Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images

Jun Wang, Yiming Bao, Yaofeng Wen, Hongbing Lu, Hu Luo, Yunfei Xiang, Xiaoming Li, Chen Liu, and Dahong Qian, *Senior Member, IEEE*

***Abstract*— We propose a conceptually simple framework for fast COVID-19 screening in 3D chest CT images. The framework can efficiently predict whether or not a CT scan contains pneumonia while simultaneously identifying pneumonia types between COVID-19 and Interstitial Lung Disease (ILD) caused by other viruses. In the proposed method, two 3D-ResNets are coupled together into a single model for the two above-mentioned tasks via a novel prior-attention strategy. We extend residual learning with the proposed prior-attention mechanism and design a new so-called prior-attention residual learning (PARL) block. The model can be easily built by stacking the PARL blocks and trained end-to-end using multi-task losses. More specifically, one 3D-ResNet branch is trained as a binary classifier using lung images with and without pneumonia so that it can highlight the lesion areas within the lungs. Simultaneously, inside the PARL blocks, prior-attention maps are generated from this branch and used to guide another branch to learn more discriminative representations for the pneumonia-type classification. Experimental results demonstrate that the proposed framework can significantly improve the performance of COVID-19 screening. Compared to other methods, it achieves a state-of-the-art result. Moreover, the proposed method can be easily extended to other similar clinical applications such as computer-aided detection and diagnosis of pulmonary nodules in CT images, glaucoma lesions in Retina fundus images, etc.**

***Index Terms*— COVID-19, pneumonia, residual learning, medical image classification, deep attention learning.**

Manuscript received April 27, 2020; revised May 9, 2020; accepted May 10, 2020. Date of publication May 15, 2020; date of current version July 30, 2020. This work was supported in part by the Biomedical Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University under Grant YG2020YQ17, in part by Chongqing Key Technology and Application Demonstration of Medical Imaging Depth Intelligent Diagnostic Platform under Grant cstc2018jszx-cyztzxX0017, in part by the National Key Research and Development Program of China under Grant 2018YFC0116402 and 2017YFC0113400. *(Jun Wang and Yiming Bao are co-first authors.) (Corresponding authors: Chen Liu; Dahong Qian.)*

Jun Wang, Yiming Bao, Yaofeng Wen, and Dahong Qian are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wjcy19870122@163.com; yiming.bao@sjtu.edu.cn; ericwene@126.com; dahong.qian@sjtu.edu.cn).

Hongbing Lu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China (e-mail: lhb@zju.edu.cn).

Hu Luo and Yunfei Xiang are with the Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital to Army Medical University, Chongqing 400038, China (e-mail: luohucy@163.com; 553288106@qq.com).

Xiaoming Li and Chen Liu are with the Department of Radiology, The First Affiliated Hospital to Army Medical University, Chongqing 400038, China (e-mail: 359261069@qq.com; liuchen@aifmri.com).

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.2020.2994908

I. Introduction

THE break of novel coronavirus pneumonia (COVID-19) has rapidly spread to most countries worldwide. To date (April 10, 2020), there have been 1,521,252 confirmed cases all around the world [1]. In clinical practice, compared to the real-time reverse-transcriptase polymerase chain reaction (RT-PCR), computed tomography (CT) is an effective tool for much faster screening of COVID-19. However, manual screening of COVID-19 from CT images is a time-consuming and labor-intensive task, since doctors must find the lesions from volumetric chest CT scans in a slice-by-slice manner. Besides, as shown in Fig. 1, the manifestations of COVID-19 in CT images are similar to other types of viral pneumonia, which makes it hard to manually distinguish COVID-19.

A reliable computer-aided diagnosis system (CADs) of COVID-19 is supposed to be useful in clinical practice, which can alleviate the doctor's workload and improve the detection efficiency. However, developing such a system is a challenging task, because the lesions of pneumonia in CT images have wide variations in appearances, sizes, and locations in the lung regions, as shown in Fig. 1. It seems difficult to design suitable methods to handle the complicated characteristics of the pneumonia lesions using just the classical image processing techniques or conventional machine learning methods [2]–[4] that rely on handcrafted descriptors.

In recent years, the development of deep convolutional neural networks (DCNNs) has led to a series of breakthroughs for image classification [5]–[8], object detection [9]–[14], and semantic segmentation [15]–[19] in the field of natural image processing. CNNs expert at automatically learning rich high-level discriminative semantic features from images, removing the need for handcrafted descriptors. These breakthroughs also revealed that deeper models can achieve superior performance [5]. Therefore, it is feasible that training very

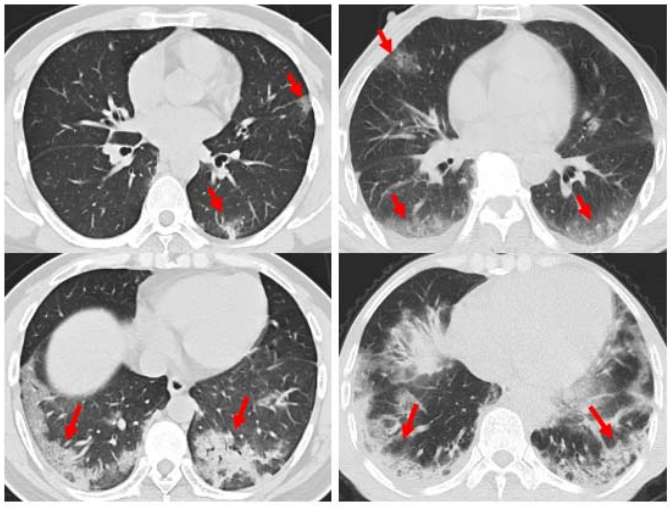


Fig. 1. Examples of (a) COVID-19 and (b) interstitial lung disease (ILD) in CT images as shown in the left and the right column, respectively. The main lesion regions are indicated with red arrows and it can be seen that the lesions have inter-class similarity and intra-class variation, which is one of the main challenges for the COVID-19 screening task.

deep CNN-based models to achieve promising performance in COVID-19 screening. Nowadays, it is very easy to construct robust deep models with more than 100 layers using residual learning blocks [5].

However, some challenges remain and should be addressed when applying the above-mentioned deep learning methods for the proposed COVID-19 screening task. First, it is very hard to collect sufficient samples together with accurately annotated labels to train very deep models in a short time, especially for object detection and segmentation models. Training of these models requires additional meticulous annotations that were manually labeled by experienced doctors. One example is that training most object detection models requires bounding boxes of desired targets, while training segmentation models requires lesion-aware masks. Labeling these annotations is also very time-consuming and impractical to doctors. Second, a volumetric CT scan has three dimensions. The computational cost and memory requirement both increase with 3D inputs. It is infeasible to train a very deep 3D CNN-based model due to the constraint of hardware resources. Third, a perplexing problem is the inter-class similarity and intra-class variation of pneumonia lesions, as demonstrated in Fig. 1. Finally, a lung image infected with pneumonia still contains a large part of non-lesion regions, which also have a wide and complicated variation of tissues. Obviously, the non-lesion regions have great negative impact on the performance. It is much more complicated than detecting objects of scenes in natural images.

To address the above-mentioned issues, we propose a novel multi-task prior-attention residual learning strategy for one-stage lesion-aware COVID-19 screening in CT images. It exhibits the following appealing properties:

(1) Two 3D-ResNet based sub-networks are integrated into a single model for pneumonia detection and its type-classification. The sub-network for the type-classification task is implemented as a binary classifier and it can identify COVID-19 from interstitial lung disease (ILD) caused by other viruses. Besides, the sub-network for the detection task is also designed as a binary classifier that can predict whether or not a given CT scan contains pneumonia. Compared to object detection or segmentation methods, the proposed method (that relies on only classification models) is much easier to implement, because it requires only weak image-level labels and fewer hyper-parameters at the training stage. Training models which use only image-level labels make it possible to collect relatively sufficient samples in a short time.

(2) Inspired by some recent advances of deep attention learning mechanisms [21]–[24], especially by the self-attention residual learning for state-of-the-art skin lesion classification [24], we designed a "prior-attention" mechanism in the proposed models. Many works [25], [26] have demonstrated that a DCNN model trained for a classification task has a remarkable localization ability that can highlight the discriminative regions in images, despite being trained with only image-level labels. Since the proposed sub-network for the detection task is designed as a binary classifier and trained using CT scans with and without pneumonia, it is supposed to have the ability to provide lesion-attention information. Therefore, we fully use its hierarchical feature maps to generate lesion-aware soft attention maps. Then, we feed the attention maps into the corresponding layers of the type-classification sub-network to make it focus on the lesion regions.

