Dual-Sampling Attention Network for Diagnosis of COVID-19 From Community Acquired Pneumonia

Xi Ouyang, Jiayu Huo, Liming Xia, Fei Shan, Jun Liu, Zhanhao Mo, Fuhua Yan, Zhongxiang Ding, Qi Yang, Bin Song, Feng Shi, Huan Yuan, Ying Wei, Xiaohuan Cao, Yaozong Gao, Dijia Wu, Qian Wang, and Dinggang Shen

***Abstract*— The coronavirus disease (COVID-19) is rapidly spreading all over the world, and has infected more than 1,436,000 people in more than 200 countries and territories as of April 9, 2020. Detecting COVID-19 at early stage is essential to deliver proper healthcare to the patients and also to protect the uninfected population. To this end, we develop a dual-sampling attention network to automatically diagnose COVID-19 from the community acquired pneumonia (CAP) in chest computed tomography (CT). In particular, we propose a novel online attention module with a 3D convolutional network (CNN) to focus on the infection regions in lungs when making decisions of diagnoses. Note that there exists imbalanced distribution of the sizes of the infection regions between COVID-19 and CAP, partially due to fast progress of COVID-19 after symptom onset. Therefore, we develop a dual-sampling strategy to mitigate the imbalanced learning. Our method is evaluated (to our best knowledge) upon the largest multi-center CT data for COVID-19 from 8 hospitals. In the training-validation stage, we collect 2186 CT scans from 1588 patients for a 5-fold cross-validation. In the testing stage, we employ another independent large-scale testing dataset including 2796 CT scans from 2057 patients. Results show that our algorithm can identify the COVID-19 images with the area under the receiver operating characteristic curve (AUC) value of 0.944, accuracy of 87.5%, sensitivity of 86.9%, specificity of 90.1%, and F1-score of 82.0%. With this performance, the proposed algorithm could potentially aid radiologists with COVID-19 diagnosis from CAP, especially in the early stage of the COVID-19 outbreak.**

***Index Terms*— COVID-19 Diagnosis, Online Attention, Explainability, Imbalanced Distribution, Dual Sampling Strategy.**

Manuscript received April 28, 2020; revised May 12, 2020; accepted May 12, 2020. Date of publication May 18, 2020; date of current version July 30, 2020. This work was supported in part by the Wuhan Science and Technology Program under Grant 2018060401011326, in part by the Hubei Provincial Novel Pneumonia Emergency Science and Technology Project under Grant 2020FCA021, in part by the Huazhong University of Science and Technology Novel Coronavirus Pneumonia Emergency Science and Technology Project under Grant 2020kfyXGYJ014, in part by the Novel Coronavirus Special Research Foundation of the Shanghai Municipal Science and Technology Commission under Grant 20441900600, in part by the Key Emergency Project of Pneumonia Epidemic of Novel Coronavirus Infection under Grant 2020sk3006, in part by the Emergency Project of Prevention and Control for COVID-19 of Central South University under Grant 60260005, in part by the National Natural Science Foundation of China under Grant 81871337 and Grant 6204100022, in part by the National Key Research and Development Program of China under Grant 2018YFC0116400, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 19QC1400600 and Grant 17411953300. *(Xi Ouyang, Jiayu Huo, Liming Xia, Fei Shan, Jun Liu, Zhanhao Mo, Fuhua Yan, Zhongxiang Ding, Qi Yang, and Bin Song contributed equally to this work.) (Corresponding authors: Qian Wang; Dinggang Shen.)*

Please see the Acknowledgment section of this article for the author affiliations.

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.2020.2995508

I. Introduction

THE disease caused by the novel coronavirus, or Coronavirus Disease 2019 (COVID-19) is quickly spreading globally. It has infected more than 1,436,000 people in more than 200 countries and territories as of April 9, 2020 [1]. On February 12, 2020, the World Health Organization (WHO) officially named the disease caused by the novel coronavirus as Coronavirus Disease 2019 (COVID-19) [2]. Now, the number of COVID-19 patients is dramatically increasing every day around the world [3]. Compared with the prior Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), COVID-19 has spread to more places and caused more deaths, despite its relatively lower fatality rate [4], [5]. Considering the pandemic of COVID-19, it is important to detect COVID-19 early, which could facilitate the slowdown of viral transmission and thus disease containment.

In clinics, real-time reverse-transcription–polymerase-chain-reaction (RT-PCR) is the golden standard to make a definitive diagnosis of COVID-19 infection [6]. However, the high false negative rate [7] and unavailability of RT-PCR assay in the early stage of an outbreak may delay the identification of potential patients. Due to the highly contagious nature of the virus, it then constitutes a high risk for infecting a larger population. At the same time, thoracic computed tomography (CT) is relatively easy to perform and can produce fast diagnosis [8]. For example, almost all COVID-19 patients have some typical radiographic features in chest CT, including ground-glass opacities (GGO), multifocal patchy consolidation, and/or interstitial changes with a peripheral distribution [9]. Thus chest CT has been recommended as a major tool for clinical diagnosis especially in the hard-hit region such as Hubei, China [6]. Considering the need for high-throughput screening by chest CT and the workload for radiologists especially in

the outbreak, we design a deep-learning-based method to automatically diagnose COVID-19 infection from the community acquired pneumonia (CAP) infection.

With the development of deep learning [11]–[15], the technology has a wide range of applications in medical image processing, including disease diagnosis [16], organ segmentation [17], etc. Convolutional neural network (CNN) [18], one of the most representative deep learning technology, has been applied to reading and analyzing CT images in many recent studies [19], [20]. For example, Koichiro et. al. use CNN for differentiation of liver masses on dynamic contrast agent–enhanced CT images [21]. Also, some studies focus on the diagnoses of lung diseases in chest CT, e.g., pulmonary nodules [22], [23] and pulmonary tuberculosis [24]. Although deep learning has achieved remarkable performance for abnormality diagnoses of medical images [16], [25], [26], physicians have concerns especially in the lack of model interpretability and understanding [27], which is important for the diagnosis of COVID-19. To provide more insight for model decisions, the class activation mapping (CAM) [28] and gradient-weighted class activation mapping (Grad-CAM) [29] methods have been proposed to produce localization heatmaps highlighting important regions that are closely associated with predicted results.

In this study, we propose a dual-sampling attention network to classify the COVID-19 and CAP infection. To focus on the lung, our method leverages a lung mask to suppress image context of none-lung regions in chest CT. At the same time, we refine the attention of the deep learning model through an online mechanism, in order to better focus on the infection regions in the lung. In this way, the model facilitates interpreting and explaining the evidence for the automatic diagnosis of COVID-19. The experimental results also demonstrate that the proposed online attention refinement can effectively improve the classification performance.

In our work, an important observation is that COVID-19 cases usually have more severe infection than CAP cases [30], although some COVID-19 cases and CAP cases do have similar infection sizes. To illustrate it, we use an established VB-Net toolkit [10] to automatically segment lungs and pneumonia infection regions on all the cases in our training-validation (TV) set (with details of our TV set provided in Section IV), and show the distribution of the ratios between the infection regions and lungs in Fig. 1. Note that the VB-Net toolkit is from previous work [10], which was trained on a different dataset. In this work, we directly use it to get the segmentation of the lung and infection masks. The proposed CNN model uses these masks as well as the intensity CT image as inputs for classification. We can see the imbalanced distribution of the infection size ratios in both COVID-19 and CAP data. In this situation, the conventional uniform sampling on the entire dataset to train the network could lead to unsatisfactory diagnosis performance, especially concerning the limited cases of COVID-19 with small infections and also the limited cases of CAP with large infections. To this end, we train the second network with the size-balanced sampling strategy, by sampling more cases of COVID-19 with small infections and also more cases of CAP with large infections within mini-batches. Finally, we apply ensemble learning to integrate the networks of uniform sampling and size-balanced sampling to get the final diagnosis results, by following the dual-sampling strategy.

