

COVID-19 Automatic Diagnosis With Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks

Wenqi Shi , Li Tong , Yuanda Zhu , and May D. Wang , *Senior Member, IEEE*

Abstract—Researchers seek help from deep learning methods to alleviate the enormous burden of reading radiological images by clinicians during the COVID-19 pandemic. However, clinicians are often reluctant to trust deep models due to their black-box characteristics. To automatically differentiate COVID-19 and community-acquired pneumonia from healthy lungs in radiographic imaging, we propose an explainable attention-transfer classification model based on the knowledge distillation network structure. The attention transfer direction always goes from the teacher network to the student network. Firstly, the teacher network extracts global features and concentrates on the infection regions to generate attention maps. It uses a deformable attention module to strengthen the response of infection regions and to suppress noise in irrelevant regions with an expanded reception field. Secondly, an image fusion module combines attention knowledge transferred from teacher network to student network with the essential information in original input. While the teacher network focuses on global features, the student branch focuses on irregularly shaped lesion regions to learn discriminative features. Lastly, we conduct extensive experiments on public chest X-ray and CT datasets to demonstrate the explainability of the proposed architecture in diagnosing COVID-19.

Index Terms—COVID-19, explainable artificial intelligence, automatic diagnosis, radiographic imaging, attention mechanism, knowledge distillation.

I. INTRODUCTION

CORONAVIRUS disease 2019 (COVID-19) has been widely spread worldwide since the beginning of 2020 [1]. COVID-19 is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), a highly contagious virus. At present, Reverse Transcription Polymerase Chain Reaction

Manuscript received September 27, 2020; revised February 8, 2021; accepted February 24, 2021. Date of publication April 21, 2021; date of current version July 20, 2021. The work was supported by The Wallace H. Coulter Distinguished Faculty Fellow, Amazon Faculty Research Fellow, Microsoft Azure Cloud Grant, and Petit Institute Faculty Fellow awards to Professor Wang. The content of this article is solely the responsibility of the authors. (*Corresponding author: May D. Wang.*)

Wenqi Shi and Yuanda Zhu are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: wshi83@gatech.edu; yzhu94@gatech.edu).

Li Tong and May D. Wang are with the Wallace H. Coulter School of Biomedical Engineering, Georgia Institute of Technology, Emory University, Atlanta, GA 30322 USA (e-mail: ltong9@gatech.edu; maywang@gatech.edu).

Digital Object Identifier 10.1109/JBHI.2021.3074893

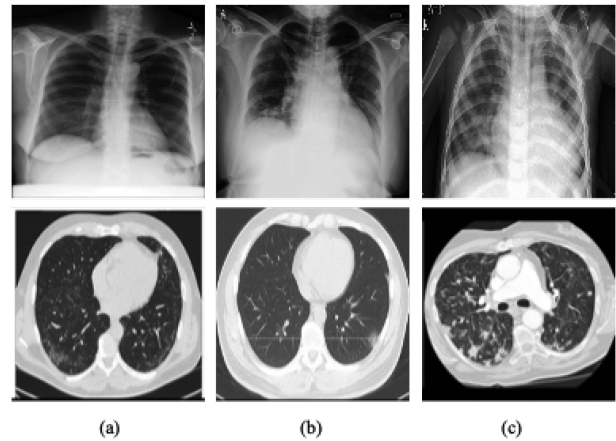


Fig. 1. Examples of X-ray scans (up) and CT scans (bottom) from the collected dataset: (a) no findings (b) COVID-19 pneumonia (c) community-acquired pneumonia (CAP). COVID-19 and CAP both cause the density of the lungs to increase, which can be seen as whiteness in the lungs on radiography, depending on the severity of pneumonia. Compared with non-COVID-19 pneumonia, COVID-19 pneumonia was more likely to have a peripheral distribution, ground-glass opacity, fine reticular opacity, and vascular thickening [5]. CAP is predominantly associated with consolidation on chest radiography [6].

(RT-PCR) is the universally applicable and effective method to diagnose COVID-19 [2]. However, there exists a conflict between the shortage of equipment for testing environments and the rapid and accurate screening of suspected subjects. Further, RT-PCR testing is also reported to be not sensitive enough in the early stage [3] and suffer from high false-negative rates [4].

As competitive candidates and important complements to RT-PCR tests, the radiological imaging techniques, e.g., chest X-ray imaging and chest computed tomography (CT) imaging (see Fig. 1.), have also demonstrated effectiveness in current diagnosis. According to existing studies [7], CT scanning serves as an essential supplement in follow-up assessment and disease evolution evaluation. Moreover, similar observations [3] also suggest that the sensitivity of CT for COVID-19 infection is 98% compared to RT-PCR sensitivity of 71%, and radiological imaging may help support early screening of COVID-19. In contrast to RT-PCR testing, radiological imaging and the corresponding diagnosis can be obtained in a much faster way. However, the manual delineation of lung infections is tedious and time-consuming work. Besides, infection diagnosis

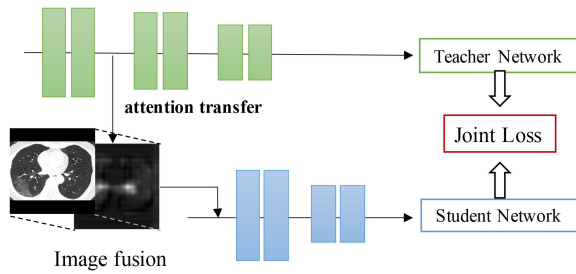


Fig. 2. Schematic representation of attention transfer network structure. The proposed architecture can be divided into a teacher network and a student network based on the attention transfer direction.

by clinicians and radiologists is a highly subjective task, often influenced by individual bias and clinical experience. To alleviate the enormous burden of reading radiological images for clinicians and improve diagnosis accuracy under the COVID-19 pandemic, automatic diagnosis systems are in great demand.

To improve the efficiency of radiological imaging-based diagnosis, automatic diagnostic systems have been developed with artificial intelligence (AI), which reads patients' X-ray or CT images as inputs and output the responding diagnostic results [7], [8]. However, most of these AI-related COVID-19 clinical decision support systems [9]–[13] are black-box deep learning models and lack proper explainability. Clinicians often feel reluctant to trust or understand such models because of non-transparency risks [14], [15], which is a significant barrier for broader adoption. Explainable Artificial Intelligence (XAI) is considered a novel research tool [16] to address some of the black-box AI system's restrictions by explaining their predictions and creating reliable models. Furthermore, XAI helps to improve the quality of predictions by introducing explainability and mitigating undesired biases [17]. A fully automatic diagnosis system without human verification would be unconscionable and potentially dangerous in piratical settings [18]. Thus, we aim to develop an explainable COVID-19 diagnosis model to enable clinical verification.

To solve the problem discussed above, we propose an explainable attention transfer network for the COVID-19 automatic diagnosis system in this study. The proposed network structure can be divided into teacher network and student network based on the attention transfer direction, as shown in Fig. 2. The teacher network extracts global features and concentrates on the infection regions with the proposed deformable attention module (DAM) to strengthen the response of infection regions and to suppress noise in irrelevant regions with an expanded reception field. Then an image fusion module combines attention knowledge transferred from teacher network to student network with the essential information in original input. While the teacher network focuses on global features, the student branch focuses on irregularly shaped lesion regions to learn discriminative features. The main contribution of our work is three-fold:

- An explainable attention transfer classification model based on the knowledge distillation network structure is designed to achieve COVID-19 automatic diagnosis with radiology. In this study, we utilize an attention mechanism

to transfer knowledge from the teacher network to the student network to improve model performance and provide network interpretation.