(3) Similar to the residual learning [5], the proposed strategy is also based on modular designment. The prior-attention mechanism is incorporated into residual blocks (referred to as PARL blocks). Thus, deep models can be easily built by stacking the PARL blocks and trained end-to-end.

(4) The afore-mentioned issues (i.e., insufficiency of training data, inter-class similarity, intra-class variation, and non-lesion regions of images) are the common challenges in the whole field of medical image processing. Among these issues, the "non-lesion regions" can aggravate the other issues and it is the main obstacle in improving performance, especially under scenarios where the non-lesion regions in medical images have complicated tissue variations. The proposed method can alleviate this issue by learning effective lesion-aware attention information from targeting lesion images (or patches) and normal images (nor background patches). Therefore, the proposed method can be also applied to a variety of similar scenarios in clinical practice, such as skin lesion classification [24], thorax disease classification [27], glaucoma detection [28], pulmonary nodule detection [29] and their malignancy prediction [30], etc.

II. Related Work

A. Semantic Segmentation

Semantic segmentation plays important role in the field of pattern recognition. Its main task is to identify all pixels that belong to objects of a specific class in an image. To this end, many DCNN-based segmentation methods [15]–[19] have been proposed in literature. Long *et al.* [15] proposed a fully convolutional networks (FCN) for semantic segmentation in natural images. Convolutional operations are stacked layer-by-layer to extract hierarchical feature maps of an input image.

The final layer of the feature maps is then used to generate a pixel-wise probability score map indicating which class the pixels belong to. Upon the FCN, several variants [16]–[19] were developed for more precise segmentation.

Recently, DCNN models were also developed for medical image segmentation. Ronneberger *et al*. [16] developed a U-Net for biomedical image segmentation. Tang *et al*. [31] modified V-Net [32] to a 3D version for lung lobe segmentation in CT images. In this study, a 3D U-Net was also trained for lobe segmentation as a pre-processing step of the COVID-19 detection.

B. Deep Attention Learning

The performance of a model is supposed to depend heavily on the model depth (i.e., the deeper, the better). To train robust models as deep as possible, many prior works have focused on either collecting large-scale datasets (e.g., the ImageNet database [33]) or developing powerful computational tricks, such as the dropout [34], normalizations [35], [36] and "shortcut connections" [5]. Among these tricks, the dropout and normalizations can effectively suppress the over-fitting issue. However, the main obstacle in training deep models is the so-called degradation problem [37]. The residual learning technique [5] successfully addresses this issue using residual learning blocks with "shortcut connections". Although these tricks have demonstrated their validity in many applications, it is still a challenge to train very deep models in some specific scenarios (e.g., the field of medical image analysis) due to the complicated application tasks and the shortage of large-scale datasets.

Recently, some works [21]–[24] have investigated that the attention mechanism is an effective technique that helps further improve the performance of DCNNs. Wang *et al*. [21] proposed a residual attention network for image classification. The network is constructed by a cascade of several attention modules. Each module contains a trainable encoder-decoder structure to learn soft attention masks, which are then multiplied to the convolutional feature maps to highlight important information. Hu *et al.* [22] designed a "Squeeze-and-Excitation" (SE) block, with the goal of improving the quality of representations from a network by explicitly modelling the interdependencies between the channels of its convolutional features. Chen *et al*. [23] took full advantage of the three characteristics of CNN features, namely spatial, channel-wise, and multi-layer, for visual attention-based image captioning. They designed a novel SCA-CNN model that learned to pay attention to every feature entry in the multi-layer 3D feature maps. Inspired by the self-attention ability of CNNs [26], Zhang *et al*. [24] designed a novel self-attention residual network for skin lesion classification. The network can work well without adding any extra learnable layers.

Although all the above-mentioned attention mechanisms effectively improve the performance of deep learning models in large-scale natural image classification tasks, they still suffer from a main drawback for medical image classification. Generally, lesions in medical images have the issue of inter-class similarity, intra-class variation and complicated contextual information as discussed in Section I. These attention mechanisms (trained using only targeted lesion images) may fail to learn rich discriminative representations of different lesions. In contrast, the proposed prior-attention mechanism can learn more effective soft-attention maps, since the training is driven by binary classification between lesion images and normal images without lesions.

C. COVID-19 Screening

Some attempts [38]–[46] have been made to develop CAD systems for COVID-19 screening in CT images. For example, Li *et al*. [44] trained a 2D convolutional neural network (CNN) for three-category classification of CT scans, i.e. COVID-19, community acquired pneumonia (CAP), and non-pneumonia. The network takes a series of CT slices as input and uses the 2D-ResNet50 as a backbone to extract CNN features from each slice of the CT series. The features are then combined using a max-pooling operation and the resulting map is fed to a fully connected layer to generate a probability score for each class. Xu *et al*. [45] first used a 3D segmentation model, i.e. V-Net [32] to segment lesion candidates from CT images. Then, the candidates were classified into COVID-19 or Influenza-A viral pneumonia using a 2D-ResNet18 model.

Although these attempts have demonstrated their validity in COVID-19 screening, some drawbacks remain in clinical application. More specifically, there are many causes of pneumonia such as infections from various types of bacteria and viruses. Xu *et al*. [45] classified pneumonia into either COVID-19 and Influenza-A. This classification task is too simple for clinical application. In contrast, the work proposed by Li *et al*. [44] seems more significant in clinical application as their model can distinguish COVID-19 from CAP, rather than just Influenza-A. However, one of the main challenges in clinical practice is identifying COVID-19 from other viral pneumonia types. The CAP cases collected by Li *et al*. [44] contain a large number of non-viral pneumonia cases. Therefore, the ability to differentiate COVID-19 from other viral pneumonia types needs further verification. Besides, they trained a single 2D-CNN for classifying non-pneumonia (Non-Pneu), CAP, and COVID-19. This training strategy may fail to learn sufficient discriminative semantic representations for effectively differentiating pneumonia types due to two main reasons: (1) Models trained for multi-class categorization tasks may suffer from the inter-class interference issue [11]. For instance, the Non-Pneu cases inevitably interfere with the training of classification between COVID-19 and other pneumonia types. (2) A lung image infected with pneumonia still contains a large part of non-lesion regions as mentioned in Section I, which also prevents the improvement of classification performance.

In summary, our contributions can be concluded as: (1) our study focuses on developing techniques for classifying COVID-19 from other types of viral pneumonia. (2) We directly use 3D CNNs to extract features from the whole 3D lung regions so that richer 3D spatial information can be learned. (3) We conduct experiments to demonstrate that the proposed method can achieve state-of-the-art performance. The main improvement of the proposed method relies on