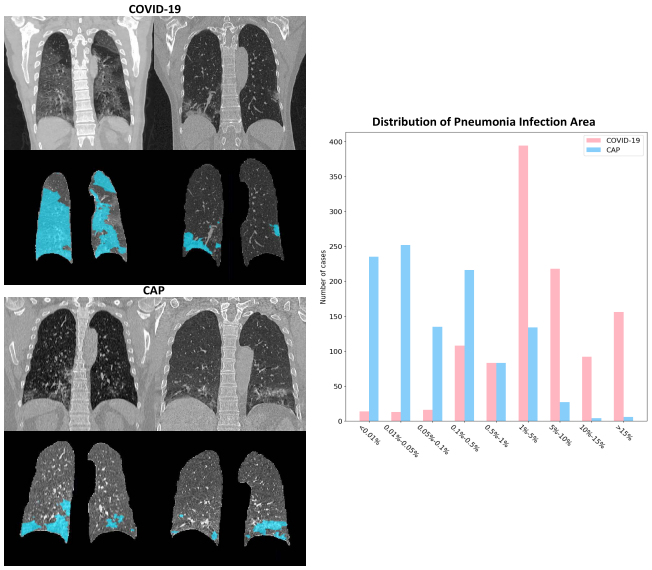

Fig. 1. Examples of CT images and infection segmentations of two COVID-19 patients (top left) and two CAP patients (bottom left), and the size distribution of the infection regions of COVID-19 and CAP in our training-validation set (right). The segmentation results of the lungs and infection regions are obtained from an established VB-Net toolkit [10]. The sizes of the infection regions are denoted by the volume ratios of the segmented infection regions and the whole lungs. Compared with CAP, the COVID-19 cases tend to have more severe infections in terms of the infection region sizes.

As a summary, the contributions of our work are in three-fold:

- We propose an online module to utilize the segmented pneumonia infection regions to refine the attention for the network. This ensures the network to focus on the infection regions and increase the adoption of visual attention for model interpretability and explainability.
- We propose a dual-sampling strategy to train the network, which further alleviates the imbalanced distribution of the sizes of pneumonia infection regions.
- To our knowledge, we have used the largest multi-center CT data in the world for evaluating automatic COVID-19 diagnosis. In particular, we conduct extensive cross-validations in a TV dataset of 2186 CT scans from 1588 patients. Moreover, to better evaluate the performance and generalization ability of the proposed method, a large independent testing set of 2796 CT scans from 2057 patients is also used. Experimental results demonstrate that our algorithm is able to identify the COVID-19 images with the area under the receiver operating characteristic curve (AUC) value of 0.944, accuracy of 87.5%, sensitivity of 86.9%, specificity of 90.1%, and F1-score of 82.0%.

II. RELATED WORKS

A. Computer-Assisted Pneumonia Diagnosis

Chest X-ray (CXR) is one of the firstline imaging modality to diagnose pneumonia, which manifests as increased

opacity [31]. The CNN networks have been successfully applied to pneumonia diagnosis in CXR images [16], [32]. As the release of the Radiological Society of North America (RSNA) pneumonia detection challenge [33] dataset, object detection methods (i.e., RetinaNet [34] and Mask R-CNN [35]) have been used for pneumonia localization in CXR images. At the same time, CT has been used as a standard procedure in the diagnosis of lung diseases [36]. An automated classification method has been proposed to use regional volumetric texture analysis for usual interstitial pneumonia diagnosis in high-resolution CT [37]. For COVID-19, GGO and consolidation along the subpleural area of the lung are the typical radiographic features of COVID-19 patients [9]. Chest CT, especially high-resolution CT, can detect small areas of ground glass opacity (GGO) [38].

Some recent works have focused on the COVID-19 diagnosis from other pneumonia in CT images [39]–[41]. It requires the chest CT images to identify some typical features, including GGO, multifocal patchy consolidation, and/or interstitial changes with a peripheral distribution [9]. Wang *et al.* [39] propose a 2D CNN network to classify between COVID-19 and other viral pneumonia based on manually delineated regions. Xu *et al.* [40] use a V-Net model to segment the infection region and apply a ResNet18 network for the classification. Song *et al.* [41] use a ResNet50 network to process all the slices of each 3D chest CT images to form the final prediction for each CT image. However, all these methods are evaluated in small datasets. In this paper, we have collected 4982 CT scans from 3645 patients, provided by 8 collaborative hospitals. To our best knowledge, it is the largest multi-center dataset for COVID-19 till now, which can prove the effectiveness of the method.

Note that, in the context of pneumonia diagnosis, lung segmentation is often an essential preprocessing step in analyzing chest CT images to assess pneumonia. In the literature, Alom *et al.* [42] utilize U-net, residual network and recurrent CNN for lung lesion segmentation. A convolutional-deconvolutional capsule network has also been proposed for pathological lung segmentation in CT images. In this paper, we use an established VB-Net toolkit for lung segmentation, which has been reported with high Dice similarity coefficient of $>98\%$ in evaluation [10]. Also, this VB-Net toolkit achieves Dice similarity coefficient of 92% between automatically and manually delineated pneumonia infection regions, showing the state-of-the-art performance [43]. For more related works, a recent review paper of automatic segmentation methods on COVID-19 could be found in [43].

B. Class Re-Sampling Strategies

For network training in the datasets with long-tailed data distribution, there exist some problems for the universal paradigm to sample the entire dataset uniformly [45]. In such datasets, some classes contain relatively few samples. The information of these cases may be ignored by the network if applying uniform sampling. To address this, some class re-sampling strategies have been proposed in the literature [46]–[50]. The aim of these methods is to adjust the numbers of the examples from different classes within mini-batches, which achieves better performance on the long-tailed dataset. Generally, class re-sampling strategies could be categorized into two groups, i.e., over-sampling by repeating data for minority classes [46]–[48] and under-sampling by randomly removing samples to make the number of each class to be equal [47], [49], [50]. The COVID-19 data is hard to collect and precious, so abandoning data is not a good choice. In this study, we adapt the over-sampling strategies [46] on the COVID-19 with small infections and also CAP with large infections to form a size-balanced sampling method, which can better balance the distribution of the infection regions of COVID-19 and CAP cases within mini-batches. However, over-sampling may lead to over-fitting upon these minority classes [51], [52]. We thus propose the dual-sampling strategy to integrate results from the two networks trained with uniform sampling and size-balanced sampling, respectively.

C. Attention Mechanism

Attention mechanism has been widely used in many deep networks, and can be roughly divided into two types: 1) activation-based attention [53]–[55] and 2) gradient-based attention [28], [29]. The activation-based attention usually serves as an inserted module to refine the hidden feature maps during the training, which can make the network to focus on the important regions. For the activation-based attention, the channel-wise attention assigns weights to each channel in the feature maps [55] while the position-wise attention produces heatmaps of importance for each pixel of the feature maps [53], [54]. The most common gradient-based attention methods are CAM [28] and Grad-CAM [29], which reveal the important regions influencing the network prediction. These methods are normally conducted offline and provide a pattern of model interpretability during the inference stage. Recently, some studies [56], [57] argue that the gradient-based methods can be developed as an online module during the training for better localization. In this study, we extend the gradient-based attention to composing an online trainable component and the scenario of 3D input. The proposed attention module utilizes the segmented pneumonia infection regions to ensure that the network can make decisions based on these infection regions.

