in the future.

In conclusion, we deal with learning from noisy labels for COVID-19 pneumonia lesion segmentation from CT images where clean labels are difficult and expensive to acquire. We first introduce a novel noise-robust Dice loss function $\mathcal{L}_{\text{NR-Dice}}$, which is a generalization of MAE loss [14] that is robust against noisy labels and Dice loss [18] that is insensitive to foreground-background imbalance. We then propose a novel COVID-19 Pneumonia Lesion segmentation network (COPLE-Net) that combines several light-weight modules for better performance. They are combined with a novel adaptive self-ensembling framework, where we introduce an adaptive teacher and an adaptive student to suppress the effect of noisy labels on training. Experimental results showed that our proposed $\mathcal{L}_{\text{NR-Dice}}$ outperformed existing noise-robust loss functions, and the COPLE-Net achieved higher performance than state-of-the-art CNNs for medical image segmentation. What's more, our adaptive self-ensembling framework significantly outperformed a standard training process and surpassed other noise-robust methods in the scenario of learning from noisy labels for COVID-19 pneumonia lesion segmentation.

REFERENCES

[1] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *New England J. Med.*, vol. 382, pp. 727–733, Jan. 2020.

[2] D. Benvenuto *et al.*, "The global spread of 2019-nCoV: A molecular evolutionary analysis," *Pathogens Global Health*, vol. 114, no. 2, pp. 64–67, 2020.

[3] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, early access, Apr. 16, 2020, doi: 10.1109/RBME.2020.2987975.

[4] WHO. (2020). *Coronavirus Disease 2019 (COVID-19) Situation Report—81*. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200410-sitrep-81-covid-19.pdf?sfvrsn=ca96eb84_2

[5] M.-Y. Ng *et al.*, "Imaging profile of the COVID-19 infection: Radiologic findings and literature review," *Radiol., Cardiothoracic Imag.*, vol. 2, no. 1, Feb. 2020, Art. no. e200034.

[6] L. Huang *et al.*, "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach," *Radiol., Cardiothoracic Imag.*, vol. 2, no. 2, Apr. 2020, Art. no. e200075.

[7] J. Lei, J. Li, X. Li, and X. Qi, "CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia," *Radiology*, vol. 295, no. 1, 2020, Art. no. 200236.

[8] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, Mar. 2020, Art. no. 200905, doi: 10.1148/radiol.2020200905.

[9] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, *arXiv:2003.04655*. [Online]. Available: http://arxiv.org/abs/2003.04655

[10] 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, Apr. 2020, Art. no. e200082.

[11] H. Wang *et al.*, "Recognizing brain states using deep sparse recurrent neural network," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1058–1068, Apr. 2019.

[12] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, Jun. 2017.

[13] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," 2019, pp. 1–17, *arXiv:1912.02911*. [Online]. Available: http://arxiv.org/abs/1912.02911

[14] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI*, 2017, pp. 1919–1925.

[15] H. Zhu, J. Shi, and J. Wu, "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation," in *Proc. MICCAI*, 2019, pp. 576–584.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[17] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. NIPS*, 2018, pp. 8778–8788.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. ICDV*, 2016, pp. 565–571.

[19] Z. Mirikharaji, Y. Yan, and G. Hamarneh, "Learning to segment skin lesions from noisy annotations," in *Proc. MICCAI MILID Workshop*, 2019, pp. 207–215.

[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NIPS*, 2017, pp. 1195–1204.

[21] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. MICCAI*, 2019, pp. 605–613.

[22] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *Proc. ICLR*, 2018, pp. 1–13.

[23] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, Jul. 2019.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. MICCAI Workshop DLMIA*, in Lecture Notes in Computer Science, vol. 11045, 2018, pp. 3–11.

[25] S. Jin *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks," *medRxiv*, pp. 1–22, Mar. 2020, doi: 10.1101/2020.03.19.20039354.

[26] J. Chen *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: A prospective study," *medRxiv*, pp. 1–27, Feb. 2020, doi: 10.1101/2020.02.25.20021568.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "SELF: Learning to filter noisy labels with self-ensembling," in *Proc. ICLR*, 2020, pp. 1–15.

[29] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. ICML*, vol. 10, 2018, pp. 6900–6909.

[30] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1280–1283.

[31] Y. Shen and S. Sanghavi, "Learning with bad training data via iterative trimmed loss minimization," in *Proc. ICML*, 2019, pp. 10075–10097.

[32] K. Lee, S. Yun, K. Lee, H. Lee, B. Li, and J. Shin, "Robust inference via generative classifiers for handling noisy labels," in *Proc. ICML*, 2019, pp. 3763–3772.

[33] P. Ostyakov *et al.*, "Label denoising with large ensembles of heterogeneous neural networks," in *Proc. ECCV*, 2018, pp. 250–261.

[34] S. Min, X. Chen, Z.-J. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proc. AAAI*, vol. 33, 2019, pp. 4578–4585.

[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[36] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Proc. MIDL*, 2018, pp. 1–10.

[37] F. Isensee *et al.*, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," in *Bildverarbeitung für die Medizin 2019*. Wiesbaden, Germany: Springer Fachmedien Wiesbaden, 2019, p. 22.

[38] C. Huang, H. Han, Q. Yao, S. Zhu, and S. K. Zhou, "3D U^2-Net: A 3D universal U-Net for multi-domain medical image segmentation," in *Proc. MICCAI*, vol. 2, 2019, pp. 291–299.

[39] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4229–4238.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[41] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 234–244.

[42] Q. Duan *et al.*, "SenseCare: A research platform for medical image informatics and interactive 3D visualization," 2020, *arXiv:2004.07031*. [Online]. Available: https://arxiv.org/abs/2004.07031

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[44] L. Fidon *et al.*, "Generalised Wasserstein Dice score for imbalanced multi-class segmentation using holistic convolutional networks," in *Proc. Int. MICCAI Brainlesion Workshop*, 2017, pp. 64–76.

[45] P. Wang and A. C. S. Chung, "Focal Dice loss and image dilation for brain tumor segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045. Cham, Switzerland: Springer, 2018, pp. 119–127.

[46] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.

[47] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9190–9200.

[48] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.