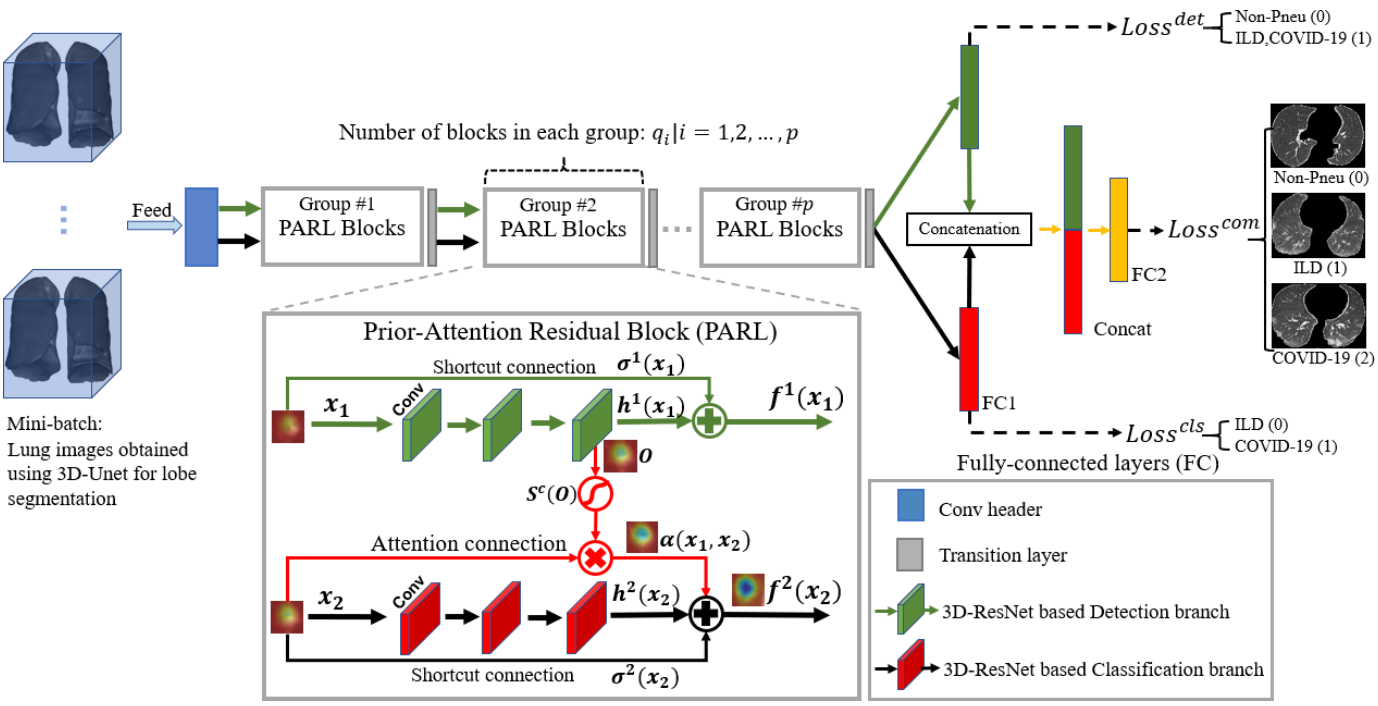


Fig. 2. The architecture of the proposed multi-task prior-attention residual learning strategy. Attention maps are transferred from the detection branch to the pneumonia type-classification branch inside the prior-attention residual learning (PARL) blocks. Three losses are integrated in the training stage, containing lesion detection loss (L^{det}), type classification loss (L^{cls}) and composed classification loss of the above two losses (L^{com}). Two hyper-parameters$p$ and $q_i|i = 1, 2, \dots, p$ are the number of groups and the number of blocks in each group. x_1 and x_2 represent the input feature maps while $f^1(x_1)$ and $f^2(x_2)$ represent the output feature maps of the two branches. $\sigma(\cdot)$ denotes the identity mapping (short-cut connection) and $S^c(\boldsymbol{O})$ is the soft attention maps produced by the detection branch. $\alpha(x_1, x_2)$ is the attention connection.

the application of prior-attention mechanism and multi-task training for learning more discriminative lesion-aware representation for the COVID-19 screening.

III. Methods

The proposed framework for the COVID-19 screening contains two main stages: (a) lobe segmentation using 3D-Unet [20] as a pre-processing step and (b) pneumonia prediction using 3D-ResNets with prior-attention mechanism. Details of the two stages are presented in Section III-A and III-B, respectively.

A. Lobe Segmentation

Lung segmentation in CT images is an important prerequisite step for automatic pneumonia detection. The left and right human lungs are divided into a total number of five lobes (i.e., two lobes in the left lung and three in the right). Previous investigators used UNet or its variants to segment lung regions or lung lobes [47]. Lobe segmentation is more complicated than lung segmentation. However, in clinical practice, lobe information can play a pivotal role as reference for doctors to locate pulmonary lesions and perform their quantitative analysis of the lesions [47]. Hence, it is a basic function in most commercial CAD systems. In this study, we also directly segment lung regions into five lobes.

To achieve this task, we trained a 3D-UNet [20] for lobe segmentation in volumetric CT scans. For a given scan, we first use thresholding and connected-component labeling algorithms to obtain a binary lung mask that indicates the coarse lung regions [29]. Then, we crop a sub-image containing lung regions covered by convex hull of the lung mask, which removes noise outside the lungs, as well as reducing the cost of GPU memory. Finally, we apply the trained 3D-UNet model on the sub-image to obtain its lobe mask.

B. Pneumonia Prediction

After the lobe mask is obtained, we crop refined lung regions according to the lobe mask. The cropped image is then resized to $96 \times 96 \times 96$ and fed into the 3D-ResNets for pneumonia prediction.

As shown in Fig. 2, two 3D-ResNet based sub-networks are designed for two tasks: pneumonia detection (as demonstrated with green cubes) and pneumonia-type classification (as demonstrated with red cubes). The detection sub-network is implemented as a binary classifier that can identify whether or not a given CT scan contains pneumonia, while the type-classification sub-network is implemented for binary classification of ILD and COVID-19. The two sub-networks are fused together using an extra fully-connected layer (as illustrated with the yellow rectangle in Fig. 2) for final three-category classification, i.e., Non-Pneu, ILD, and COVID-19. To enhance the COVID-19 screening, the convolutional layers of the two sub-networks are closely combined via a prior-attention mechanism. The inference procedure can be expressed as:

$$\boldsymbol{P} = f\left(\boldsymbol{I}, \boldsymbol{W}_{det}, \boldsymbol{W}_{cls} | S\left(\boldsymbol{W}_{det}\right), \boldsymbol{W}_{fc}\right), \quad (1)$$

where $\boldsymbol{I}$ is the volumetric lung image that fed into the model f. $\boldsymbol{W}_{det}$ and $\boldsymbol{W}_{cls}$ indicate the learned convolutional weights of the detection and the type-classification sub-network, respectively. $\boldsymbol{W}_{fc}$ denotes the learned weights of the fully-connected layers. $S(\cdot)$ denotes an attention function. The output $\boldsymbol{P}$ is a softmax probability vector:

$$\boldsymbol{P} = \left[p^{non}, p^{ild}, p^{cvd}\right], \tag{2}$$

where p^{non}, p^{ild}, and p^{cvd} are the probabilities corresponding to the three classification categories (i.e., Non-Pneu, ILD, and COVID-19), respectively.

Normally, the lung areas in a CT image contain a large part of non-lesion regions, where complicated variation of lung tissues exist, e.g., vessels and fibers. Obviously, these non-lesion regions have negative impact on the type-classification. To alleviate this issue, we generate soft lesion-aware maps using the convolutional feature maps of the detection sub-network who has remarkable lesion localization ability. The soft maps are then fed into the type-classification sub-network to make it pay attention to the lesion regions. Since the attention information is generated from another model, rather than the type-classification model itself, we call it "prior-attention".

In practice, both sub-networks can be trained independently, i.e., training the detection sub-network followed by training the type-classification sub-network and the extra fully-connected layer that is used to fuse the two sub-networks. However, this training strategy has two main drawbacks. First, multi-stage training, rather than end-to-end training, is much more time-consuming. Second, it is complicated to implement the prior-attention mechanism that transfers the attention information from a pre-trained detection model to a type-classification model. Accordingly, we designed a new residual learning block that incorporated with the prior-attention mechanism (i.e., the PARL block as shown in Fig. 2) to make it possible that hierarchical prior-attention information can be transferred inside the basic blocks. Benefitting from this proposed modular network design, deep attention models can be easily built by cascading the blocks and trained end-to-end using multi-task loss. Details are introduced in Section III-C and III-D.

C. PARL Block

As shown in Fig. 2, each PARL block has two branches: a branch for the pneumonia detection task (demonstrated by the green cubes) and another branch for the type-classification task (demonstrated by the red cubes).