- We propose a deformable attention module to focus on irregularly shaped infection regions and their neighborhood in radiological images. Combining with local information, it can help deep networks pay more attention to infection regions and suppress noise in irrelevant regions with expanded reception fields. Specifically, the proposed attention module can serve as an interpretation tool, which can be flexibly inserted into existing convolutional architectures.
- We conduct extensive experiments on public available chest X-ray and CT datasets to evaluate the proposed multi-class classification model differentiating COVID-19, normal, and CAP cases. Moreover, our algorithm achieves the-state-of-art performance and improves the model's explainability by saliency map, severity assessment, and prediction confidence.

The rest of the study is organized as follows: Section II briefly introduces the related works; Section III shows the whole model structure and further explains proposed modules; Section IV presents experimental results with analysis; and Section V discusses the insights and future work.

II. RELATED WORKS

A. Computer-aided COVID-19 Diagnosis Research

While reading radiological image in diagnosing COVID-19, qualitative interpretation accompanied with quantitative analysis should be conducted to make radiology reporting much more comprehensive. To this end, investigators look for help from computer-aided methods to read and analyze X-rays and CT scans, aiming to diagnose and monitor COVID-19. Besides relatively high diagnosis accuracy, AI-related techniques can also play a role in exploring potential infection regions and other clinical tasks in radiological images.

Motivated by the high demand for rapid interpretation of chest X-ray images, many researchers seek help from deep learning models [9], [10], [19]–[21] to diagnose cases infected with COVID-19. Wang *et al.* [19] has proposed COVID-Net with a deep CNN designed to classify COVID-19 infection, pneumonia viral, pneumonia bacterial, and normal (non-COVID19 infection) X-ray imaging datasets. COVID-Net achieves an overall accuracy of 83.5% for four-category classification task and 92.5% of three-category (COVID-19, normal, and non-COVID pneumonia cases) classification task. Ozturk *et al.* [21] has presented a DarkCovidNet for automatic COVID-19 identification with DarkNet backbone to provide accurate diagnostics for multi-class classification and binary classification. Besides the traditional supervised learning methods, some researchers also make use of semi-supervised learning techniques [22] under the consideration that the number of COVID-19 cases is not abundant enough and is smaller compared to that of normal cases for traditional supervised learning.

Clinicians and radiologists can also read and analyze CT slices to identify certain characteristic visual features in the

lungs related to COVID-19, such as the bilateral and peripheral ground-glass opacity (GGO) in the early and the pulmonary consolidation opacity in the late stage [4], [23]. Xu *et al.* [11] has established an early screening model to differentiate COVID-19 from pneumonia and healthy cases using 618 pulmonary CT samples and achieved a total accuracy of 86.7%. Song *et al.* [12] has proposed a deep learning based automatic diagnosis system named DeepPneumonia with an overall accuracy of 86.0% and 94.0% for multi-class and binary classification, respectively.

Aside from the pure diagnosis tasks, there appears more related applications, like severity assessment [24], [25], large-scale screening [26], lung infection quantification [27] and uncertainty related problems in AI-based diagnosis [28], [29]. Among the models and methods mentioned above, we find that many methods lack comprehensive interpretations towards their results. Thus, we are willing to propose an explainable diagnosis system, aiming to form a thorough interpretation from different angles.

B. Attention Mechanism

To generate more interpretable results and help the model make more reasonable classifications, we borrow the idea from the attention mechanism, which is widely used in artificial intelligence related applications. The attention mechanism is a crucial aspect of human perception [30], enabling human beings to selectively focus on essential parts of the image, instead of processing the whole scene in its entirety. Simulating such selective attention mechanism of Human Visual System (HVS) is also vital for understanding mechanisms behind black-box neural networks. Attention mechanism has been widely applied to computer vision areas [30]–[34] and plays a gradually important role in more applications, equipping the model with some new characteristics: a) decide which part of the inputs to focus on; b) allocate limited computing resources to more important components. The attention modules can be roughly separated into channel-wise attention module, spatial-wise attention module, and self-attention module.

In spatial-wise mechanism [30], [31], [33], attention module in convolution neural networks localizes key information by utilizing the inter-spatial relationship from different location of feature maps. Jaderberg *et al.* [34] has proposed a spatial transformer networks (STN) module, which explicitly allows the spatial manipulation of data and equips the model the ability to transform feature maps within the network spatially.

In channel-wise mechanism [32], [35], attention module exploits inter-channel relationship with additional convolution layers, which represents the correlation between the current channel and the key information. The larger the weight is, the more attention we should pay to the channel. Squeeze-and-Excitation (SE) Networks [32] is proposed to model the importance of each channel via different learned weights. Furthermore, many researchers have combined spatial and channel attention mechanisms together [36]–[38] to take advantage of both. Fu *et al.* [37] has proposed dual attention networks (DANet) to adaptively integrate local features with their global dependencies in two types of attention modules.

Vaswani *et al.* [39] has introduced self-attention in transformer to draw global dependencies between input and output relying entirely on attention mechanism. Then Wang *et al.* [31] propose self-attention mechanism to computer vision areas with a non-local attention module. Non-local operation calculates the response at a position as a weighted sum of the features at all positions and implements correlation matrix to obtain the final attention map.

Attention mechanism has shown its superiority in a variety of medical image analysis tasks [40]–[42]. In particular, some state-of-the-art methods have been proposed to leverage the attention mechanisms to enhance the discriminative capability of classification models for both X-ray [43] and CT [44] image analysis task. Although attention has been widely applied in image processing and biomedical tasks, it has fewer applications to COVID-19 automatic diagnosis. Besides, the non-local attention module related methods [31] only calculate the pixel-to-pixel relationship and ignore the context information, which may not highlight infected regions in radiological images.

C. Visual Interpretability of Deep Networks

While AI-based models are extraordinarily powerful, adopting these algorithms in the medical domain has been limited [16]: even if physicians and regulators try to understand the implicit mathematical principles inside such models, they still need explicit declarative knowledge representation and explanatory structures to verify the prediction results. This means we need to build up systems to make decisions transparent, understandable, and explainable via XAI methods.

Visualization of deep network is the most direct way to explore visual patterns hidden in a neural unit [45]. Gradient-based methods [46]–[48] are widely applied in network visualization, mainly compute gradients of the score and estimate the image appearance that maximizes the unit score. In computer vision tasks, many prominent XAI methods [47], [48] often explain with a heatmap on input image, which generated from back propagation to serve as an overlay by building an abstract feature importance. Specifically, saliency analysis [47] is proposed to compute the gradient of the class score with respect to the input image to provide the final data-driven results; Layer-wise Relevance Propagation (LRP) [48] related methods help identify where neurons contribute the most to the higher-layer and generate explainable visualization results with the conservative relevance redistribution procedure.