III. METHOD

The overall framework is shown in Fig. 2. The input for the network is the 3D CT images masked in lungs only. We use an established VB-Net toolkit [10] to segment the lungs for all CT images, and perform auto-contouring of possible infection regions as shown in Fig. 3. The VB-Net toolkit is a modified network that combines V-Net [58] with bottleneck layers to reduce and integrate feature map channels. The toolkit is capable of segmenting the infected regions as well as the lung fields, achieving Dice similarity coefficient of 92% between automatically and manually delineated infection regions [10]. By labeling all voxels within the segmented regions to 1, and the rest part to 0, we can get the corresponding lung mask and then input image by masking the original CT image with the corresponding lung mask.

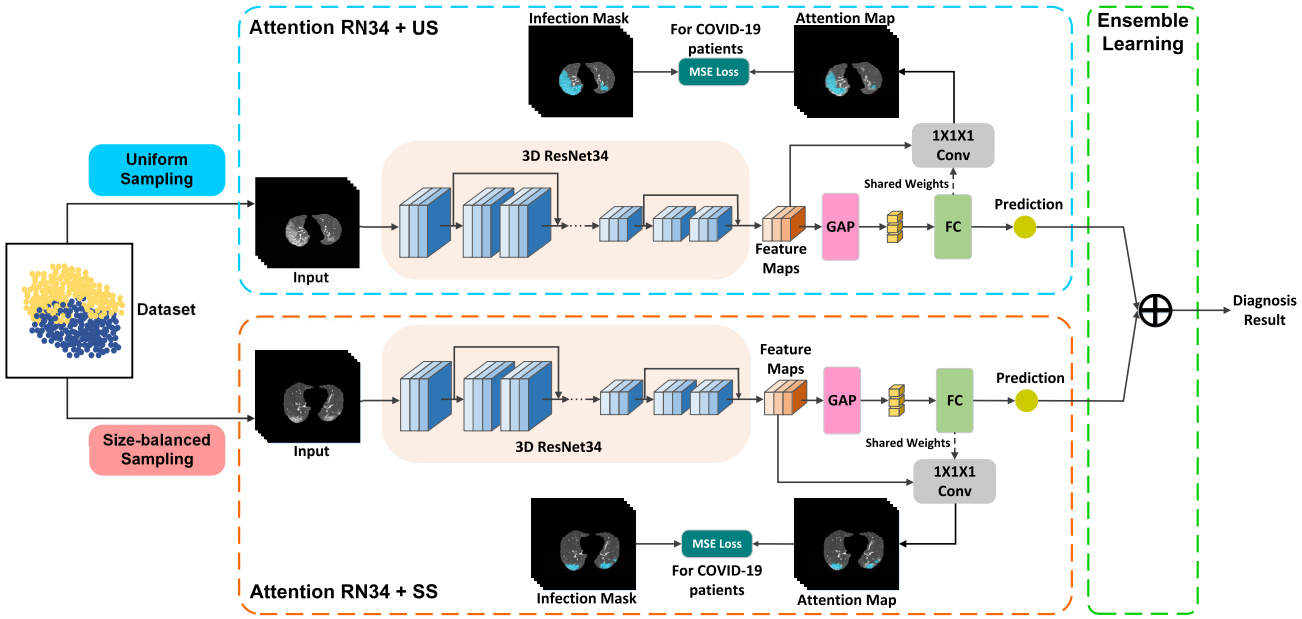

Fig. 2. Illustration of the pipeline of the proposed method, including two steps. 1) We train two 3D ResNet34 networks [44] with different sampling strategies. Also, the online attention mechanism generates attention maps during training, which refer to the segmented infection regions to refine the attention localization. 2) We use the ensemble learning to integrate predictions from the two trained networks. In this figure, "Attention RN34 + US" means the 3D ResNet34 (RN34) with attention module and uniform sampling (US) strategy, while "Attention RN34 + SS" means the 3D ResNet34 with attention module and size-balanced sampling (SS) strategy. "GAP" indicates the global average pooling layer, and "FC" indicates the fully connected layer. "$1 \times 1 \times 1$ Conv" refers to the convolutional layer with $1 \times 1 \times 1$ kernel, and takes the parameters from the fully connected layer as the kernel weights. "MSE Loss" refers to the mean square error function.

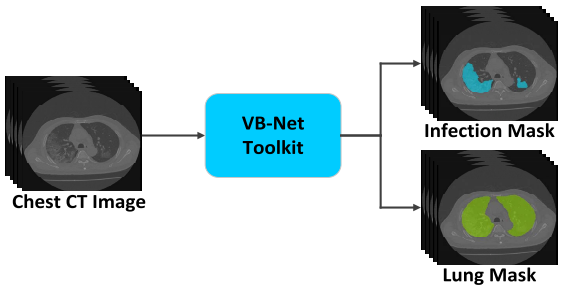

Fig. 3. The pneumonia infection region (upper right) and the lung segmentation (bottom right) from the VB-Net toolkit [10].

As shown in Fig. 2, the training pipeline of our method consists of two stages: 1) using different sampling strategies to train two 3D ResNet34 models [44] with the online attention module; 2) training an ensemble learning layer to integrate the predictions from the two models. The details of our method are introduced in the following sections.

A. Network

We use the 3D ResNet34 architecture [44] as the backbone network. It is the 3D extended version of residual network [13], which uses the 3D kernels in all the convolutional layers. In 3D ResNet34, we set the stride of each dimension as 1 in the last residual block instead of 2. This makes the resolution of the feature maps before the global average pooling (GAP) [59] operation into $1/16$ of the input CT image in each dimension. Compared with the case of downsampling the input image by a factor of 32 in each dimension in the original 3D ResNet34, it can greatly improve the quality of the generated attention maps based on higher-resolution feature maps.

B. Online Attention Module

To exhaustively learn all features that are important for classification, and also to produce the corresponding attention maps, we use an online attention mechanism of 3D class activation mapping (CAM). The key idea of CAM [28], [29], [56] is to back-propagate weights of the fully-connected layer onto the convolutional feature maps for generating the attention maps. In this study, we extend this offline operation to become an online trainable component for the scenario of 3D input. Let f denote the feature maps before the GAP operation and also w denote the weight matrix of the fully-connected layer. To make our attention generation procedure trainable, we use w as the kernel of a $1 \times 1 \times 1$ convolution layer and apply a ReLU layer [60] to generate the attention feature map A as:

$$A = \mathrm{ReLU}\left(\mathrm{conv}\left(f, w\right)\right), \tag{1}$$

where A has the shape $X \times Y \times Z$, and X, Y, Z is $1/16$ of corresponding size of the input CT images. Given the attention feature map A, we first upsample it to the input image size, then normalize it to have intensity values between 0 and 1, and finally perform sigmoid for soft masking [57], as follows:

$$T(A) = \frac{1}{1 + \exp(-\alpha(A - \beta))}, \tag{2}$$

where values of α and β are set to 100 and 0.4 respectively. $T(A)$ is the generated attention map of this online attention module, where A is defined in Eq. 1. During the training, the parameters in the $1 \times 1 \times 1$ convolution layer are always copied from the fully-connected layer and only updated by the binary cross entropy (BCE) loss for the classification task.