The classification branch is a prior-attention residual learning unit that is composed of three stacked 3D convolutional layers, a "shortcut connection", and an "attention connection" (each convolutional layer is followed by a batch normalization layer and a ReLU activation layer which are not drawn in Fig. 2 for simplicity reasons). If the shortcut connection, the attention connection, and the underlying mapping fitted by the convolutional layers are denoted as $\sigma^2(\boldsymbol{x}_2)$, $\alpha(\boldsymbol{x}_1, \boldsymbol{x}_2)$and $h^2(\boldsymbol{x}_2, \boldsymbol{W}_2)$, respectively, the output of the unit can be expressed as follows:

$$f^2(\boldsymbol{x}_2) = \sigma^2(\boldsymbol{x}_2) + h^2(\boldsymbol{x}_2, \boldsymbol{W}_2) + \gamma \times \alpha(\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{3}$$

where $\boldsymbol{x}_2$, $\boldsymbol{W}_2$ represent the input feature map of the unit and the weights learned by the convolutional layers in the classification branch, respectively. $\boldsymbol{x}_1$ is the input feature map of the residual unit in the detection branch. γ is a weighting factor that controls a trade-off between the attention feature map and other two feature maps. In our implementation, γ is set to 1.0 by default for simplicity reasons. According to the original residual learning [5], the short connection can be simply implemented as an identity mapping:

$$\sigma^2(\boldsymbol{x}_2) = \boldsymbol{x}_2. \tag{4}$$

In (3), the attention connection $\alpha(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is the key factor to improve the classification performance. It is obtained by multiplying a soft attention map to the input feature map on an element-wise basis:

$$\alpha(\boldsymbol{x}_1, \boldsymbol{x}_2) = S(\boldsymbol{O}) \cdot \boldsymbol{x}_2, \tag{5}$$

where $\boldsymbol{O} = h^1(\boldsymbol{x}_1)$ denotes the feature maps of the final layer in the detection branch. The term of $S(\cdot)$ represents a normalization function used to generate the soft attention map from the feature map $\boldsymbol{O}$:

$$S(\boldsymbol{O}) = \left\{ m \,\middle|\, m^c_{i,j,k} = \frac{e^{O^c_{i,j,k}}}{\sum_{i',j',k'} e^{O^c_{i',j',k'}}} \right\}, \tag{6}$$

where (i, j, k) and c represent the spatial coordinates and the channel index of $\boldsymbol{O}$, respectively. $S(\cdot)$ uses a spatial softmax function to highlight the important regions in each channel.

Note that the channel number of $\boldsymbol{O}$ should be equal to that of $\boldsymbol{x}_2$ to satisfy the element-wise multiplication (this is our default implementation for simplicity reasons). Else, a $1 \times 1 \times 1$ convolutional operation can be performed on $\boldsymbol{O}$ to harmonize the channel number.

D. Model Building and Loss Function

A deep model with arbitrary depth can be easily constructed by stacking PARL blocks as shown in Fig. 2. Similar to the original residual learning [5], multiple blocks are grouped together followed by a transition layer (CNN operations with stride 2) to reduce spatial size of the feature maps. In our implementation, two main hyper-parameters are used for building the model: p and $q_i|i = 1, 2, \ldots, p$, which are used to control the number of groups and the number of PARL blocks in each group, respectively.

To train the model, a mini-batch of samples, including normal images without pneumonia, images with ILD, and images with COVID-19 are fed into the model per iteration. The model is optimized by minimizing an objective function of multi-task loss that is defined as:

$$L = \frac{1}{M}\sum_{i=1}^{M} L^{det}(y_i^{det}, \hat{y}_i^{det}) + \frac{1}{M_{cls}}\sum_{i=1}^{M_{cls}} y_i^{det} L^{cls}(y_i^{cls}, \hat{y}_i^{cls}) + \frac{1}{M}\sum_{i=1}^{M} L^{com}(y_i^{com}, \hat{y}_i^{com}), \tag{7}$$

where the first term, the second term, and the third term are the cross entropy for the detection, the binary type-classification,

and the final combined three-category classification task, respectively. M and M_{cls} are mini-batch size and the number of positive samples (i.e., ILD and COVID-19) in the mini-batch. y_i and $\hat{y}_i$ represent the ground truth and the predicted label. For computing the loss of the detection branch, y_i^{det} is set to 0 if the sample is a normal image, and is set to 1 if the sample is an image infected with ILD or COVID-19. For computing the loss of the classification branch, the negative samples (i.e., the normal images) are directly ignored and y_i^{cls} is set to 0 or 1 if the positive sample is infected with ILD or COVID-19. The term $y_i^{det} L^{cls}(y_i^{cls}, \hat{y}_i^{cls})$ means the binary classification loss is activated only for positive samples (i.e., $y_i^{det} = 1$) and disabled otherwise ($y_i^{det} = 0$). For computing the loss of the final combined three-category classification task, y_i^{com} is set to 0, 1 or 2 for Non-Pneu, ILD, and COVID-19, respectively.

IV. Materials

For this study, ethical approval was obtained, and the informed consent requirement was waived (Approval Number: KY2020036). We collected CT scans of 4657 patients (F/M, 1946/2711; mean age: 46 ± 17 years) from several cooperative hospitals, including a total of 936 normal scans, 2406 scans with ILD caused by viruses, and 1315 scans with COVID-19. All the pneumonia diseases were confirmed as positive by RT-PCR or serum antibody test besides COVID-19. The ILD patient inclusion or exclusion criteria was executed based on "An official American Thoracic Society/European Respiratory Society statement" by two experienced respiratory physicians (HL with 10 years of experience and FX with 15 years of experience). All the ILD CT images were independently reviewed by two experienced radiologists in CT diagnostics (XL with 8 years of experience and CL with 10 years of experience). The ILD CT images must have the pulmonary fibrosis features. In clinical practice, there were patients who underwent several scans. For each of these patients, we selected only the scan that was firstly reconstructed with the thinnest slice-thickness for building the dataset.

CT examinations were performed using scanners from different manufacturers with standard chest imaging protocols. Each scan contained 96-539 slices with a varying slice-thickness from 0.5 mm to 3 mm. The reconstruction matrix of each slice was 512 × 512 with in-plane pixel spatial resolution from 0.63 mm × 0.63 mm to 0.83 mm × 0.83 mm. From these collected scans, we randomly selected 60 scans (20 scans of each class) for online-evaluation, 600 scans (200 scans of each class) for offline-test, and the rest 3997 scans for training and 5-fold cross-validation.

In order to train the 3D-Unet for the lobe segmentation, we collected a total of 251 chest CT scans with corresponding voxel-level lobe labels. Among these scans, 51 cases were pneumonia-free and publicly available. They were chosen from the LUNA-16 dataset [48] and annotated by Tang *et al.* [31]. The 3D-Unet trained using just these scans were not reliable for segmentation of scans infected with pneumonia. Hence, we collected additional 200 scans with pneumonia to augment the training dataset. These scans were annotated by the two radiologists (i.e., XL and CL) and were not included in the above-mentioned 4657 scans.

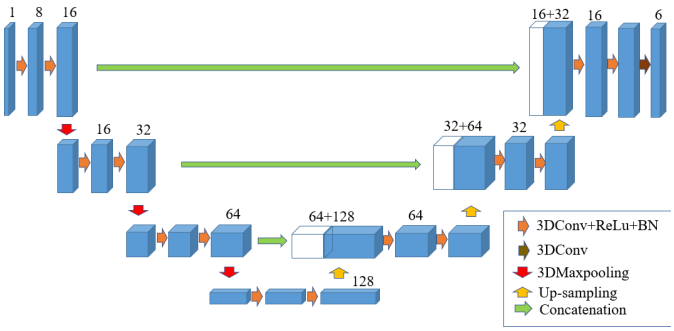


Fig. 3. The 3D-UNet architecture for lobe segmentation. Input image size is 128 × 96 × 128 and output size is 128 × 96 × 128 × 6 where the number of channels (i.e., 6) correspond to 6 categories, including non-lung regions and 5 lobes.

To reduce the variations such as slice-thickness between the scans, we interpolated each scan to 1 mm × 1 mm × 1mm and converted CT numbers (Hounsfield units) to gray-scale values using lung window (L/W: -500 HU/1500 HU).

