Multiscale Attention Guided Network for COVID-19 Diagnosis Using Chest X-Ray Images

Jingxiong Li, Yaqi Wang, Shuai Wang, Jun Wang, Jun Liu, Qun Jin, and Lingling Sun

***Abstract*—Coronavirus disease 2019 (COVID-19) is one of the most destructive pandemic after millennium, forcing the world to tackle a health crisis. Automated lung infections classification using chest X-ray (CXR) images could strengthen diagnostic capability when handling COVID-19. However, classifying COVID-19 from pneumonia cases using CXR image is a difficult task because of shared spatial characteristics, high feature variation and contrast diversity between cases. Moreover, massive data collection is impractical for a newly emerged disease, which limited the performance of data thirsty deep learning models. To address these challenges, Multiscale Attention Guided deep network with Soft Distance regularization (*MAG-SD*) is proposed to automatically classify COVID-19 from pneumonia CXR images. In *MAG-SD*, *MA-Net* is used to produce prediction vector and attention from multiscale feature maps. To improve the robustness of trained model and relieve the shortage of training data, attention guided augmentations along with a soft distance regularization are posed, which aims at generating meaningful augmentations and reduce noise. Our multiscale attention model achieves better classification performance on our pneumonia CXR image dataset. Plentiful experiments are proposed for *MAG-SD* which demonstrates its unique advantage in pneumonia classification over cutting-edge models. The code is available at https://github.com/JasonLeeGHub/MAG-SD.**

***Index Terms*—COVID-19, x-ray radiology, multiscale attention, convolutional neural network.**

Manuscript received July 13, 2020; revised October 29, 2020 and January 3, 2021; accepted February 4, 2021. Date of publication February 9, 2021; date of current version May 11, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61827806 and in part by Biomedical Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University under Grant YG2020YQ17. *(Corresponding author: Yaqi Wang.)*

Jingxiong Li, Jun Liu, and Lingling Sun are with the Key Lab of RF Circuits and Systems of Ministry of Education, Microelectronics CAD Center, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China (e-mail: jingxiong.li2019@outlook.com; ljun77@hdu.edu.cn; sunll@hdu.edu.cn).

Yaqi Wang is with the College of Media Engineering, Communication University of Zhejiang, Hangzhou, Zhejiang 310018, China (e-mail: wangyaqi@hdu.edu.cn).

Shuai Wang is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: shuaiwang.tai@gmail.com).

Jun Wang is with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wjcy19870122@sjtu.edu.cn).

Qun Jin is with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokorozawa 359-1192, Japan (e-mail: jin@waseda.jp).

Digital Object Identifier 10.1109/JBHI.2021.3058293

I. Introduction

THE coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is one of the most devastating infectious diseases after millennium [1]. This new type of coronavirus is announced in late December, 2019, then spread globally in 2020. It has been declared as a pandemic by World Health Organization (WHO) according to its high contagiosity and unprecedented pressure bought to public healthcare system [2]. The current gold-standard for screening COVID-19 is polymerase chain reaction (PCR) laboratory test, however, the test capacity is extremely limited and requires professional equipment [3]. [4] also reports that PCR tests suffers from high false negative rate.

Radiological images collected by X-ray and computed tomography (CT) are important complements to PCR tests. The virus leads to pneumonia, which is an inflammatory condition of the lung's air sacs [5]. Radiological signs show ground-glass opacity, airspace opacities and later consolidation with bilateral, peripheral, and lower zone predominant distributions [6]. Comparing with CT imaging, CXR diagnosis provides a low-cost and time-saving diagnosis method [7]. Besides, underdeveloped regions can hardly have sufficient CT scanners, making CT based COVID-19 screening impossible. X-rays are the most common diagnostic imaging equipment available even in rural regions, which means X-ray diagnosis can cover larger susceptible population [8].

Diagnosis accuracy of COVID-19 and radiography based infection localization are critical for treatment planning and follow-up evaluations [9]. However, pressure of pandemic forces physicians to evaluate in limited time, which raises misdiagnosis rate implicitly [10]. As a result, accurate and robust classification methods are required. This is challenging as COVID-19 is a new type of disease which has low amount of data comparing with available datasets, such as image data published by [11] or [12]. In addition, the COVID-19 shares characteristic with other types of pneumonia, which requires the method focus on both global and local features [13]. Moreover, varied parameter settings causes imparities when collecting X-ray image from different devices.