C. Size-Balanced Sampling

The main idea of size-balanced sampling is to repeat the data sampling for the COVID-19 cases with small infections and also the CAP cases with large infections in each mini-batch during training. Normally, we use the uniform sampling in the entire dataset for the network training (i.e., "Attention RN34 + US" branch in Fig. 2). Specifically, each sample in the training dataset is fed into the network only once with equal probability within one epoch. Thus, the model can review the entire dataset when maintaining the intrinsic data distribution. Due to the imbalance of the distribution of infection size, we train a second network via the size-balanced sampling strategy (i.e., "Attention RN34 + SS" branch). It aims to boost the sampling possibility of the small-infection-area COVID-19 and also large-infection-area CAP cases in each mini-batch. To this end, we split the data into 4 groups according to the volume ratio of the pneumonia infection regions and the lung: 1) small-infection-area COVID-19, 2) large-infection-area COVID-19, 3) small-infection-area CAP, and 4) large-infection-area CAP. For COVID-19, we define the cases that meet the criteria of <0.030 as small-infection-area COVID-19, and the rest as large-infection-area COVID-19. For CAP, we define the cases with the ratio >0.001 as large-infection-area CAP and the rest as small-infection-area CAP. We define the numbers of samples for the 4 groups as $[N^{covid}_{small}, N^{covid}_{large}, N^{cap}_{small}, N^{cap}_{large}]$. Then, inspired by the class-resampling strategy in [46], we define the weights $[W^{covid}_{small}, W^{covid}_{large}, W^{cap}_{small}, W^{cap}_{large}]$ for 4 groups as $[N^{covid}_{large}/N^{covid}_{small}, 1, 1, N^{cap}_{small}/N^{cap}_{large}]$. Since the numbers of small-infection-area COVID-19 and large-infection-area CAP are relatively small, the weights W^{covid}_{small} and W^{cap}_{large} are higher than 1. The values of these two weights are approximately 1.5 in each training fold. Then, the sampling possibilities for 4 groups are calculated by the weight of each group divided by the sum of all weights, W_{sum}. In a mini-batch, we randomly select a group according to the refined possibilities for each group $[W^{covid}_{small}/W_{sum}, 1/W_{sum}, 1/W_{sum}, W^{cap}_{large}/W_{sum}]$, and uniformly pick up a sample from the selected group. This strategy ensures to have more possibility to sample cases from the two groups of 1) COVID-19 with small infections and 2) CAP with large infections. We conduct the size-balanced sampling strategy for all mini-batches when training the "Attention RN34 + SS" model.

D. Objective Function

Two losses are used to train "Attention RN34 + US" and "Attention RN34 + SS" models, i.e., the classification loss L_c and the extra attention loss L_{ex} for COVID-19 cases, respectively. We adopt the binary cross entropy as constrain for the COVID-19/CAP classification loss L_c. For the COVID-19 cases, given the pneumonia infection segmentation mask M, we can use them to directly refine the attention maps from our model and L_{ex} is thus formulated as:

$$L_{ex} = \frac{\sum_{ijk} (T(A_{ijk}) - M_{ijk})^2}{\sum_{ijk} T(A_{ijk}) + \sum_{ijk} M_{ijk}}, \tag{3}$$

where $T(A_{ijk})$ is the attention map generated from our online attention module (Eq. 2), and i, j and k represent the $(i, j, k)^{th}$ voxel in the attention map. The proposed L_{ex} is modified from the traditional mean square error (MSE) loss, using the sum of regions of attention map $T(A_{ijk})$ and the corresponding mask M_{ijk} as an adaptive normalization factor. It can adjust the loss value dynamically according to the sizes of pneumonia infection regions. Then, the overall objective function for training "Attention RN34 + US" and "Attention RN34 + SS" models is expressed as:

$$L_{total} = L_c + \lambda L_{ex}, \tag{4}$$

where λ is a weight factor for the attention loss. It is set to 0.5 in our experiments. For the CAP cases, only the classification loss L_c is used for model training.

E. Ensemble Learning

The size-balanced sampling method could gain more attention on the minority classes and remedy the infection area bias in COVID-19 and CAP patients. A drawback is that it may suffer from the possible over-fitting of these minority classes. In contrast, the uniform sampling method could learn feature representation from the original data distribution in a relatively robust way. Taking the advantages of both sampling methods, we propose a dual-sampling method via an ensemble learning layer, which gauges the weights for the prediction results produced by the two models.

After training the two models with different sampling strategies, we use an ensemble learning layer to integrate the predictions from two models into the final diagnosis result. We combine the prediction scores with different weights for different ratios of the pneumonia infection regions and the lung:

$$P_{final} = wP_{US} + (1 - w)P_{SS}, \tag{5}$$

where, w is the weight factor. In our experiment, it is set to 0.35 for the case where the ratio meets the criterion <0.001 or >0.030, and 0.96 for the rest cases. The factor values are determined with a hyperparameter search on the TV set. Then, P_{final} is the final prediction result of the dual-sampling model. As presented in Eq. 5, the dual-sampling strategy combines the characteristics of uniform sampling and size-balanced sampling. For the minority classes, i.e., COVID-19 with small infections as well as CAP with large infections, we assign extra weights to the "Attention RN34 + SS" model. For the rest cases, more weights are assigned to the "Attention RN34 + US" model.

TABLE I
DEMOGRAPHIC OF THE TRAINING-VALIDATION (TV) DATASET AND TEST DATASET

Characteristics	TV set	Test set
No. (images (patients))		
COVID-19	1094 (960)	2295 (1605)
CAP	1092 (628)	501 (452)
Total	2186 (1588)	2796 (2057)
Median age in years (range)		
COVID-19	50.0 (14-89)	50.0 (8-95)
CAP	57.0 (12-94)	42.0 (15-98)
Total	53.0 (12-94)	49.0 (8-98)
Female/Male		
COVID-19	479/481	800/805
CAP	322/306	255/197
Total	801/787	1055/1002

IV. EXPERIMENTAL RESULTS

A. Dataset

In this study, we use a large multi-center CT data for evaluating the proposed method in diagnosis of COVID-19. In particular, we have collected a total of 4982 (<2mm) chest CT images from 3645 patients, including 3389 COVID-19 CT images and 1593 CAP CT images. All recruited COVID-19 patients were confirmed by RT-PCR test. Here, the images were provided by the Tongji Hospital of Huazhong University of Science and Technology, Shanghai Public Health Clinical Center of Fudan University, the Second Xiangya Hospital of Central South University, the Third Hospital of Jilin University, Ruijin Hospital of Shanghai Jiao Tong University School of Medicine, Hangzhou First People's Hospital of Zhejiang University, the Beijing Chaoyang Hospital of Capital Medical University, and Sichuan University West China Hospital. According to the data collection dates, we separate them into two datasets. The first dataset (TV dataset) is used for training and cross-validation, which includes 1094 COVID-19 images and 1092 CAP images. The second dataset serves for independent testing, including 2295 COVID-19 images and 501 CAP images. Note that the split is done on patient level, which means the images of the same subject are kept in the same group of training or testing. More details are shown in Table I.

Thin-slice chest CT images are used in this study with the CT thickness ranging from 0.625 to 1.5mm. CT scanners include uCT 780 from UIH, Optima CT520, Discovery CT750, LightSpeed 16 from GE, Aquilion ONE from Toshiba, SOMATOM Force from Siemens, and SCENARIA from Hitachi. Scanning protocol includes: 120 kV, with breath hold at full inspiration. All CT images are anonymized before sending them for conducting this research project. The study is approved by the Institutional Review Board of participating institutes. Written informed consent is waived due to the retrospective nature of the study.