In addition, Zhou *et al.* [49] has proposed Class Activation Map (CAM) to accurately compute the receptive field of neural activations in feature maps. It learns quite complete local abstraction with a one-to-one correspondence to input space from the convolutional layers. GradCAM [50] improves the original CAM method by using gradient information flowing into the last convolutional layer to understand each neuron for a decision of interest. The following GradCAM++ [51] provides better visual explanations in multiple object instances within a single image. Aside from other global-wise algorithms, Local Interpretable Model-agnostic Explanations (LIME) [52] uses a linear model

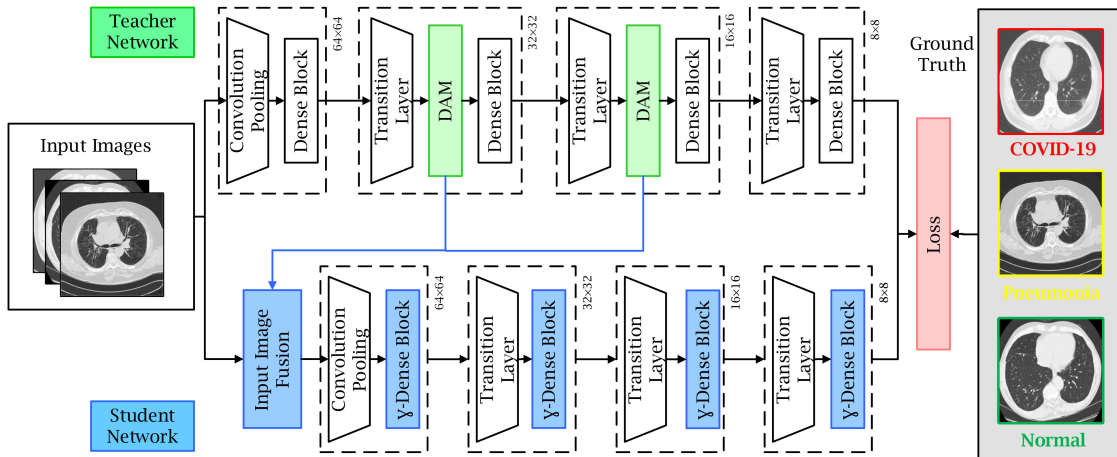


Fig. 3. Illustration of our proposed explainable COVID-19 classification model. The teacher network extracts global features and concentrates on the infection regions with the proposed DAM. The outputs of DAMs transfer attention knowledge and combine with essential information in original input via an image fusion module. Training with teacher network jointly, the student branch with weighted dense connectivity can focus on irregularly shaped lesion regions to learn discriminative features and improve network performance.

to approximate the black-box model in the vicinity of a specific input.

Several visualization methods have been adopted to improve the interpretable ability of AI-based COVID-19 models. Saliency detection methods [53] have been applied to the COVID-19 diagnosis system to interpret the proposed classification model. CAM [49] and its improvement models Grad-CAM [50], GradCAM++ [51] have also been widely utilized in COVID-19 diagnosis to establish explainable classification modules [20], [21], [54]. In this study, attention mechanism is utilized to generate visual explanations for classification model and it can also serve as an interpretation tool when inserted into other convolutional architectures flexibly.

III. THE PROPOSED METHOD

In this paper, we propose an explainable classification model to automatically differentiate COVID-19, CAP from healthy lungs in radiographic images. Following the original attention transfer network [55], the proposed network structure can be divided into teacher network and student network. The teacher network extracts global features and concentrates on the infection regions with the proposed deformable attention modules. Attention knowledge transfers from the teacher network to the student via an image fusion module. With knowledge borrowed from the teacher network, the student branch with weighted dense connectivity can focus on irregularly shaped lesion regions to learn discriminative features and improve network performance. We optimize the two networks jointly by minimizing the proposed joint loss function. The overall framework of our method is illustrated in Fig. 3.

A. Teacher Network with Deformable Attention Module

The teacher network extracts global features through a deep network and concentrates on the infection regions with the deformable attention modules. Following the basic architecture of DenseNet-169 [56], the teacher network comprises of four

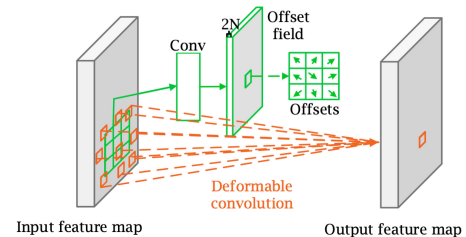


Fig. 4. Illustration of deformable convolution with learning offsets. The deformable convolution adds 2D offsets to the regular grid sampling locations in the standard convolution, enabling free deformation.

dense blocks, two deformable attention modules, and following transition layers. Dense block is composed of 6, 12, 24, 16 densely connected layers connects each layer to every other layer in a feed-forward fashion to ensure information reuse. Two deformable attention modules are inserted to estimate infected regions and track long-distance dependency information. Each transition layer does convolution and pooling with a batch normalization layer and a 1×1 convolutional layer followed by a 2×2 average pooling layer to reduce the dimension and channel of feature maps.

Considering the boundary of infected regions is irregularly shaped, we propose a deformable attention module to emphasize features in infection region through calculating the response at each position and generating a weighted sum of features. We implement deformable convolution operation [57] to learn a global spatial weights matrix, where each element indicates the cross correlation between regions in feature maps. As shown in Fig. 4, compared with standard convolution, the deformable convolution adds 2D offsets to the regular grid sampling locations, which enables free form deformation. In particular, the offsets are learned from the preceding feature maps via an additional convolution layer, including both horizontal and vertical directions. In a basic convolution layer, the sampling location p_k with a basic 3×3 convolution kernel can be expressed as

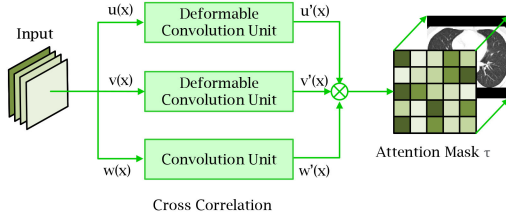


Fig. 5. Illustration of proposed deformable attention module. The input feature map first pass through two deformable convolution in horizontal and vertical direction separately, and then implemented by a standard convolution layer. The attention map is calculated by a correlation matrix.

$\mathbf{p}_k \in \mathcal{K} = \{(-1, -1), (0, -1), \dots, (1, 0), (1, 1)\}$. Unlike uniform sampling with fixed \mathbf{p}_k in normal convolutions, the regular grid \mathcal{K} is augmented with offsets $\Delta\mathbf{p}$ in deformable convolution, which enlarges the receptive fields around the infected regions gradually. Concretely, for a location \mathbf{p} in the input feature map \mathbf{x} , the output feature map $\mathbf{y}(\mathbf{p})$ of deformable convolution layer can be expressed as:

$$\mathbf{y}(\mathbf{p}) = \sum_{k=1}^{|\mathcal{K}|} \mathbf{w}(\mathbf{p}_k) \cdot \mathbf{x}(\mathbf{p} + \mathbf{p}_k + \Delta\mathbf{p}_k), \quad (1)$$

where the summation of sampled values are weighted by $\mathbf{w}(\mathbf{p})$.

The proposed deformable attention module implements two deformable convolution layers to generate the output attention map, as shown in Fig. 5. The input feature map first pass through two deformable convolution in horizontal and vertical direction separately, and then implemented by a 1×1 standard convolution. We utilize the deformable convolution layers to generate the attention mask towards the feature map obtained from convolution layer. Concretely, the input feature $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ of DAM is transferred into $u(\mathbf{x}), v(\mathbf{x}) \in \mathbb{R}^{\frac{C}{r} \times H \times W}$ through deformable convolution layers in two directions and $w(\mathbf{x}) \in \mathbb{R}^{C \times H \times W}$ through a standard convolution layer, with a dimensional reduction ratio $r = 4$. The spatial weights matrix can be computed by a normalized cross correlation (NCC) matrix τ :

$$\tau(i, j) = \frac{1}{n} \sum_{i, j} \frac{1}{\sigma_u \sigma_v} u'(i, j) v'(i, j), \quad (2)$$

where $u', v' \in \mathbb{R}^{\frac{C}{r} \times N}$ is arranged into sequences by u, v ; n is the number of elements in $u'(i, j)$ and $v'(i, j)$ and σ_u, σ_v is the corresponding standard deviation of u', v' . We implement the obtained spatial weights matrix on the feature map $w(\mathbf{x})$ gathered by a 1×1 convolution layer, and the output attention map \mathbf{A} of the DAM can be formulated by:

$$\mathbf{A} = \sum_{i=1}^N w(\mathbf{x}) \tau_{j, i}, \quad s.t. \sum_i \tau_{j, i} = 1. \quad (3)$$

Combining context with local information, the proposed attention module helps the teacher network highlight the response of infection regions and reduce noise in irrelevant regions in an expanded reception field achieved by deformable convolution. Compared with the self-attention module in Wang *et al.* [31], our deformable attention module can adaptively concentrate on irregularly shaped infection regions and its neighborhood to

integrate global and local information. Furthermore, the proposed attention module can be flexibly transferred into any deep neural networks to capture long-distance dependency and achieve explainable results.