Massive radiological data and rapid developing computational power give artificial intelligence (AI) a chance to assist clinical diagnosis. Recently, classification of COVID-19 from radiological images have been explored. Wang and Wong [14] present a COVID-Net operated on CXR images to classify COVID-19 from pneumonia and normal cases. COVID-19 cases

are extracted from online COVID-19 datasets published by [15] and [16]. Non-COVID-19 image includes 1591 pneumonia images and 1203 normal images released by National Institutes of Health Clinical Center [17]. The experimental results showed that classification method with residual projection-expansion-projection-extension (PEPX) design pattern achieves 93.3% accuracy, which is better than general deep models such as VGG-19 (83.0% Accuracy) and ResNet-50 (90.6% Accuracy). The authors illustrate the locations focused by their model to visualize its decision making process.

Ghoshal and Tucker [18] present a Bayesian CNN to make diagnosis through model uncertainty. It is trained on 68 COVID-19 cases from [15] and Non-COVID-19 cases from Kaggle's Chest X-ray Images (Pneumonia) [19], which improve the classification accuracy of a standard ResNet50V2 model from 86.0% to 89.8%. The authors further discuss the effectiveness of uncertainty-aware classification by decision visualization.

Zhang *et al.* [20] design a screening method based on ResNet to detect COVID-19 and find abnormalities from CXR images. Images are evaluated by an abnormity detecting module producing reference score to optimize classification loss. The model is trained on 70 COVID-19 images and 1008 non-COVID-19 images, which reaches 96.0%, 70.7%, 95.2% in Sensitivity, Specificity and AUC respectively.

Generally, current studies operated on CXR images mostly depends on online datasets with limited COVID-19 cases. Insufficient data can hardly evaluate the robustness of the models and restricted their generalizability. Models trained on extremely imbalanced dataset also lead to long-tail distribution problems. Although plenty of works have discussed diagnosing COVID-19 by AI, few works address the problem of imbalanced data and limited size of dataset because of several issues: 1) Models trained by imbalanced data tend to classify all the targets to the dominant class which has overwhelmingly more labels than other classes. 2) Unique labels on X-ray image, such as L/R position labels, easily attract model's attention then mislead the predictions. 3) COVID-19 cases share features with non-COVID cases, which requires a sensitive and robust model to do classification.

These challenges inspired us to treat pneumonia classification as a Fine-Grained Visual Classification (FGVC) problem, which aims at classifying sub-level categories under a basic-level category. FGVC cases are similar apart from some minor differences and also has the problem of lacking training data. Classic Convolutional Neural Networks (CNNs), including VGG [21], ResNet [22] and Inception [23], has difficulties handling this problem. We propose a novel Multiscale Attention Guided deep network with Soft Distance regularization (*MAG-SD*) for COVID-19 CXR image classification. To balance the quantity of different data, a weakly-supervised method is presented, which requires a few labeled data to do effective augmentations. Multiscale strategy is applied to attention generator, producing detailed scalar matrix for prediction. Our classification model is motivated from the fact that clinical diagnosis of COVID-19 follows a procedure which firstly evaluates the regional appearance, then makes diagnosis exclusively. Thus, we propose a multiscale attention module which estimates both shallow and deep layers. Comparing with using feature maps from only highest level features, the utilization of lower features could increase its ability of finding fine-grained features. Moreover, a soft distance regularization method is integrated to refine classification result by adaptively adjusting classification loss. In a nutshell, contribution of this paper is threefold:

1) We design a novel deep network, *MA-Net*, to treat COVID-19 diagnosis as a FGVC problem. Multiscale attention is introduced to assess attention maps on multi level features. Composed attention maps are used as guidance for training steps. Attention pooling is proposed to utilize attention maps for classification.

2) We address data shortage by proposing attention guided data augmentation and multi-shot training phase. It includes attention mix-up, attention patching and attention dimming that could enhance and search local feature then generating data. Models are trained on imbalanced COVID-19 datasets and achieve the state-of-the-art.

3) Without introducing other modules or parameters, we formulate a new regularization term utilizing soft distance between predictions, which works as a constraint to limit classifiers from producing contradicted output for one target.

This paper is organized as follows. In Section II, we introduce insightful works which have high relevance with our contribution. Section III presents the proposed method. In Section IV, database and experimental setup are reported in detail, then results are presented and discussed individually. The last section concludes this study and highlights the future work.

II. Related Works

Related works are introduced in this section, including X-ray appearance for typical pneumonia, fine-grained visual classification, attention mechanism for CNNs and multiscale feature fusion utilized in computer vision.