B. Image Pre-Processing

Data are pre-processed in the following steps before feeding them into the network. First, we resample all CT images and the corresponding masks of lungs and infection regions to the same spacing (0.7168mm, 0.7168mm, 1.25mm for the x, y, and z axes, respectively) for the normalization to the same voxel size. Second, we down-sample the CT images and segmentation masks into the approximately half sizes considering efficient computation. To avoid morphological change in down-sampling, we use the same scale factor in all three dimensions and pad zeros to ensure the final size of $138 \times 256 \times 256$. We should emphasize that our method is capable of handling full-size images. Third, we conduct "window/level" (window: 1500, level: -600) scaling in CT images for contrast enhancement. We truncate the CT image into the window $[-1350, 150]$, which sets the intensity value above 150 to 150, and below -1350 to -1350. Finally, following the standard protocol of data pre-processing, we normalize the voxel-wise intensities in the CT images to the interval [0, 1].

C. Training Details and Evaluation Methods

We implement the networks in PyTorch [61], and use NVIDIA Apex for less memory consumption and faster computation. We also use the Adam [62] optimizer with momentum set to 0.9, a weight decay of 0.0001, and a learning rate of 0.0002 that is reduced by a factor of 10 after every 5 epochs. We set the batch size as 20 during the training. In our experiments, all the models are trained from scratch. In the TV set, we conduct 5-fold cross-validation. In each fold, the model is evaluated on the validation set in the end of each training epoch. The best checkpoint model with the best evaluation performance within 20 epochs is used as the final model and then evaluated on the test set. All the models are trained in 4 NVIDIA TITAN RTX graphics processing units, and the inference time for one sample is approximately 4.6s in one NVIDIA TITAN RTX GPU. For evaluating, we use five different metrics to measure the classification results from the model: area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1-score. AUC represents degree or measure of separability. In this study, we calculated the accuracy, sensitivity, specificity, and F1-score at the threshold of 0.5.

D. Results

First, we conduct 5-fold cross-validation on the TV set. The experimental results are shown in Table II, which combines the results of all 5 validation sets. The receiver operating characteristic (ROC) curve is also shown in Fig. 4(A). We can see that the models with the proposed attention refinement technique can improve the AUC and sensitivity scores. At the same time, we can see that "Attention RN34 + DS" achieves the highest performance in AUC, accuracy, sensitivity, and F1-score, when combining the two models with different sampling strategies. As for the specificity, the performance of the dual-sampling method is a little bit lower than that of ResNet34 with uniform sampling.

We further investigate the generalization capability of the model by deploying the five trained models of five individual folds on the independent testing dataset. From Fig. 4(B-F), we can see that the trained model of each

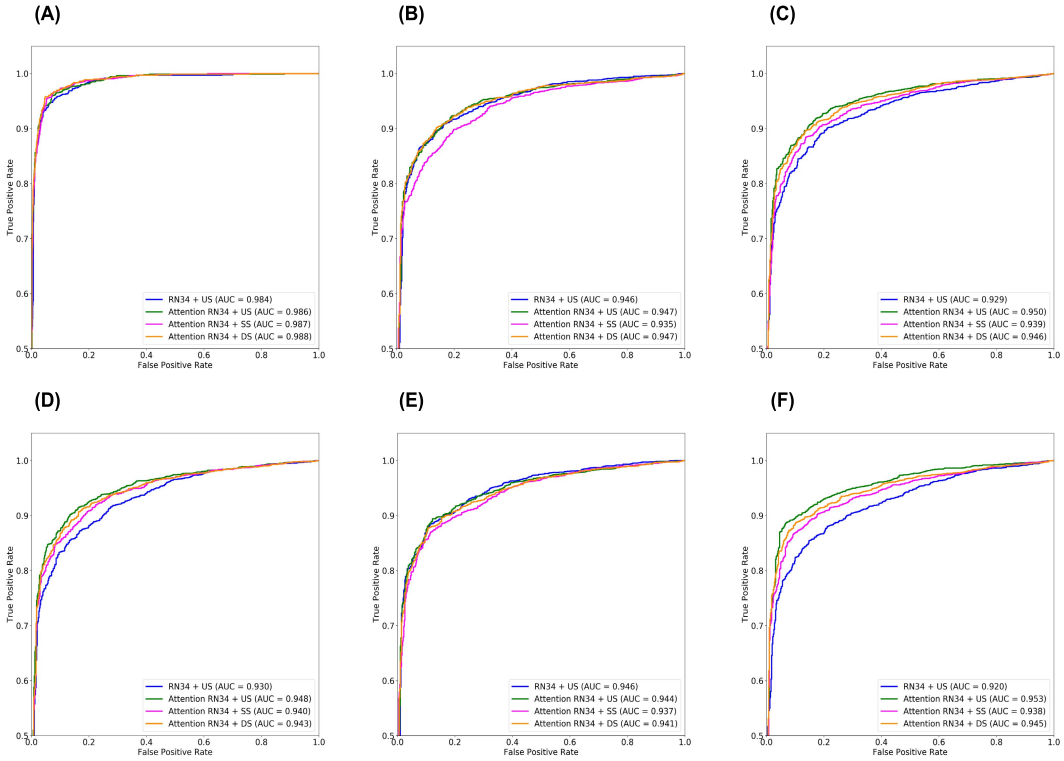

Fig. 4. ROC curves of the TV set and the test set. (A) ROC curves of TV set for 5 folds. (B) ROC curve of test set by using the model from TV set fold 1. (C) ROC curve of test set by using the model from TV set fold 2. (D) ROC curve of test set by using the model from TV set fold 3. (E) ROC curve of test set by using the model from TV set fold 4. (F) ROC curve of test set by using the model from TV set fold 5.

TABLE II

COMPARASION OF CLASSIFICATION RESULTS OF DIFFERNET MODELS ON THE TV SET AND TEST SET (RN34: 3D RESNET34; US: UNIFORM SAMPLING; SS: SIZE-BALANCED SAMPLING; DS: DUAL-SAMPLING). THE RESULTS OF AUC, ACCURACY, SENSITIVITY, SPECIFICITY AND F1-SCORE ARE PRESENT IN THIS TABLE. THE RESULTS ON TV SET ARE THE COMBINED RESULTS OF 5 VALIDATION SETS. FOR RESULTS ON TEST SET, WE SHOW MEAN±STD (STANDARD DEVIATION) SCORES OF FIVE TRAINED MODELS OF EACH TRAINING-VALIDATION FOLD