B. Student Network with Attention Transfer

Knowledge distillation with neural networks was pioneered by Hinton *et al.* [58], aiming to improve the performance of a student network through the knowledge borrowed from a powerful teacher network. In this paper, we aim to improve a student network's training by relying on the attention knowledge borrowed from an instructive teacher network. Therefore, the original input fuses with the attention map which guides to infection regions for further improvement in performance in the student network. Additionally, we propose weighted dense connectivity in the original dense block to further aggregate the information flow between layers and improve network performance.

In image fusion processing, the prior attention information borrowed from the teacher network serves as the input of the student network to train it effectively. Considering the borrowed attention maps may lose some essential information in the early stage, we also involve the original input of teacher network \mathbf{I}_0^T in the fusion module. The fused input image for the student network \mathbf{I}_0^S is defined as:

$$\mathbf{I}_0^S = \frac{\mathbf{I}_0^T \oplus (\mathbf{I}_0^T \odot A'_0) \oplus (\mathbf{I}_0^T \odot A'_1)}{3}, \quad (4)$$

where A'_0, A'_1 indicates the average attention map among all channels of A_0, A_1 from 1st, 2nd DAM. The \oplus represents channel-wise dot addition operation and \odot indicates element-wise dot product operation. As the dimension of the attention maps A_0, A_1 is reduced by transition layers, we need to first resize A_0, A_1 (with resolution $32 \times 32, 16 \times 16$) to fuse with original image \mathbf{I}_0^T .

After fused with output attention map from teacher network, \mathbf{I}_0^S is fed to next γ -dense block with weighted dense connectivity. For the following DenseNet, it comprises L layers, each of which implements a non-linear transformation $\mathbf{H}_\ell(\cdot)$, where ℓ indexes the layer. $\mathbf{H}_\ell(\cdot)$ consists of a batch normalization, a Rectified Linear Units (ReLU), and a convolution of ℓ^{th} layer as in standard dense block. We denote the output and the corresponding weight scalar of ℓ^{th} layer as \mathbf{x}_ℓ and γ_ℓ , respectively. Consequently, the weighted dense connectivity of the ℓ^{th} layer can be formulated as:

$$\mathbf{x}_\ell = \mathbf{H}_\ell([\gamma_0 \mathbf{x}_0, \gamma_1 \mathbf{x}_1, \dots, \gamma_{\ell-1} \mathbf{x}_{\ell-1}]), \quad (5)$$

where $[\gamma_0 \mathbf{x}_0, \gamma_1 \mathbf{x}_1, \dots, \gamma_{\ell-1} \mathbf{x}_{\ell-1}]$ indicates the feature map's weighted concatenation generated in layers $0, 1, \dots, \ell - 1$. All the weights are set to 1 in the initialization and optimized with iteration. Specifically, we concatenate the multiple inputs of $\mathbf{H}_\ell(\cdot)$ in eq. (5) into a single tensor for ease of implementations.

The obtained attention map works as guider to transfer knowledge from teacher network to student. With input image fusion module, the outputs of two DAMs in teacher branch combine

with essential information in original image to help student network focuses on infection regions. Weighted dense connectivity assists to introduce adaptive connections from any layer to all subsequent layers. Trained with teacher network jointly, the student network can pay more attention to irregular shaped lesion regions to learn discriminative features and improve network performance.

C. Overall Objective Function

The cross-entropy loss together with softmax activation is considered as one of the most widely used loss functions in image classification tasks. The cross-entropy loss function can be formulated as $L_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_j^T x_i + b_j}}{\sum_{j=1}^n e^{W_j^T x_i + b_j}}$, where

$x_i \in \mathbb{R}^d$ denotes the extracted deep feature of the i^{th} sample and y_i represents the Ground Truth (GT). $W_j \in \mathbb{R}^d$ indicates the weights for j^{th} class in the Fully Connected (FC) layers, and $b_j \in \mathbb{R}^n$ is the bias term. The batch size and class number are N and n , respectively. Despite its simplicity and popularity, the cross-entropy loss does not explicitly optimize the embedding feature to maximize the inter-class margin distance and encourage discriminative learning of features.

To address this issue, we modify the cosine loss [59] in face recognition to propose a Discriminative Cosine (DC) loss for the separate training of each branch. For simplicity, we fix the bias $b_j = 0$ as in the following and transfer the logit as $W_j^T x_i = \|W_j^T\| \|x_i\| \cos \theta_j$, where θ_j is the angle between weight W_j^T and feature x_i . Then we fix the individual weight $W_j^T = 1$ and re-scale embedding feature $x_i = s$ by l_2 normalization. The learned embedding features are thus distributed on a hypersphere with a radius of s . Besides, the regularization term is also involved to enlarge the inter-class variances and enlarge weights discrepancy in fully connected layer. Therefore, the proposed DC loss can be formulated as:

$$L_{DC} = \frac{1}{N} \sum_i -\log \frac{e^{s(\cos(\theta_{y_i}, i) - m)}}{e^{s(\cos(\theta_{y_i}, i) - m)} + \sum_{j \neq y_i} e^{s \cos(\theta_j, i)}} + \frac{1}{n(n-1)} \sum_{y_i=1}^n \sum_{j \neq y_i}^n W_{y_i}^T W_j, \quad (6)$$

where $\cos(\theta_j, i) = W_j^T x_i$ and hyper-parameter m indicates the the angular margin between different classes.

In distillation, knowledge is learnt by the teacher network and then transferred to the student network by minimizing a loss function, where the target is the distribution of class probabilities predicted by the teacher model. Intuitively, prediction results obtained from either teacher network or student network should maintain the same. Therefore, we utilize the Jensen-Shannon (JS) divergence [60] to measure the difference between two distributions from teacher network and student network. We denote the probabilities of class j from teacher and student network as p_j^t, p_j^s and $p_{avg} = \frac{p_j^t + p_j^s}{2}$ for simplification, where

$p_j = \frac{e^{s \cos(\theta_j, i)}}{\sum_{j=1}^n e^{s \cos(\theta_j, i)}}$. Then the JS loss can be calculated by:

$$L_{JS} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n \frac{1}{2} p_j^t \log \frac{p_j^t}{p_{avg}} + \frac{1}{2} p_j^s \log \frac{p_j^s}{p_{avg}}. \quad (7)$$

The lower the JS divergence value, the better the two prediction distributions have matched with each other.

To combine the two loss function terms, we introduce a time-dependent weighting function $w(t)$ to scale the JS loss. It assists to effectively train the student when the teacher network converges to a relatively stable situation. The overall objective function can be formulated as:

$$L = L_{DC} + w(t) \cdot L_{JS}. \quad (8)$$

We freeze and update parameters in teacher/student network alternatively in turn to optimize both networks simultaneously until convergence. With DC loss, either network acquires discriminative features with larger inter-class variances. With JS loss function, the student network is optimized collaboratively with the teacher network and gradually matches the probability distribution by minimizing L_{JS} .