V. Experiments

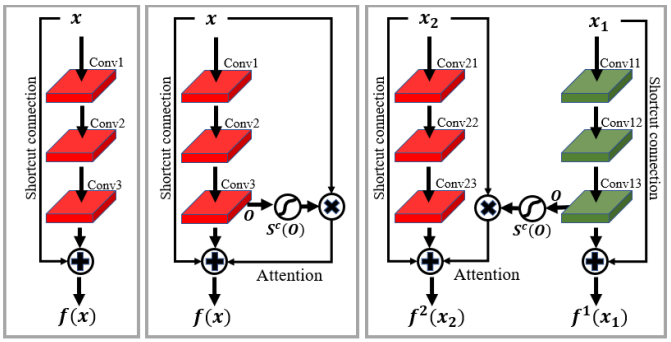
A. Model Configurations

Network architecture of the 3D-UNet trained for lobe segmentation is shown in Fig. 3. The input image size is 128 × 96 × 128 (Z × Y × X) and the output size is 128 × 96 × 128 × 6. The six channels of output map correspond to predicted probabilities of six categories, including non-lung regions, upper and inferior lobes of left lung, and upper, middle, and inferior lobes of right lung, respectively. As introduced in Section III-A, to remove most non-lung regions, each scan is pre-segmented using a coarse lung segmentation method. The resulting image has a wider side in the X direction than the Y direction. Hence, we set the anisotropic input size (i.e., 128 × 96 × 128) empirically in our implementation to keep the shape and the size of the image as much as possible.

During the training stage of 3D-Unet, a mini-batch size of 2 samples were fed into the model. In this study, we focused only on the pneumonia classification tasks, rather than the lobe segmentation task. More details of 3D-UNet and lobe segmentation can be found in the paper proposed by Cicek *et al.* [20] and the paper proposed by Tang *et al.* [31], respectively.

For classification tasks, we compared the proposed multi-task prior attention residual learning strategy for the COVID-19 screening with two baselines. One is the residual learning without attention and another is the residual learning with the self-attention mechanism [24]. All of these strategies are based on modular designment. The major difference of these residual learning blocks is illustrated in Fig. 4.

In our experiments, a total of five models were built for comparison, namely WA-66, SA-66, WA-66-M, SA-66-M, and PA-66-M as tabulated in Table I. The letters "WA", "SA", and "PA" in the model names are the abbreviations of the without-attention, the self-attention, and the prior-attention strategy, respectively. The number "66" in the model names



(a) Without-attention (b) Self-attention (c) Proposed prior-attention

Fig. 4. The main difference between the residual blocks (a) without attention (WARL), (b) with self-attention (SARL) [24], and (c) with the proposed multi-task prior-attention (PARL).

indicates the number of convolutional layers in each model. To guarantee comparison consistency, all models have the same input image size ($96 \times 96 \times 96$) and have the same magnitude of parameters. Each model contains four groups of corresponding blocks (i.e., $p = 4$). Similar to many previous works [43]–[45], the WA-66 and SA-66 were trained as classifiers that directly identify three categories, i.e., Non-Pneu, ILD and COVID-19. The WA-66-M, SA-66-M, and PA-66-M models were trained using the proposed multi-task learning strategy for ablation studies. Both the WA-66-M and SA-66-M models have identical network architecture to the PA-66 model but without the prior-attention mechanism. In the SA-66-M model, the self-attention mechanism was incorporated in the pneumonia-type classification branch.

B. Training Details and Evaluation Metrics

All classification models were trained using Google Tensorflow (version 2.0 with Keras API) on NVIDIA RTX 2080Ti GPUs. During the training stage, the loss of each model was minimized using the momentum optimizer with a learning rate of 0.0001, decaying every 500 iterations using an exponential rate of 0.95. The total number of iterations was 30k (300 epochs multiply by 100 iterations).

At each iteration, a mini-batch of 10 samples were fed into the models, including 4 normal scans, 3 scans with ILD, and 3 scans with COVID-19. We augmented the samples in real time by randomly rotating each sample to 0, 90, 180, and 270 degrees along the Z axis, and randomly flipping them in the X, Y, and Z directions. Once an epoch was completed, we performed online-evaluations using 60 samples (20 samples for each type). To avoid the over-fitting issue, we saved just the model that achieved the maximum online accuracy.

During the testing stage, the predicted label $\hat{y}_i$ of a specific sample was set to $\text{argmax}(p_k | k \in \{0, 1, 2\})$, where p_k is the estimated probability corresponding to each category, i.e., Non-Pneu ($k = 0$), ILD ($k = 1$), and COVID-19 ($k = 2$).

Ablation experiments were conducted on a total of 3997 scans using 5-fold cross-validation as mentioned in Section IV. The scans were randomly split into five subsets $\boldsymbol{S}_i | i = 1, 2, \ldots, 5$, which were used to train five independent models $\boldsymbol{M}_j | j = 1, 2, \ldots, 5\}$. Each model $\boldsymbol{M}_j$ was trained using four subsets $\boldsymbol{S}_i | i = 1, 2, \ldots, 5$ and $i \neq j$ and evaluated using the rest subset. The performance of each model was assessed in terms of accuracy, recall (sensitivity), specificity, precision, F1-value, and AUC. Then, the overall performance of the proposed method was assessed by calculating the mean and standard deviation of cross-validation metrics.

To further analyze the proposed method and make comparisons with existing methods, models were also trained using the above-mentioned 3997 scans and evaluated using the testing dataset containing a total of 600 scans. Confusion matrices were used for quantitative analysis.

VI. Results

The results of 5-fold cross-validation are tabulated in Table II. By observing the results, two main conclusions can be drawn: (1) For classification of all three target categories, all the PA-66-M, the WA-66-M, and the SA-66-M models achieve higher AUC compared to the WA-66 model and the SA-66 model. This phenomenon demonstrates that the multi-task learning strategy can suppress the inter-class interference issue by splitting the three-category classification task into two binary classification tasks, and thus the performance is improved. (2) For pneumonia-type classification (i.e., ILD or COVID-19), the SA-66-M model outperforms the WA-66-M model. However, the improvement is very minor. In contrast, the proposed PA-66-M model improves the performance by a large margin. The AUC value corresponding to the ILD and the COVID-19 achieved by the PA-66-M model are 95.7% and 97.3%, respectively, which are much higher than 93.2% and 92.8% achieved by the SA-66-M model. This phenomenon demonstrates that the proposed prior-attention mechanism, compared to the self-attention, can further improve the performance.

The above analysis reveals that, compared to just designing deeper models, developing novel techniques such as attention mechanisms and multi-task learning strategies also improve classification performance, especially under scenarios where large-scale dataset is hard to collect. As listed in Table I, both the WA-66 and the SA-66 models (containing 19 corresponding residual blocks) are much deeper than the other three models that contain only 8 corresponding residual blocks but have wider network architectures (i.e., the WA-66-M, the SA-66-M, and the PA-66-M models). However, the performances of the WA-66 and the SA-66 models are inferior to that of the other three models.

To further validate the proposed method, we trained additional WA-66-M, SA-66-M, and PA-66-M models using all the 3997 scans. The main difference between the WA-66-M, the SA-66-M, and the PA-66-M models is the attention mechanism that was used in the pneumonia-type classification branch. To demonstrate the effectiveness of the attention mechanisms, Fig. 5 shows the training loss curves of the pneumonia-type classification branch (i.e., L^{cls}) corresponding to each model. Evidently, the variation tendencies of curves corresponding to the WA-66-M and the SA-66-M model are very close to each other. But minor differences can still be observed: the SA-66-M model converges faster than the