Results		TV set	Test set
AUC	RN34 + US	0.984	0.934±0.011
	Attention RN34 + US	0.986	**0.948±0.003**
	Attention RN34 + SS	0.987	0.938±0.002
	Attention RN34 + DS	**0.988**	0.944±0.003
Accuracy	RN34 + US	0.945	0.859±0.013
	Attention RN34 + US	0.947	**0.879±0.012**
	Attention RN34 + SS	0.951	0.869±0.008
	Attention RN34 + DS	**0.954**	0.875±0.009
Sensitivity	RN34 + US	0.931	0.856±0.029
	Attention RN34 + US	0.941	**0.872±0.018**
	Attention RN34 + SS	0.953	0.868±0.020
	Attention RN34 + DS	**0.954**	0.869±0.016
Specificity	RN34 + US	**0.959**	0.870±0.071
	Attention RN34 + US	0.953	**0.907±0.029**
	Attention RN34 + SS	0.948	0.876±0.048
	Attention RN34 + DS	0.954	0.901±0.025
F1-score	RN34 + US	0.945	0.798±0.011
	Attention RN34 + US	0.947	**0.825±0.013**
	Attention RN34 + SS	0.951	0.811±0.004
	Attention RN34 + DS	**0.954**	0.820±0.008

fold achieves similar performance, implying consistent performance with different training data. Compared with the results on the TV set in Fig. 4(A), the AUC score of the models with the proposed attention module ("Attention RN34 + DS") on the independent test set drops from 0.988 to 0.944, while the AUC score of "RN34 + US" drops from 0.984 to 0.934. This indicates the strong robustness of our model, trained with our attention module, against possible over-fitting. The proposed attention module can also ensure that the decisions made by the model depend mainly on the infection regions, suppressing the contributions from the non-related parts in the images. All 501 CAP images in the test set are from a single site that was not included in the TV set. "Attention RN34 + US" and "Attention RN34 + DS" models achieves ≥90.0% in specificity for these images. We can see that our algorithm maintains a great performance on the data acquired from different centers. In the next section, the effects of different sampling strategies are presented. In order to confirm whether there exist significant differences when using the proposed attention module or not, paired t-tests are applied. The p-values between "RN34 + US" and the three proposed methods are calculated. All the p-values are small than 0.01, implying that the proposed methods have significant improvements compared with "RN34 + US".

E. Detailed Analysis

To demonstrate the effectiveness in diagnosing pneumonia of different severity, we use the VB-Net toolkit [10] to get the lung mask and the pneumonia infection regions for all CT images. Based on the quantified volume ratio of pneumonia infection regions over the lung, we roughly divide the data

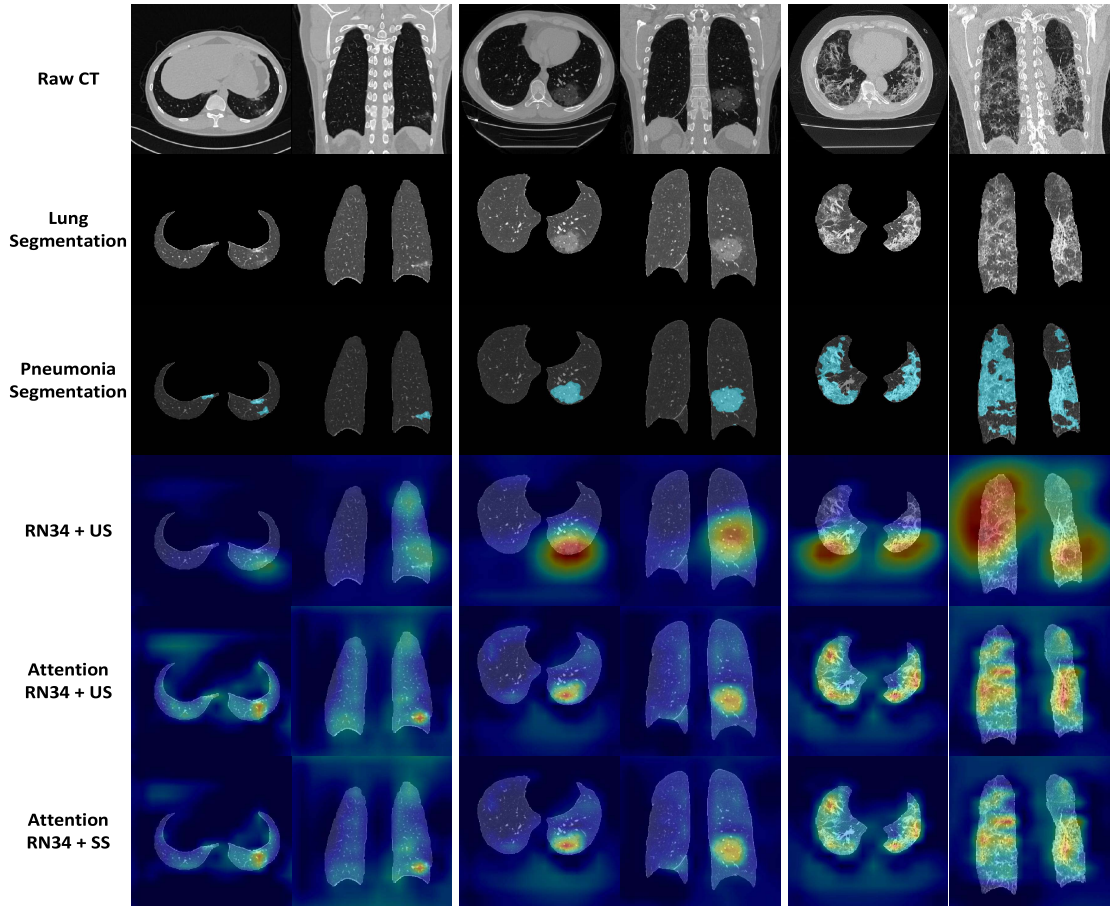

Fig. 5. Visualization results of our methods on three COVID-19 cases from small-infection group (<0.005), median-infection group (0.005 − 0.030) and large-infection group (>0.030) of the test set are shown from left to right, respectively. For each case, we show the visualization results in both axial view and coronal view. We show the original images (first row), and the segmentation results of the lung and pneumonia infection regions (2^{nd} and 3^{rd} rows) by the VB-Net tookit [10]. For the attention results, we show the Grad-CAM results of "RN34 + US" (4^{th} row), and the attention maps obtained by our proposed attention module of "Attention RN34 + US" and "Attention RN34 + SS" models (5^{th} and 6^{th} rows).

into 3 groups in both the TV set and the test set, according to the ratios, i.e., 1) <0.005, 2) 0.005 − 0.030, and 3) >0.030. As shown in Table III, most of COVID-19 images have high ratios (higher than 0.030), while most CAPs are lower than 0.005, which may indicate that the severity of COVID-19 is usually higher than that of CAP in our collected dataset. Furthermore, the classification results of COVID-19 is highly related with the ratio. In Table III, we can see that the sensitivity scores are relatively high for the high infected region group (>0.030), while the specificity scores are relatively low for the small infection region group (<0.005). This performance matches the nature of COVID-19 and CAP in the collected dataset.

As size-balanced sampling strategy ("Attention RN34 + SS") is applied in the training procedure, we can find that the sensitivity of the small infected region group (<0.005) increases from 0.534 to 0.569, compared with the case of using the uniform sampling strategy ("Attention RN34 + US"). And also the specificity of the large infected region group (>0.030) increases from 0.642 to 0.667. These results demonstrate that the size-balanced sampling strategy can effectively improve the classification robustness when the bias of the pneumonia area exists. However, if we only utilize the size-balanced sampling strategy in the training process, the sensitivity of the large infected region group (>0.030) will decrease from 0.965 to 0.955, and the specificity of the small infected region group (<0.005) will decrease from 0.933 to 0.896. This reflects that some advantages of the network may be sacrificed in order to achieve specific requirements. To achieve a dynamic balance between the two extreme conditions, we present the results using the ensemble learning with the dual-sampling model (i.e., "Attention RN34 + DS"). From the sensitivity and specificity in both small and large infected region groups, dual sampling strategy can preserve the classification ability obtained by uniform sampling, and slightly improve the classification performance of the COVID-19 cases in the small infected region group and the CAP cases in the large infected region group. Furthermore, the p-values between "Attention RN34 + US" and "Attention RN34 + DS" in both small-infected-region group (<0.005) and high-infected-region group (>0.030) are calculated. All the p-values are smaller than 0.01, which also proves the effectiveness and necessity of the dual sampling strategy.