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We implement proposed model to COVID-19 related X-ray and CT dataset to test the effectiveness and provide explainable results of our method. Due to GPU memory constraints, we utilize 128×128 dimension images to train the network. General radiological image pre-processing techniques like histogram equalization are implemented with python scikit-image library to enhance image quality. We implement data augmentation with random flipping and rotation on training data, and no augmentation is performed in test data. We use Adam optimizer [61] available in TensorFlow with hyperparameter values $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Training is performed using a batch size of 16. Hyperparameters in loss function are set to $s = 64$ and $m = 0.15$. Following [62], we ramp up the weight parameter $w(t)$ and learning rate during the first 50 epochs with weight $w(t) = \exp[-5(1 - \frac{t}{50})^2]$ and ramp down the learning rate during the rest epochs with $w(t) = \exp[-12.5(1 - \frac{100-t}{50})^2]$.

A. Data Collection

Our experiments are conducted on both COVID-19 related chest CT and X-ray image dataset separately. Examples of both CT and X-ray images from the proposed dataset are shown in Fig. 1.. In our experiments, each collected dataset is randomly shuffled into three subsets: 70% for training, 10% for validation, 20% for test.

In this study, CT images from several public datasets are utilized to train and evaluate our algorithm. We adopt 349 COVID-19 positive CT images and 384 normal class (no findings) CT images in COVID-CT-Dataset [63]. Furthermore, 304 chest CT images labeled as CAP are collected from Radiopaedia [64] to build a relatively balanced CT image dataset.

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-ART METHODS ON CHEST X-RAY DATASET.

Method	Class Label	Recall	Precision	F1-score	Overall Accuracy (OA)	CK-score
Khan et al. [68]: CORONet	Normal	0.9727	0.9453	0.9588	91.98%	0.8564
	COVID-19	0.8901	0.7713	0.8265		
	CAP	0.8757	0.7613	0.8145		
Wang et al. [19]: COVID-Net	Normal	0.9789	0.9601	0.9700	92.52%	0.8673
	COVID-19	0.8901	0.7874	0.8356		
	CAP	0.8817	0.7766	0.8258		
Karim et al. [20]: DeepCOVID-Explainer	Normal	0.9628	0.9533	0.9580	92.86%	0.8776
	COVID-19	0.9040	0.8733	0.8884		
	CAP	0.9015	0.8563	0.8784		
Ours	Normal	0.9702	0.9657	0.9679	93.44%	0.8885
	COVID-19	0.9231	0.8904	0.9065		
	CAP	0.9023	0.8732	0.8875		

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-ART METHODS ON CT DATASET.

Method	Class Label	Recall	Precision	F1-score	Overall Accuracy (OA)	CK-score
ResNet [69]	Normal	0.9342	0.8902	0.9117	83.17%	0.8248
	COVID-19	0.7297	0.8308	0.7770		
	CAP	0.8276	0.7385	0.7805		
Xu et al. [11]: ResNet + Location attention	Normal	0.9605	0.8824	0.9198	86.54%	0.7966
	COVID-19	0.8108	0.8902	0.8486		
	CAP	0.8103	0.8103	0.8103		
Wang et al. [70]: COVID-Net-CT	Normal	0.9868	0.8929	0.9375	87.50%	0.8105
	COVID-19	0.8514	0.8750	0.8630		
	CAP	0.7931	0.8462	0.8188		
Ours	Normal	0.9737	0.8916	0.9308	87.98%	0.8248
	COVID-19	0.8649	0.9014	0.8828		
	CAP	0.7758	0.8491	0.8108		

Regarding X-ray data collection, we collect 450 COVID-19 X-ray images diagnosed with COVID-19 from the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 database [65] and a COVID-19 X-ray dataset developed by Cohen *et al.* [66] using images from various open access sources. Besides, 1800 normal class (no findings) and 1837 CAP class frontal chest X-ray images are randomly adopted from the National Institutes of Health (NIH) Chest X-ray Dataset [67].

B. Evaluation Metrics

We make a confirmation and definition of the evaluation metrics utilized to assess our model's performance. In this study, we are targeting to solve a three-category classification problem separating COVID-19, normal, CAP cases. We utilize precision, recall, F1-score per class and overall accuracy (OA) to evaluate the classification performance. Additionally, we also evaluate the multi-classification results using Cohen's kappa (CK) score. Cohen's kappa statistic measures inter-rater reliability when faced with imbalanced-class or multi-class classification problems. It is generally considered as good multi-class classification algorithm with Cohen's kappa score (usually ranges from 0 to 1) above 0.8.

C. Comparison with State-of-the-art Methods

To further get to know how our model performs, we offer **Table I** and **Table II** as our multi-classification performances. The tables also include comparison on different metrics with other state-of-art methods. More specifically, **Table I** shows the main results on chest X-ray dataset and **Table II** illustrates the performances of different models on CT dataset. For settings on chest X-ray dataset, we discuss the following three existing literature:

- **CORONet** [68] is a deep CNN model using Xception architecture for automatic COVID-19 diagnosis focusing on detecting lung infections in chest X-ray data;
- **COVID-Net** [19] utilizes a DenseNet-similar structure to track long-distance connectivity and applies GSInspire to achieve the critical factors leading to the classification;
- **DeepCOVIDExplainer** [20] is constructed based on the ResNet18 structure to accomplish the classification task while it also utilize GradCAM for visualization;

On the other hand, we also present the results of three the-state-of-art method experimented on CT dataset:

- **ResNet** [69] is a widely applied model and architecture in computer vision tasks, which is also very popular in many CT-based COVID-19 diagnosis systems, playing an important role in classification functional modules;

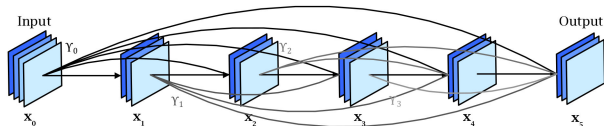


Fig. 6. Illustration of proposed γ -dense block with weighted dense connectivity. All the weights γ are set to 1 in the initialization and optimized with iterations.

- **ResNet+Location attention** [11] takes use of location attention mechanism to make the features learnt from ResNet18 and ResNet23 explicitly;
- **COVID-Net-CT** [70] share similar structure with COVID-Net, only make some small adjustments towards CT dataset.

Quantitatively, for experiments on chest X-ray dataset, our proposed method obtains the recall of 0.9231, precision of 0.8904, F1-score of 0.9065, a 93.44% overall accuracy and a 0.8885 *CK*-score for the COVID-19 category outperforms all the other state-of-art methods. For the other classes, our model performs the best in all the metrics (recall, precision, and F1-score) on pneumonia and the precision score of normal people class. For the other experiment setting on CT dataset, our proposed method achieves highest COVID-19 recall 0.8649, highest COVID-19 precision 0.9014, highest COVID-19 F1-score 0.8828, highest overall accuracy 87.98% and highest *CK*-score 0.8248 among all the state-of-art methods. So does our model on pneumonia precision score.

The main misclassification often occurs between COVID-19 positive and CAP cases, which are sometimes challenging for an experienced clinician. Without pre-trained on large scale pathological radiography imaging datasets like [19], it is acceptable for our proposed method does not perform the best on some of the metrics regarding normal and pneumonia class. The classification results validate that our method can effectively extract discriminative features for COVID-19 radiological images and make relatively high-accuracy predictions for automatic diagnosis.