TABLE I
FIVE NETWORK ARCHITECTURES FOR COMPARISON

Layer name	Output size	WA-66	SA-66	WA-66-M	SA-66-M	PA-66-M
Conv header	48 × 48 × 48	C16, (5, 5, 5), /2	C16, (5, 5, 5), /2	C16, (5, 5, 5), /2	C16, (5, 5, 5), /2	C16, (5, 5, 5), /2
Group #1	48 × 48 × 48	C16, WARLs ($q = 3$)	C16, SARLs ($q = 3$)	C16, MWARLs ($q = 2$)	C16, MSARLs ($q = 2$)	C16, PARLs ($q = 2$)
Transition	24 × 24 × 24	C32, (3, 3, 3), /2; C32, (3, 3, 3), /1	C32, (3, 3, 3), /2; C32, (3, 3, 3), /1	C32, (3, 3, 3), /2; C32, (3, 3, 3), /1	C32, (3, 3, 3), /2; C32, (3, 3, 3), /1	C32, (3, 3, 3), /2; C32, (3, 3, 3), /1
Group #2	24 × 24 × 24	C32, WARLs ($q = 4$)	C32, SARLs ($q = 4$)	C32, MWARLs ($q = 2$)	C32, MSARLs ($q = 2$)	C32, PARLs ($q = 2$)
Transition	12 × 12 × 12	C64, (3, 3, 3), /2; C64, (3, 3, 3), /1	C64, (3, 3, 3), /2; C64, (3, 3, 3), /1	C64, (3, 3, 3), /2; C64, (3, 3, 3), /1	C64, (3, 3, 3), /2; C64, (3, 3, 3), /1	C64, (3, 3, 3), /2; C64, (3, 3, 3), /1
Group #3	12 × 12 × 12	C64, WARLs ($q = 9$)	C64, SARLs ($q = 9$)	C64, MWARLs ($q = 2$)	C64, MSARLs ($q = 2$)	C64, PARLs ($q = 2$)
Transition	6 × 6 × 6	C128, (3, 3, 3), /2; C128, (3, 3, 3), /1	C128, (3, 3, 3), /2; C128, (3, 3, 3), /1	C128, (3, 3, 3), /2; C128, (3, 3, 3), /1	C128, (3, 3, 3), /2; C128, (3, 3, 3), /1	C128, (3, 3, 3), /2; C128, (3, 3, 3), /1
Group #4	6 × 6 × 6	C128, WARLs ($q = 3$)	C128, SARLs ($q = 3$)	C128, MWARLs ($q = 2$)	C128, MSARLs ($q = 2$)	C128, PARLs ($q = 2$)
Transition	3 × 3 × 3	C256, (3, 3, 3), /2; C256, (3, 3, 3), /1	C256, (3, 3, 3), /2; C256, (3, 3, 3), /1	C256, (3, 3, 3), /2; C256, (3, 3, 3), /1	C256, (3, 3, 3), /2; C256, (3, 3, 3), /1	C256, (3, 3, 3), /2; C256, (3, 3, 3), /1
Flatten	1 × 1 × 1	[6912]	[6912]	[6912, 6912]	[6912, 6912]	[6912, 6912]
FC1	1 × 1 × 1	[1488]	[1488]	[512, 512]	[512, 512]	[512, 512]
Concat	1 × 1 × 1	None	None	[1024]	[1024]	[1024]
FC2	1 × 1 × 1	[512]	[512]	[512]	[512]	[512]
Y_Preds	1 × 1 × 1	[3]	[3]	[3]	[3]	[3]
Total parameters		15,959,851	15,959,851	15,988,903	15,988,903	15,988,903

Input image size is 96 × 96 × 96. WARLs, SARLs, and PARLs are the residual blocks illustrated in Fig. 4. MWARLs and MSARLs mean the multi-task residual learning block without attention and with self-attention, respectively. q is the number of blocks in each group. C· is the number of features. (·, ·, ·) is the kernel size and /· is the stride. [·] means fully connections.

TABLE II
RESULTS OF 5-FOLD CROSS-VALIDATION OF THE FIVE CLASSIFICATION MODELS FOR COMPARISON

Lesion Type	Models	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-value (%)	AUC (%)
Non-Pneu	WA-66	86.0 ± 1.6	81.4 ± 7.3	87.0 ± 3.1	58.3 ± 4.0	67.6 ± 2.3	92.6 ± 1.7
	SA-66	84.0 ± 4.4	83.8 ± 16	84.0 ± 8.2	56.9 ± 10.9	65.4 ± 4.2	92.9 ± 2.2
	WA-66-M	82.4 ± 5.7	89.9 ± 7.7	80.8 ± 8.5	54.2 ± 14.2	65.8 ± 6.9	93.8 ± 0.7
	SA-66-M	78.4 ± 2.7	**95.1 ± 1.6**	74.8 ± 3.1	45.4 ± 3.4	61.4 ± 3.3	93.9 ± 1.5
	PA-66-M	**91.5 ± 1.0**	82.3 ± 4.7	**93.5 ± 1.6**	**73.8 ± 4.2**	**77.6 ± 2.1**	**95.3 ± 0.8**
ILD	WA-66	71.1 ± 7.8	49.1 ± 15.1	**97.7 ± 1.3**	96.5 ± 1.0	63.6 ± 14.0	91.2 ± 2.9
	SA-66	75.5 ± 7.0	59.0 ± 14.1	95.3 ± 1.7	94.0 ± 0.9	71.4 ± 11.0	90.7 ± 2.0
	WA-66-M	78.6 ± 3.8	64.7 ± 8.3	95.3± 3.3	94.6 ± 3.0	76.4 ± 5.8	92.2 ± 1.9
	SA-66-M	75.3 ± 4.6	57.0 ± 9.3	97.4 ± 1.4	**96.6 ± 1.6**	71.2 ± 7.4	93.2 ± 1.7
	PA-66-M	**89.4 ± 1.2**	**88.5 ± 1.5**	90.6 ± 2.6	91.9 ± 1.9	**90.2 ± 1.1**	**95.7** ± 1.2
COVID-19	WA-66	76.9 ± 8.1	**93.4 ± 3.4**	70.7 ± 12.2	56.6 ± 10.0	69.8 ± 6.8	92.1 ± 1.9
	SA-66	81.0 ± 10.9	83.7 ± 9.2	80.0 ± 18.0	67.4 ± 15.5	72.5 ± 8.4	92.2 ± 1.7
	WA-66-M	85.7 ± 8.2	79.7 ± 11.1	87.9 ± 14.3	77.8 ± 15.8	76.6 ± 7.6	92.9 ± 4.3
	SA-66-M	87.1 ± 3.5	81.1 ± 2.1	89.4 ± 4.9	75.4 ± 8.8	77.8 ± 4.7	92.9 ± 1.8
	PA-66-M	**93.3 ± 0.8**	87.6 ± 4.3	**95.5 ± 2.1**	**88.4 ± 4.1**	**87.8 ± 1.5**	**97.3 ± 1.1**

The highest score in each column of each lesion type is shown in bold.

WA-66-M, especially after 100 epoch iterations. In contrast, the PA-66-M converges much faster than both the other models, especially during the stage of the first 100 epochs. This phenomenon mainly stems from the fact that the proposed prior-attention mechanism can learn lesion-attention information more efficiently than the self-attention mechanism.

We evaluated these models using the offline-testing dataset containing 600 scans (200 scans for each category) and used confusion matrices for quantitative analysis. The matrices are shown in Fig. 6. Each row in a confusion matrix represents an actual ground truth class, while each column represents a predicted class. A better classifier which can predict more correct samples would have larger values on the diagonal of its confusion matrix (highlighted as red in Fig. 6).

By observing the confusion matrices in Fig. 6, the superiority of the proposed PA-66-M model is evident compared to the WA-66-M and the SA-66-M models. The superiority mainly reflects in the classification of the ILD and

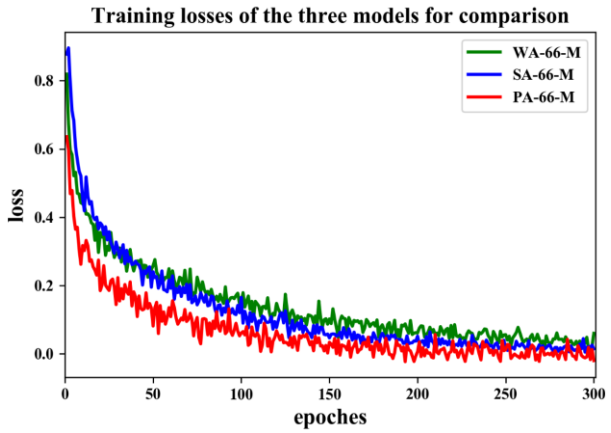


Fig. 5. Training loss curves of the pneumonia-type classification branch corresponding to the three models (i.e., WA-66-M, SA-66-M, and PA-66-M) trained using the multi-task learning strategy. It can be observed that the convergence speed of the proposed PA-66-M model is much faster than that of other models.

COVID-19 categories. The PA-66-M model achieves 191 and 176 correct predictions out of 200 ILD and 200 COVID-19, respectively, which is higher than 125 and 169 achieved by the WA-66-M model, and 122 and 167 achieved by the SA-66-M model. This phenomenon further demonstrates that the proposed prior-attention mechanism can significantly enhance pneumonia-type classification performance. However, all models have misclassifications between the Non-Pneu and the pneumonia categories. By analyzing the original images of these misclassified cases, we found that most cases with pneumonia looked similar to the normal scans, as the pneumonia lesions in these cases were not severe. It was difficult to differentiate scans with light pneumonia lesions from normal scans.