Finally, we show typical attention maps obtained by our models (Fig. 5) trained in one fold. For comparison, we show the attention results of naive ReNset34 ("RN34 + US") in the same fold without both the online attention module

TABLE III

GROUP-WISE RESULTS ON TV SET AND TEST SET. BASED ON THE VOLUME RATIO OF PNEUMONIA REGIONS AND THE LUNG, THE DATA IS DIVIDED INTO 3 GROUPS: THE VOLUME RATIOS THAT MEET THE CRITERIA OF <0.005, $0.005-0.030$, AND >0.030, RESPECTIVELY

Results		TV set			Test set		
		<0.005	$0.005-0.030$	>0.030	<0.005	$0.005-0.030$	>0.030
No. of images	COVID-19	151	318	625	363	718	1214
	CAP	838	183	71	436	41	24
	Total No.	989	501	696	799	759	1238
AUC	RN34 + US	0.949	0.975	0.972	0.796±0.032	0.914±0.021	0.905±0.011
	Attention RN34 + US	0.958	0.974	0.986	**0.835±0.012**	**0.923±0.005**	0.906±0.016
	Attention RN34 + SS	0.958	**0.981**	0.986	0.816±0.007	0.919±0.004	0.906±0.014
	Attention RN34 + DS	**0.960**	**0.981**	**0.987**	0.830±0.011	0.919±0.004	**0.907±0.015**
Accuracy	RN34 + US	0.930	0.930	0.976	0.719±0.015	0.848±0.037	0.955±0.007
	Attention RN34 + US	0.932	0.930	**0.981**	0.752±0.017	**0.871±0.017**	**0.965±0.008**
	Attention RN34 + SS	0.938	**0.942**	0.974	0.747±0.006	0.858±0.018	0.955±0.009
	Attention RN34 + DS	**0.941**	**0.942**	**0.981**	**0.755±0.012**	0.859±0.016	0.962±0.007
Sensitivity	RN34 + US	0.675	0.925	0.995	0.514±0.093	0.851±0.042	0.962±0.007
	Attention RN34 + US	0.722	0.937	**0.997**	0.534±0.050	**0.875±0.021**	**0.972±0.008**
	Attention RN34 + SS	**0.815**	**0.953**	0.987	**0.569±0.061**	0.862±0.020	0.960±0.010
	Attention RN34 + DS	0.795	**0.953**	0.994	0.549±0.049	0.863±0.018	0.968±0.008
Specificity	RN34 + US	**0.976**	**0.940**	0.803	0.889±0.074	**0.810±0.078**	0.617±0.062
	Attention RN34 + US	0.970	0.918	0.845	**0.933±0.024**	0.785±0.090	0.642±0.037
	Attention RN34 + SS	0.961	0.923	0.859	0.896±0.051	0.785±0.047	**0.667±0.051**
	Attention RN34 + DS	0.968	0.923	**0.873**	0.926±0.025	0.790±0.051	0.650±0.037
F1-score	RN34 + US	0.853	0.926	0.928	0.698±0.022	0.643±0.035	0.663±0.018
	Attention RN34 + US	0.863	0.925	0.946	0.732±0.022	**0.662±0.015**	**0.702±0.026**
	Attention RN34 + SS	0.882	**0.938**	0.929	0.732±0.009	0.648±0.020	0.671±0.018
	Attention RN34 + DS	**0.885**	**0.938**	**0.947**	**0.737±0.017**	0.649±0.017	0.692±0.018

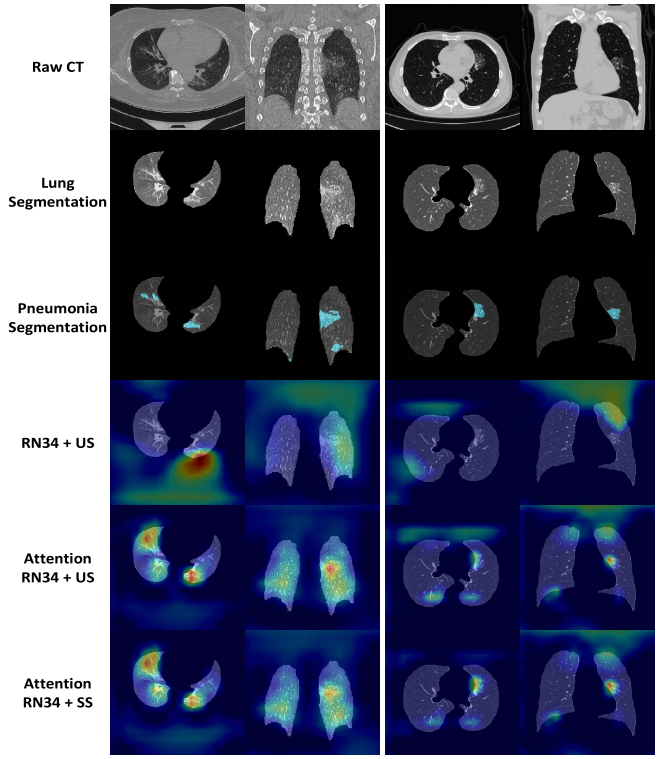

Fig. 6. Visualization results of two failure cases.

and the infection mask refinement, and perform the model explanation techniques (Grad-CAM [29]) to get the heatmaps for classification. We can see that the output of Grad-CAM roughly indicates the infection localization, yet sometimes appears far outside of the lung. However, the attention maps from our models ("Attention RN34 + US" and "Attention RN34 + SS") can reveal the precise locations of the infection. These conspicuous areas in attention maps are similar to the infection segmentation results, which demonstrates that the final classification results determined by our model are reliable and interpretable. The attention maps thus can be possibly used as the basis to derive the COVID-19 diagnosis in clinical practice.

F. Failure Analysis

We also show two failure cases in Fig. 6, where the COVID-19 cases are classified as CAP by mistake for all the models. As can be observed from the results shown in Fig. 5, the attention maps from all the models incorrectly get activated on many areas unrelated to pneumonia. "RN34 + US" model even generates many highlighted areas in the none-lung region instead of focusing on lungs. With the proposed attention constrain, the attention maps of "Attention RN34 + US" and "Attention RN34 + SS" have partially alleviated this problem. But still the visual evidence is insufficient to reach a final correct prediction.

V. DISCUSSION AND CONCLUSION

For COVID-19, it is important to get the diagnosis result at soon as possible. Although RT-PCR is the current ground truth to diagnose COVID-19, it will take up to days to get the final results and the capacity of the tests is also limited in many places especially in the early outbreak [8]. CT is shown as a powerful tool and could provide the chest scan results in several minutes. It is beneficial to develop an automatic diagnosis method based on chest CT to assist the COVID-19 screening. In this study, we explore a deep-learning-based method to perform automatic COVID-19 diagnosis from CAP in chest CT images. We evaluate our method by the largest

multi-center CT data in the world, to the best of our knowledge. To further evaluate the generalization ability of the model, we use independent data from different hospitals (not included in the TV set), achieving AUC of 0.944, accuracy of 87.5%, sensitivity of 86.9%, specificity of 90.1%, and F1-score of 82.0%. At the same time, to better understand the decision of the deep learning model, we also refine the attention module and show the visual evidence, which is able to reveal important regions used in the model for diagnosis. Our proposed method could be further extended for differential diagnosis of pneumonia, which can greatly assist physicians.