D. Explainable Results

1) **Model Interpretation and Visualization:** The final output attention map of proposed model can be calculated as $\mathbf{A} = \mathbf{A}_0 + \mathbf{A}_1$, where $\mathbf{A}_0, \mathbf{A}_1$ indicates the output of the 1st, 2nd DAM. We compare our attention map combining the outputs of two DAMs with several state-of-art model interpretation method GradCAM [50], GradCAM++ [51], and LRP [48]. In general, the more accurate an algorithm is, the more consistent the visualizations of attention maps will be. Key features can then easily be identified based on where the attention maps are overlapping. As shown in Fig. 7, the heat-maps are overlapped on the original images with the red color highlighting the activation region associated with the predicted class. The other three model interpretation methods are implemented on the COVID-Net [19] with good classification performance to provide explainable results. It can be seen that our calculated attention maps successfully highlight more detailed infection regions while other approaches sometimes fail to capture key features (e.g. (b)

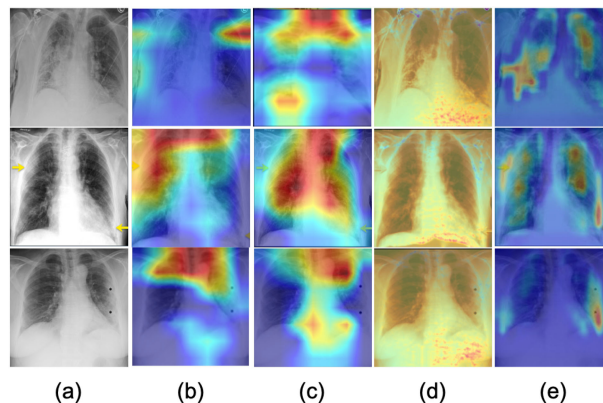


Fig. 7. Comparison of different network visualization and model interpretation methods. The heatmaps are overlapped on the original COVID-19 images, the red color highlights the activation region associated with the predicted class. The intensity of colors on the heatmap corresponds to importance of features for the prediction of COVID-19 positivity. (a) Original image (b) GradCAM [50] (c) GradCAM++ [51] (d) LRP [48] (e) Ours

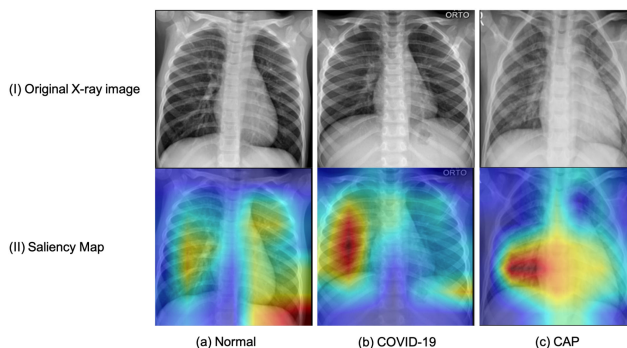


Fig. 8. Representative chest X-ray images (I) and corresponding saliency maps (II) of (a) the healthy, (b) COVID-19, and (c) CAP. The heatmaps are overlapped on the original image, where the red color highlights the activation region associated with the predicted class.

GradCAM in the 1st row) or get distracted by some irrelevant information outside of lungs area like skeletal structure (e.g. (b) GradCAM, (c) GardCAM++ in the 3rd row). Moreover, since our method combining with local information and takes features around infection regions, the obtained attention maps enlarged the reception field to expand highlighted infection regions.

From the heatmaps in Fig. 8, we can verify that our model is not making decisions based on inappropriate regions of the radiology images. We can observe that the focus area for (b) COVID-19 and (c) CAP was explicitly different from the one for (a) the healthy. According to the saliency maps, we also found that the network focused on different regions when classifying COVID-19 and CAP. For CAP cases, the model paid more attention to effusion and consolidation adjacent to the pleura. On the other hand, the network focused more on GGO rather than consolidation for COVID-19 subjects.

Consequently, precise decisive feature localization is crucial for both model interpretation and rapid confirmation of reliability of outcomes [20]. Attention map highlights the critical regions on the radiological image and provides an explainable result of a prediction model. It offers insight to clinicians with

TABLE III

RADIOLOGICAL SCORING PERFORMED BY THREE BLINDED EXPERTS, INDICATING THE EXTENT OF GROUND-GLASS OPACITY IN EACH LUNG (RIGHT AND LEFT LUNG).

Score	Involvement
1	<25% involvement
2	25-50% involvement
3	50-75% involvement
4	>75% involvement

TABLE IV

COMPARISON OF SEVERITY SCORE PERFORMANCE METRICS OF OUR METHOD WITH THE-STATE-OF-ART.

Method	Correlation Coefficient	R^2
DLNet1 [25]	0.75	0.38
DLNet2 [25]	0.83	0.67
Ours	0.76	0.57

the process of making more accurate diagnosis and correcting the potential misdiagnosis in AI-based model. Besides ensuring trustworthiness, it will also provide new insights and visual indicators of potential clinical factors regarding COVID-19.

2) *Severity Score*: For COVID-19 cases, the proposed X-ray dataset contains 94 PA chest X-ray images from Cohen *et al.* [66] assigned with severity score [24] and lung mask [71]. Each image is assigned with a severity score by three experts, indicating the extent of GGO or consolidation in each lung (right and left lung). Calculated by right lung and left lung together, the final extent score ranges from 0 to 8, as shown in Table III. Following the definition in [25], we compute the ‘‘pneumonia ratio’’ for each lung. We divide the ‘‘pneumonia ratio’’ into four levels following the same GT criterion, and the total score is summed for both lungs. Pneumonia ratio is calculated by:

$$\text{Pneumonia Ratio} = \frac{\text{Area}_{\text{pneumonia}}}{\text{Area}_{\text{lungs}}}, \quad (9)$$

where the $\text{Area}_{\text{lungs}}$ is computed according to the total number of pixels involved in the lung mask. Specifically, the attention map is multiplied by the lung mask to restrict pneumonia infected regions to the lung area. We restrict the attention map in $\text{Area}_{\text{lungs}}$ provided by lung mask and discard pixels outside. We then further define $\text{Area}_{\text{pneumonia}}$ by the area in the normalized attention map (re-scale to 1) with a lower bound restricted by threshold $T = 0.5$.

As shown in Table IV, we can observe the predicted severity scores against ground truth with correlation coefficient of the fitted model being 0.76 and $R^2 = 0.57$. Compared with specific pneumonia localization algorithm DLNet[25], our attention map achieves a relatively similar performance which indicates the effectiveness of our severity assessment method and attention map. The severity of COVID-19 pneumonia is directly associated with its extent in the lungs; thus, we can improve the performance with an accurate segmentation of regions infected with pneumonia in the future work. Besides, the limitation of sample size prevents proper cohort selection. It is possible to further improve the generalization of our model with more radiology data labelled the same severity score criterion.





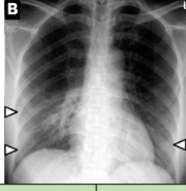

GT: COVID-19		GT: Normal		GT: CAP	
					
COVID-19	0.6774	COVID-19	0.2103	COVID-19	0.3677
Normal	0.1774	Normal	0.6445	Normal	0.0556
CAP	0.1452	CAP	0.1452	CAP	0.5767
GT: CAP		GT: COVID-19		GT: COVID-19	
					
COVID-19	0.4893	COVID-19	0.3498	COVID-19	0.3910
Normal	0.0807	Normal	0.4272	Normal	0.1596
CAP	0.4300	CAP	0.2230	CAP	0.4493

Fig. 9. Model confidence towards different sample points. GT stands for ground truth of corresponding figure. The three chest X-ray images in the upper line (with GT label in green) are classified correctly by the proposed model, while images in the bottom (with GT label in red) indicate misdiagnosed cases. In each case, the predicted label with highest prediction confidence score is changed to bold text with correct prediction in green and wrong in red.