We also reviewed relevant state-of-the-art studies on the CT-based COVID-19 screening task, as listed in Table III. Most existing studies focused on developing methods for identifying COVID-19 from other types of pneumonia, including non-viral pneumonia. The studies of Wang *et al.* [42] and Xu *et al.* [45] are closer to our work in distinguishing COVID-19 from other viral pneumonia. However, the main drawback of their works was that too few metrics were measured, which is insufficient to accurately reflect the overall performance of the classification.

VII. Discussion

Classification techniques are more feasible alternatives than object detection and segmentation-based methods for developing COVID-19 screening CADs in a relatively short time. This is because training classification models require only image-level ground truth labels. Therefore, analogous to most previous works [38]-[46], we also adopted classification techniques to implement our CT-based COVID-19 screening task.

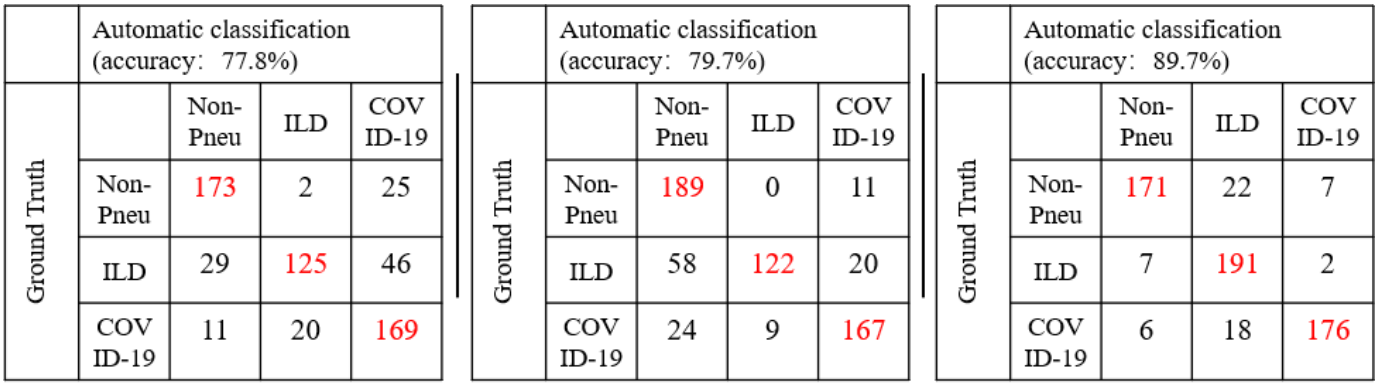
Compared to prior works, our method can achieve superior performance. We attribute the success to two main aspects: (1) We collected more clinical cases from multiple hospitals to train our models. (2) We developed a prior-attention residual learning strategy for training models. In the proposed method, two 3D-ResNet based sub-networks were integrated into a single model for both pneumonia detection and lesion type classification. Since the detection network was trained as a binary classifier using normal images and pneumonia-infected images, it can highlight lesion regions more accurately than models trained using just pneumonia-infected images. Hence, prior-attention information generated by the detection model can more effectively guide the lesion-type classification than self-attention information generated by the type classification model itself.

Fig. 7 shows two clinical cases that are infected with COVID-19 and ILD, respectively. To illustrate the effectiveness of the proposed prior-attention strategy, we created a heatmap from the convolutional feature maps of the type-classification sub-network corresponding to a specific model (i.e., the WA-66-M model, the SA-66-M model, or the PA-66-M model) using a visualization method [26] and applied the heatmap to the original input image. By comparing the heatmaps as shown in Fig. 7 (b), (c), and (d), it can be observed that the proposed PA-66-M model can highlight the lesion regions more accurately than both the WA-66-M model and the SA-66-M model, and the heatmap of PA-66-M has larger red areas (high attention) inside the main lesion regions as indicated with the red dashed rectangles.

Moreover, to obtain the final three-category classification results, the two sub-networks were fused using a learnable fully-connected layer (refereed as to late-fusion strategy), rather than using voting strategies that were commonly adopted in ensemble learning methods (refereed as to committee-fusion strategy). A. A. A. Setio *et al.* [51] have demonstrated that the late-fusion strategy can achieve higher performance than the committee-fusion strategy.

However, the proposed classification model currently may fail to screen out scans with COVID-19 lesions at an early stage and misclassify normal scans to pneumonia category. The lesions in these non-severe scans normally appear as relatively small ground-glass nodules (GGN) that are very difficult to identify from the whole volumetric lung images. To alleviate this issue, pulmonary nodule detection [29] can be adopted as a compensation method. Besides, knowledge about the location information of pulmonary lesions in lung lobe regions (i.e. which lobe the lesions are located at) is useful in clinical practice, e.g. guiding diagnosis or surgery [49]. However, currently, lobe information obtained using the 3D-UNet is only used to segment lung regions as a pre-processing step for pneumonia detection. In the future, we will try to fully use the lobe segmentation result to determine the lobe location of predicted pneumonia lesions.

It is also worth mentioning that the weighting factor γ (as defined in Equation 3) was set to a fixed value (i.e., 1.0) in our experiments for simplicity and fairness of the comparison between the proposed prior-attention strategy and the self-attention strategy proposed by Zhang *et al.* [24]. However, in [24], the weighting factor was implemented as a learnable parameter that can be used to adaptively adjust the contribution of the attention feature map, avoiding the interference by a bad attention feature map obtained at the early stage of model training. Actually, the proposed prior-attention mechanism can effectively avoid this issue even with



	Automatic classification (accuracy: 77.8%)			
Ground Truth		Non-Pneu	ILD	COVID-19
	Non-Pneu	173	2	25
	ILD	29	125	46
	COVID-19	11	20	169

(a) WA-66-M

	Automatic classification (accuracy: 79.7%)			
Ground Truth		Non-Pneu	ILD	COVID-19
	Non-Pneu	189	0	11
	ILD	58	122	20
	COVID-19	24	9	167

(b) SA-66-M

	Automatic classification (accuracy: 89.7%)			
Ground Truth		Non-Pneu	ILD	COVID-19
	Non-Pneu	171	22	7
	ILD	7	191	2
	COVID-19	6	18	176

(c) PA-66-M

Fig. 6. Confusion matrices corresponding to the WA-66-M, the SA-66-M, and the PA-66-M models. The accuracy demonstrate that the PA-66-M achieves much higher performance than that of both the other two models.

TABLE III
STATE-OF-THE-ART STUDIES ON THE COVID-19 SCREENING TASK

Literature	Material Partition (case)	Task	Method	Result
Chen *et al*. [38]	Total 106 Not clearly mentioned	Classification: 1. COVID19 2.others (viral, bacterial Pneu.)	2D-UNet++	95.2% (Accuracy) 100% (Sensitivity) 93.6% (Specificity)
Zheng *et al*. [39]	499 for training 132 for validation	Classification: 1.COVID-19 2.others (viral, bacterial Pneu.)	2D-U-Net 2D-CNN	90.7% (Sensitivity) 91.1% (Specificity)
Jin *et al*. [40]	312 for training 104 for validation 1255 for testing	Classification: 1.COVID-19 2.others (viral, bacterial Pneu.)	2D-CNN	94.1% (Sensitivity) 95.5% (Specificity)
Jin *et al*. [41]	1136 for training 282 for validation	Classification: 1.COVID-19 2.others (viral, bacterial Pneu.)	2D-UNet++ 2D-CNN	97.4% (Sensitivity) 92.2% (Specificity)
Wang *et al*. [42]	Total 250 Not clearly mentioned	Classification: 1.COVID-19 2.Viral Pneu.	2D-CNN	82.9% (Accuracy)
Song *et al*. [43]	164 for training 27 for validation 83 for testing	Classification: 1.COVID-19 2.Bacterial Pneu. 3. Normal	2D-ResNet-50	86.0% (Accuracy)
Li *et al*.[44]	3920 for training 436 for testing	Classification: 1.COVID-19 2.CAP (viral, bacterial Pneu.) 3.Normal	2D-ResNet-50	90.0% (Sensitivity) 96.0% (Specificity)
Xu *et al*.[45]	528 for training 90 for testing	Classification: 1.COVID-19 2.Influenza-A. 3.Normal	2D-CNN	86.7% (Accuracy)
Shi *et al*.[46]	Total 2685 5-fold cross validation	Classification: 1.COVID-19 2.CAP (viral, bacterial Pneu.)	Random Forest	87.9% (Accuracy) 90.7% (Sensitivity) 83.3% (Specificity)
Proposed	3997 for 5-fold cross validation 60 for validation 600 for testing	Classification: 1.COVID-19 2.ILD (only viral Pneu.) 3.Normal	3D-ResNets (with prior-attention)	93.3% (Accuracy) 87.6% (Sensitivity) 95.5% (Specificity)

a constant weighting factor, which has been demonstrated in Fig. 5: the PA-66-M model converges much faster than the SA-66-M model, especially at the early stage of model training. Definitely, in the future, the performance may be further improved if the weighting factor is also implemented as a learnable parameter in the proposed prior-attention strategy.