There also exist several limitations in this study. First, when longitudinal data becomes ready, the proposed model should be tested for its consistency tracking the development of the COVID-19 during the treatment, as considered in [63]. Second, although the proposed online attention module could largely improve the interpretability and explainability in COVID-19 diagnosis, in comparison to the conventional methods such as Grad-CAM, future work is still needed to analyze the correlation between these attention localizations with the specific imaging signs that are frequently used in clinical diagnosis. There also exist some failure cases that the visualization results do not appear correctly at the pneumonia infection regions, as shown in Fig. 6. This motivates us to further improve the attention module to better focus on the related regions and reduce the distortion from cofounding visual information to the classification task in the future research. Third, we also notice that the accuracy of the small-infection-area COVID-19 is not quite satisfactory. This indicates the necessity of combining CT images with clinical assessment and laboratory tests for precise diagnosis of early COVID-19, which will also be covered by our future work. The last but not least, the CAP cases used in this study do not include the subtype information, i.e., bacterial, fungal, and non-COVID-19 viral pneumonia. To assist the clinical diagnosis of pneumonia subtypes would also be beneficial.

To conclude, we have developed a 3D CNN network with both online attention refinement and dual-sampling strategy to distinguish COVID-19 from the CAP in the chest CT images. The generalization performance of this algorithm is also verified by the largest multi-center CT data in the world, to our best knowledge.

ACKNOWLEDGMENT

Xi Ouyang and Jiayu Huo are with the Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China, and also with Shanghai United Imaging Intelligence Company, Ltd., Shanghai 201807, China (e-mail: xi.ouyang@sjtu.edu.cn; jiayu.huo@sjtu.edu.cn).

Liming Xia is with the Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xialiming2017@outlook.com).

Fei Shan is with the Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai 200433, China (e-mail: shanfei_2901@163.com).

Jun Liu is with the Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha 410083, China, and also with the Department of Radiology, Quality Control Center, Changsha 410011, China (e-mail: junliu123@csu.edu.cn).

Zhanhao Mo is with the Department of Radiology, The Third Hospital, Jilin University, Changchun 130012, China (e-mail: mozhanhao@jlu.edu.cn).

Fuhua Yan is with the Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 201101, China (e-mail: yfh11655@rjh.com.cn).

Zhongxiang Ding is with the Department of Radiology, Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou 310027, China (e-mail: hangzhoudzx73@126.com).

Qi Yang is with the Beijing Chaoyang Hospital, Capital Medical University, Beijing 100069, China (e-mail: yangyangqiqi@gmail.com).

Bin Song is with the Department of Radiology, West China Hospital, Sichuan University, Chengdu 610017, China (e-mail: anicesong@vip.sina.com).

Feng Shi, Huan Yuan, Ying Wei, Xiaohuan Cao, Yaozong Gao, Dijia Wu, and Dinggang Shen are with the Department of Research and Development, Shanghai United Imaging Intelligence Company, Ltd., Shanghai 201807, China (e-mail: feng.shi@united-imaging.com; huan.yuan@united-imaging.com; ying.wei@united-imaging.com; xiaohuan.cao@united-imaging.com; yaozong.gao@united-imaging.com; dijia.wu@united-imaging.com; dinggang.shen@gmail.com).

Qian Wang is with the School of Biomedical Engineering, Institute for Medical Imaging Technology, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: wang.qian@sjtu.edu.cn).

REFERENCES

[1] *Coronavirus Disease 2019 (COVID-19): Situation Report, 80*, WHO, Geneva, Switzerland, 2020.

[2] *Who Director-General's Remarks at the Media Briefing on 2019-NCOV*, WHO, Geneva, Switzerland, Feb. 2020.

[3] *Coronavirus Disease (COVID-2019) Situation Reports*, WHO, Geneva, Switzerland, 2020.

[4] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention," *Jama*, vol. 323, pp. 1239–1242, Feb. 2020.

[5] E. Mahase, "Coronavirus: COVID-19 has killed more people than SARS and MERS combined, despite lower case fatality rate," *BMJ, Clin. Res. Ed.*, vol. 368, p. m641, 2020.

[6] Z. Y. Zu *et al.*, "Coronavirus disease 2019 (COVID-19): A perspective from China," *Radiology*, Feb. 2020, Art. no. 200490.

[7] J. F.-W. Chan *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster," *Lancet*, vol. 395, no. 10223, pp. 514–523, Feb. 2020.

[8] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, Feb. 2020, Art. no. 200642.

[9] M. Chung *et al.*, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, Apr. 2020, Art. no. 200230.

[10] 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

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[13] 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.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[15] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating CT image from MRI data using 3D fully convolutional networks," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 170–178.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

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

[18] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[19] T. Pang, S. Guo, X. Zhang, and L. Zhao, "Automatic lung segmentation based on texture and deep features of HRCT images with interstitial lung disease," *BioMed Res. Int.*, vol. 2019, pp. 1–8, Nov. 2019.

[20] B. Park, H. Park, S. M. Lee, J. B. Seo, and N. Kim, "Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks," *J. Digit. Imag.*, vol. 32, no. 6, pp. 1019–1026, Dec. 2019.

[21] K. Yasaka, H. Akai, O. Abe, and S. Kiryu, "Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study," *Radiology*, vol. 286, no. 3, pp. 887–896, Mar. 2018.

[22] P. Huang *et al.*, "Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study," *Radiology*, vol. 286, no. 1, pp. 286–295, 2018.

[23] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Med.*, vol. 25, no. 6, pp. 954–961, Jun. 2019.

[24] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017.

[25] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597.

[26] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2013, pp. 403–410.

[27] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[30] F. Shi *et al.*, "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," 2020, *arXiv:2003.09860*. [Online]. Available: http://arxiv.org/abs/2003.09860

[31] T. Franquet, "Imaging of community-acquired pneumonia," *J. Thoracic Imag.*, vol. 33, no. 5, pp. 282–294, 2018.

[32] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[33] Radiological Society of North America. (2018). *RSNA Pneumonia Detection Challenge*. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[34] 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.

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[36] M. O. Wielpütz, C. P. Heußel, F. J. Herth, and H.-U. Kauczor, "Radiological diagnosis in lung disease: Factoring treatment options into the choice of diagnostic modality," *Deutsches Ärzteblatt Int.*, vol. 111, no. 11, p. 181, 2014.

[37] A. Depeursinge *et al.*, "Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution CT," *Investigative Radiol.*, vol. 50, no. 4, p. 261, 2015.

[38] H. Macmahon *et al.*, "Guidelines for management of incidental pulmonary nodules detected on CT images: From the fleischner society 2017," *Radiology*, vol. 284, no. 1, pp. 228–243, Jul. 2017.

[39] S. Wang *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *MedRxiv*, Jan. 2020.

[40] X. Xu *et al.*, "Deep learning system to screen coronavirus disease 2019 pneumonia," 2020, *arXiv:2002.09334*. [Online]. Available: http://arxiv.org/abs/2002.09334

[41] Y. Song *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *MedRxiv*, Jan. 2020.

[42] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: http://arxiv.org/abs/1802.06955

[43] 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.

[44] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[45] G. Van Horn and P. Perona, "The devil is in the tails: Fine-grained classification in the wild," 2017, *arXiv:1709.01450*. [Online]. Available: http://arxiv.org/abs/1709.01450

[46] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," 2019, *arXiv:1912.02413*. [Online]. Available: http://arxiv.org/abs/1912.02413

[47] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[48] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 467–482.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[50] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[51] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9268–9277.

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[53] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[54] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[55] 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.

[56] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 10.

[57] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.

[58] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[59] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[60] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[61] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[63] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: Consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388–399, Apr. 2006.