3) *Prediction Confidence*: As is known to all, the COVID-19 diagnosis process is a complicated decision-making system and it may cause disastrous consequences if incorrect conclusion is obtained. Thus, the results output from the proposed model will be more convincing and explainable if an uncertainty or confidence score is given accompanying the normal classification prediction. We utilize a simple Softmax function to form a primary model confidence score towards each data, which also represents the probability that the model thinks its answer might be correct:

$$\text{Conf}_i = \max_{j \in [0, K-1]} \frac{e^{z_{ij}}}{\sum_{k=0}^{K-1} e^{z_{ik}}}, \quad i \in [0, N-1]. \quad (10)$$

where z_{ij} represents the j -th class prediction score of the i -th data obtained from our proposed model; N represents the data size and K represents the number of classes, which is 3 in this application. With the confidence score, we can see how confident the model is towards different given data.

For better understanding, we give a brief view of examples with the X-ray imaging dataset and offer the corresponding confidence score in Fig. 9. The figures on the first row with green ground-truth labels are correctly-classified images and the figures on the second row with red labels are mis-classified images. For the mis-classified figures, the labels and scores in red text are the wrong predictions given by our proposed model and the green labels and scores are the ground-truth labels. It can be easily noticed that the mild COVID-19 patients with smaller shadow area in lungs in the second row can be classified as normal by mistake. When diagnosing these patients, the model seems to be not that confidence compared to the correctly-classified severe patients in the first row.

TABLE V
ABLATION STUDY METRICS ON CT DATASET.

	COVID-19 Recall	COVID-19 Precision	COVID-19 F1-score	Overall Accuracy
DenseNet	0.7297	0.8852	0.7999	0.8365
DenseNet w/ γ -DB	0.7432	0.8871	0.8088	0.8413
DenseNet w/DAM	0.8378	0.9000	0.8678	0.8606
DenseNet w/DC loss	0.8108	0.8955	0.8510	0.8558
Ours	0.8649	0.9014	0.8828	0.8798

With the aid of the confidence score, we may avoid making wrong decisions and classifications at the early stage of the disease. If the model confidence score is below certain threshold (like 0.5 in this three-class classification scenery), we can consult some experts and clinicians for advise. Besides misdiagnosed/failure cases, we should also be alert to correct prediction with relatively low confidence, which may indicate a random guess of model. The combination of AI automatic diagnose system and the experience of doctors can both reduce the uncertainty and risk of making wrong diagnosis and improve the efficiency of the whole diagnosing process.

E. Ablation Study

To validate the effectiveness of various components in the proposed architecture, we design an ablation experiment based on DenseNet structure to assess each module’s characteristics. Here we focus on precision, recall, and F1-score of COVID-19 positive class in the CT dataset, which is relatively class-balanced than X-ray. To analyze the contributions of the proposed model, Table V quantitatively shows the performance of the baseline models and our proposed method. On the 1st row, we show the results of the backbone DenseNet-169 [56]. To further verify each component’s validity, we conducted several comparative experiments on combining backbone and a single functional module and the corresponding results are shown from 2nd to 4th row. At last, to show the validity of the composition method of these functional methods, we also offer the results of our method on the 5th row. Specifically, γ -DB in the second row denotes the γ -dense block with weighted dense connectivity. In general, it is clear that each proposed single functional module: DAM, weighted dense connectivity, and DC loss all make contribution to the promotion of performance. With either component involving, it achieves relatively better performance compared with original DenseNet backbone. From the last row in Table V, we can notice that the proposed model outperforms the other ablation models. Thus, we may confirm that both components make improvements and work well with joint network structure.

1) *The Effectiveness of DAM:* We analyze the effectiveness of DAM by visualizing the extracted attention maps and learned offsets fields from two DAMs in the teacher network. In Fig. 10, column (a) shows the original input CT images, (b),(c) show the learned offset fields, and (d),(e) represent the obtained attention maps. From (b),(c) in this figure, it can be observed that the respective fields become larger at infection regions. Extended reception help to provide comprehensive attention maps and improve classification performance. From the (d),

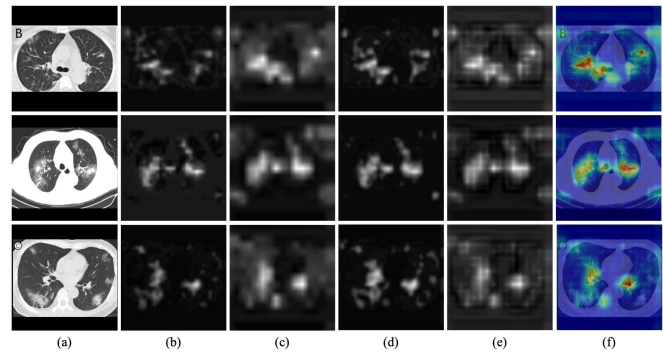


Fig. 10. Visualization of proposed deformable attention module. (a) original input images. (b),(c) learned offset fields of the 1st and 2nd DAM. (d),(e) attention maps of the 1st and 2nd DAM. (f) the final attention maps.

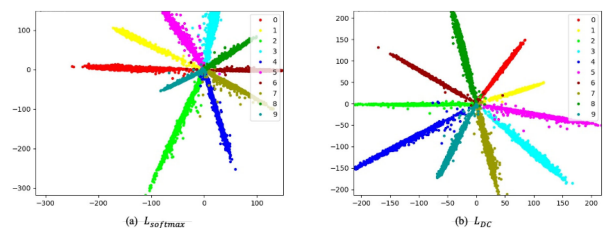


Fig. 11. Visualization of features learned with different (a) softmax loss L_{CE} (b) proposed L_{DC} functions on MNIST dataset. Each color denotes one class from digit 0 to 9.

(e) columns, it is clear to see that both attention maps are able to highlight infected regions in original image. The attention map at large scale includes more detailed context information, while the smaller one contains more structure information. To quantitatively evaluate the performance of DAM, we compare the teacher network (single DenseNet structure with DAM) with original DenseNet. It is evident that the proposed DAM performs better with an improvement of 2.41% in overall accuracy.

2) *The Effectiveness of DC Loss:* We replace the original softmax cross entropy loss in baseline model with DC loss to evaluate its performance. The relatively significant improvement of about 8.11% in recall, 5.11% in F1-score, and 1.93% in overall accuracy compared with baseline indicates that DC loss can facilitate the distinction of learned features and strengthen the proposed model’s robustness. Cosine loss helps further maximize the decision margin in the angular space to solve the original softmax loss lacks the power of discrimination; the regularization term also gives assistance to enlarge the inter-class variances. Moreover, visualization of features learned with cross entropy loss and DC loss on the MNIST dataset in Fig. 11 can also demonstrate the improvement of DC loss. Compared with softmax loss, converged DC loss can further increase the inter-class separability and the intra-class compactness. The quantitative experiment and feature visualization in general dataset demonstrate the discriminative ability of DC loss.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose an explainable attention transfer classification model based on the knowledge distillation network

structure to automatically differentiate COVID-19, CAP from healthy lungs with radiographic imaging. The teacher network extracts global features and concentrates on the infection regions to generate attention maps. We propose a deformable attention module to reinforce the response of infection regions and reduce noises in irrelevant regions with expanded reception field. Moreover, combining essential information in original input, attention knowledge transfers from teacher network to student via an image fusion module. Trained with teacher network jointly, the student branch with weighted dense connectivity can focus more on irregularly shaped lesion regions to learn discriminative features and improve network performance. Lastly, we conduct extensive experiments on public chest X-ray and CT imaging datasets to demonstrate the explainability of the proposed architecture in diagnosing COVID-19.