Finally, it is worth noting that the proposed prior-attention residual learning strategy can also be applied to other medical image classification tasks. For example, in a pulmonary nodule detection system, the detection sub-network can be trained using positive nodule samples and negative background images that are randomly cropped from normal

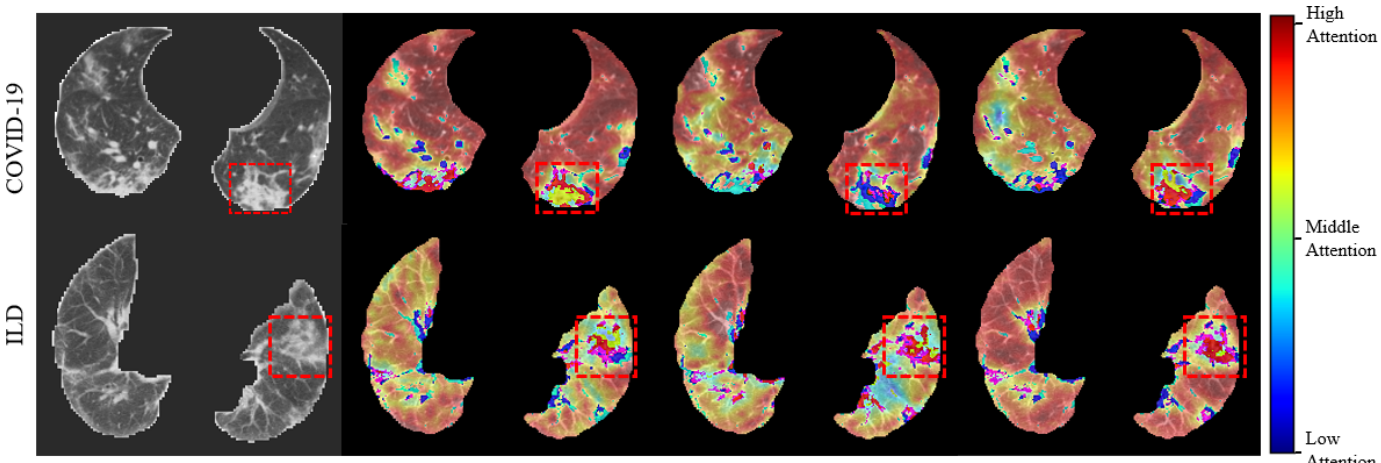


Fig. 7. Two clinical example patients infected with COVID-19 and ILD, respectively. To demonstrate the effectiveness of the proposed method, heatmaps are created from feature maps of each deep model and imposed to the original image. By comparing the heatmaps of (b), (c), and (d), it can be observed that the proposed PA-66-M model can highlight the lesion regions more accurately than both the WA-66-M and the SA-66-M models (see the regions indicated with the red dashed rectangles).

lung regions for false positive reduction [50], [51]. Also, it can be used to strengthen the classification sub-network that is trained for a specific task such as malignancy prediction of detected pulmonary nodules [30].

VIII. Conclusion

In this paper, we presented a novel multi-task prior-attention learning strategy to implement COVID-19 screening in volumetric chest CT images. Specifically, we integrated two ResNet-based branches into one model framework for end-to-end training by designing a prior-attention residual learning (PARL) block. Inside these blocks, hierarchical attention information from lesion region detection branch was transferred to COVID-19 classification branch for learning more discriminative representations. Compared to other methods with self-attention and without attention, our method located lesion regions more correctly so that the extra supervision information is more effective to enhance the performance of COVID-19 classification tasks. Experimental results demonstrated that our method surpassed other state-of-the-art COVID-19 screening methods. In the near future, more efforts will be devoted to exploring how to identify COVID-19 in the early stages and how this prior-attention mechanism can be applied in other medical image analysis problems.

References

[1] WHO. (Apr. 10, 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

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[3] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Aug. 1985.

[4] L. Breiman and A. Cutler. (2007). *Random forests-Classification Description: Random Forests*. [Online]. Available: http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm

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

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comut. Vis.*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[12] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, Netherlands, 2016, pp. 21–37.

[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[14] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 845–853.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, in Lecture Notes in Computer Science, vol. 9351. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[18] L. Chen, G. Papandreou, L. Kokkinos, L. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," 2016, *arXiv:1606.00915*. [Online]. Available: https://arxiv.org/abs/1606.00915

[19] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," 2016, *arXiv:1611.06612*. [Online]. Available: http://arxiv.org/abs/1611.06612

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," 2016, *arXiv:1606.06650*. [Online]. Available: http://arxiv.org/abs/1606.06650

[21] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[22] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*. [Online]. Available: http://arxiv.org/abs/1709.01507

[23] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6298–6306.

[24] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[25] 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.

[26] 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.*, Oct. 2017, pp. 618–626.

[27] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*. [Online]. Available: http://arxiv.org/abs/1801.09927

[28] L. Li *et al.*, "A large-scale database and a CNN model for attention-based glaucoma detection," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 413–424, Feb. 2020.

[29] J. Wang *et al.*, "Pulmonary nodule detection in volumetric chest CT scans using CNNs-based nodule-size-adaptive detection and classification," *IEEE Access*, vol. 7, pp. 46033–46044, 2019.

[30] J. Wang *et al.*, "Feature-shared adaptive-boost deep learning for invasiveness classification of pulmonary sub-solid nodules in CT images," *Med. Phys.*, vol. 47, no. 4, pp. 1738–1749, Apr. 2020. [Online]. Available: https://apm.onlinelibrary.wiley.com/doi/10.1002/mp.14068

[31] H. Tang, C. Zhang, and X. Xie, "Automatic pulmonary lobe segmentation using deep learning," 2019, *arXiv:1903.09879*. [Online]. Available: http://arxiv.org/abs/1903.09879

[32] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016, *arXiv:1606.04797*. [Online]. Available: http://arxiv.org/abs/1606.04797

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami FL, USA, Jun. 2009, pp. 248–255.

[34] N. Srivastava, G. Hiton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[36] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[38] J. Chen *et al. Deep Learning-Based Model For Detecting 2019 Novel Coronavirus Pneumonia on High-Resolution Computed Tomography: A Prospective Study*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org//10.1101/2020.02.25.20021568v2

[39] C. Zheng *et al. Deep Learning-Based Detection For COVID-19 From Chest CT Using Weak Label*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.12.20027185v1.full.pdf

[40] C. Jin *et al. Development and Evaluation of an AI System For COVID-19 Diagnosis*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.20.20039834v2.

[41] S. Jin *et al. Ai-Assisted CT Imaging Analysis For COVID-19 Screening: Building and Deploying A Medical AI System in Four Weeks*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.19.20039354v1

[42] S. Wang *et al. A Deep Learning Algorithm Using CT Images to Screen for Corona Virus Disease (COVID-19)*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.14.20023028v4

[43] Y. Song *et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images*. Accessed: Apr. 3, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.23.20026930v1

[44] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, to be published. Accessed: Apr. 3, 2020. [Online]. Available: https://pubs.rsna.org/doi/pdf/10.1148/radiol.2020200905

[45] 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

[46] 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

[47] 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.

[48] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.

[49] E. M. van Rikxoort, B. de Hoop, S. van de Vorst, M. Prokop, and B. van Ginneken, "Automatic segmentation of pulmonary segments from volumetric chest CT scans," *IEEE Trans. Med. Imag.*, vol. 28, no. 4, pp. 621–630, Apr. 2009.

[50] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[51] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.