In this work, we have applied network visualization to highlight the potentially infected regions to explain the classification of CAP or COVID-19 infected regions from non-infected ones. Our goal is to provide clinicians with explainable AI tools so that they can diagnose COVID-19 cases more quickly and objectively. In the next phase of this project, we plan to include newly released data in the teacher network to build a semi-supervision model [72], which is expected to fully utilize the advantage of knowledge distillation network structure and further improve prediction accuracy. Moreover, it remains challenging to quantify correctness of model interpretability as current evaluation approaches mainly require subjective input from humans [73]. We aim to extend the quantitative evaluation of interpretation results (e.g., Dice similarity coefficient [74]) when the ground truth data of infection regions is available in the future. In addition, we will explore specific image features that can separate COVID-19 from CAP, and will combine the radiology findings with other data such as epidemiological histories, clinical characteristics, and hematological analysis to further improve the diagnosis accuracy in a multi-modality integration framework [75].

REFERENCES

- [1] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *Lancet*, vol. 395, no. 10223, pp. 470–473, 2020.
- [2] W. Wang *et al.*, "Detection of SARS-CoV-2 in different types of clinical specimens," *JAMA*, vol. 323, no. 18, pp. 1843–1844, May 2020.
- [3] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, 2020, 296(2), E115–E117, Art. no. 200432.
- [4] 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*, 2020, 296(2), E32–E40, Art. no. 200642.
- [5] H. X. Bai *et al.*, "Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT," *Radiology*, vol. 296, no. 2, pp. E46–E54, 2020.
- [6] J. Cleverley, J. Piper, and M. M. Jones, "The role of chest radiography in confirming COVID-19 pneumonia," *BMJ*, vol. 370, 2020.
- [7] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021.
- [8] X. He *et al.*, "Sample-efficient deep learning for COVID-19 diagnosis based on CT scans," *medRxiv*, 2020.
- [9] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images," 2020, *arXiv:2003.11055*.
- [10] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, pp. 635–640, 2020.
- [11] X. Xu *et al.*, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.
- [12] Y. Song *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *medRxiv*, 2020.
- [13] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, "Advancing medical imaging informatics by deep learning-based domain adaptation," *Yearbook Med. Inform.*, vol. 29, no. 1, p. 129, 2020.
- [14] W. Hsu, M. K. Markey, and M. D. Wang, "Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 6, pp. 1010–1013, 2013.
- [15] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proc. 8th ACM Int. Conf. Bioinf., Comput. Biol., Health Inform.*, 2017, pp. 233–240.
- [16] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" 2017, *arXiv:1712.09923*.
- [17] Y. Zhang, H. Wu, H. Liu, L. Tong, and M. D. Wang, "Improve model generalization and robustness to dataset bias with bias-regularized learning and domain-guided augmentation," 2019, *arXiv:1910.06745*.
- [18] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, "Biological interpretation of morphological patterns in histopathological whole-slide images," in *Proc. ACM Conf. Bioinf., Comput. Biol. Biomed.*, 2012, pp. 218–225.
- [19] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," 2020, *arXiv:2003.09871*.
- [20] M. R. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, and O. D. Beyan, "DeepCOVIDExplainer: Explainable COVID-19 predictions based on chest X-ray images," 2020, *arXiv:2004.04582*.
- [21] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, 2020, Art. no. 103792.
- [22] J. Zhou, B. Jing, and Z. Wang, "SODA: Detecting COVID-19 in chest X-rays with semi-supervised open set domain adaptation," 2020, *arXiv:2005.11003*.
- [23] Z. Ye, Y. Zhang, Y. Wang, Z. Huang, and B. Song, "Chest CT manifestations of new coronavirus disease 2019 (COVID-19): A pictorial review," *Eur. Radiol.*, vol. 30, pp. 4381–4389, 2020.
- [24] J. P. Cohen *et al.*, "Predicting COVID-19 pneumonia severity on chest X-ray with deep learning," 2020, *arXiv:2005.11856*.
- [25] R. Amer, M. Frid-Adar, O. Gozes, J. Nassar, and H. Greenspan, "COVID-19 in CXR: From detection and severity scoring to patient disease monitoring," 2020, *arXiv:2008.02150*.
- [26] F. Shi *et al.*, "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," 2020, *arXiv:2003.09860*.
- [27] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, *arXiv:2003.04655*.
- [28] D. Di *et al.*, "Hypergraph learning for identification of COVID-19 with CT imaging," 2020, *arXiv:2005.04043*.
- [29] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection," 2020, *arXiv:2003.10769*.
- [30] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.
- [31] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
- [32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [33] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 680–697.
- [34] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
- [35] A. Bastidas and H. Tang, "Channel attention networks," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 881–888.
- [36] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [37] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

- [38] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.
- [39] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [40] B. D. de Vós, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2017, pp. 204–212.
- [41] I. Yoo, D. G. Hildebrand, W. F. Tobin, W.-C. A. Lee, and W.-K. Jeong, "ssEMnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features," in *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2017, pp. 249–257.
- [42] W. Shi, L. Tong, Y. Zhuang, Y. Zhu, and M. D. Wang, "EXAM: An explainable attention-based model for COVID-19 automatic diagnosis," in *Proc. 11th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2020, pp. 1–6.
- [43] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2016–2027, Jul. 2020.
- [44] R. Xu *et al.*, "Pulmonary textures classification via a multi-scale attention network," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2041–2052, Jul. 2020.
- [45] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.
- [46] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
- [47] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
- [48] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.
- [49] 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.*, 2016, pp. 2921–2929.
- [50] 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.*, 2017, pp. 618–626.
- [51] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.
- [52] M. T. Ribeiro, S. Singh, and C. Guestrin, "“why should i trust you?” explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.
- [53] Y.-H. Wu *et al.*, "JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation," 2020, *arXiv:2004.07054*.
- [54] X. Ouyang *et al.*, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, Aug. 2020.
- [55] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.
- [56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
- [57] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
- [58] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [59] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.
- [60] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
- [61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [62] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.
- [63] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*.
- [64] C. H. Henry Knipe, "Radiopaedia pneumonia dataset," Website, 2020. [Online]. Available: <https://radiopaedia.org/articles/pneumonia>
- [65] I. S. of Medical and I. R. (SIRM), "COVID-19 database," 2020. [Online]. Available: <https://www.sirm.org/category/senza-categoria/COVID-19/>
- [66] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 image data collection: Prospective predictions are the future," 2020, *arXiv:2006.11988*.
- [67] 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.*, 2017, pp. 2097–2106.
- [68] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images," *Comput. Methods Prog. Biomed.*, 2020, Art. no. 105581.
- [69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [70] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images," 2020, *arXiv:2009.05383*.
- [71] R. Selvan *et al.*, "Lung segmentation from chest X-rays using variational data imputation," *ICML Workshop Art Learn. Missing Values*, Jul. 2020, *arXiv:2020.2005.10052*.
- [72] L. Tong, H. Wu, and M. D. Wang, "CAESNet: Convolutional autoencoder based semi-supervised network for improving multiclass classification of endomicroscopic images," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1286–1296, 2019.
- [73] Y.-S. Lin, W.-C. Lee, and Z. B. Celik, "What do you see? evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors," 2020, *arXiv:2009.10639*.
- [74] M. Nei and W.-H. Li, "Mathematical model for studying genetic variation in terms of restriction endonucleases," *Proc. Nat. Acad. Sci. USA*, vol. 76, no. 10, pp. 5269–5273, 1979.
- [75] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.