A. Pneumonia X-Ray Imagery

Chest X-ray is a widely used imaging modality providing high-resolution pictures to visualize the pathological changes of thoracic diseases. Diagnosis could be made according to the visual patterns demonstrated on CXR images [19]. Clinical research from Katz and Leung [24] demonstrated that typical image pattern for bacterial pneumonia includes opacity of single lobe and pleural effusion. Viral pneumonia also has radiological appearance such as pulmonary edema, small area of effusions, consolidation or lobe mass. Reports from [25], [26], demonstrated that the most common pattern on CXR in COVID-19 was consolidation or ground-glass opacity. It is notable that COVID-19 shares some visual feature with viral pneumonia while viral and bacterial pneumonia can hardly be differentiated because of similar spatial appearance.

B. Fine-Grained Visual Classification

Mass application of CNNs revealed its advantage in solving large scale image classification problem [27] and illuminated a promising way to settle FGVC tasks by using CNN models to explore inconspicuous local features. Some models relied on

local annotations to train part-based detectors, localizing certain parts before prediction [28], [29]. However, local feature annotation requires expensive human labor, limiting its reproducibility in reality. Recently, approaches only require global labels also emerged whose motivation was to first localize the corresponding parts and then compare their local features [30]. Fu *et al.* [31] introduced WS-DAN, which was a weakly supervised deep network handling FGVC by posing attention to enhance local feature and guide augmentation. FGVC was also a common problem in medical image because of spatial similarity between infections. Qin *et al.* [32] proposed a fine-grained classification CNN for different types of lung cancer in PET and CT Images.

C. Attention for CNNs

For visual task, attention usually indicates a scalar matrix representing the relative importance and inner relevance of local feature [33]. This nonuniform representation was produced by special designed modules [34]. Works reported that applying attention on classification oriented CNN could provide an intuitional way to localize target object, helping to identify visual properties through local representation. An attention guided method demonstrated by Gondal *et al.* [35] reported that attention mechanism is helpful in Diabetic Retinopathy (DR) localization and recognition. Zhang *et al.* [36] regulated the attention of deep model by training self-attention blocks for skin lesion classification and surpassed the baselines. Generally, attention mechanism guide the models to analyze global and local features simultaneously then generate believable classification results.

D. Multiscale Feature Fusion

Extracting hybrid feature maps from multi-resolution input image is a common strategy in computer vision since the the era of hand-engineered features. CNNs have an inherent multiscale feature in pyramidal shape, which is advantageous in producing semantically strong representations if effective feature fusion is operated. Models such as U-Net [37] and V-Net [38] exploited skip connections to associate feature maps across resolutions. FPN [39] leveraged the prediction of multiscale hierarchy by generating multiple prediction. For CXR image, Huang *et al.* [40] presented weight concatenation method to cooperate global and local feature. Thriving of spatial attention gave inspiration to extract attention from multi-resolution feature map. Sedai *et al.* [41] proposed A-CNN for chest pathologies localization, which utilized multiscale attention by calculating convex combination from weighted average of the feature maps.

III. Method

In this episode, we propose our approach that explore multiscale fine-grain feature adaptively. We first produce an overview for our *MAG-SD*. Then *MA-Net* is presented in terms of network architecture with attention modules. A weakly supervised data augmentation module, *Attention Guided Augmentation*, is introduced to address the shortage of COVID-19 cases. At last, *Soft Distance Regularization* is proposed to erase noise imported by augmentations.

A. Overview

COVID-19 CXR images are less distinctive comparing with other pneumonia cases, which requires a model to extract features for fine-grained feature of input image. WS-DAN [31], which is competitive in fine-grained image classification has been adopted for this topic. The architecture includes a feature extractor (*i.e.* ResNet50), an attention generator operated on feature map and an augmentation generator producing local-enhanced and noise-blended image. An overview of our *MAG-SD* is shown in Fig. 1. In primary training route, preprocessed CXR image I_0' is fed into *MA-Net* for prediction vector P and attention map A. *Attention Guided Augmentation* is operated on I_0', using A to produce augmented data I_1, I_2, I_3. In Auxiliary training routes, I_1, I_2, I_3 are pushed into *MA-Net* for prediction vectors p_1, p_2, p_3. All the vectors (*i.e.* P, p_1, p_2, p_3) are utilized by *Soft Distance Regularization* for a proper loss.

B. Multiscale Attention Guided Network (MA-Net)

1) Network Architecture: Fig. 2 presents a demonstration of our proposed *MA-Net*. As observed, a CNN based encoder is operated on augmented images. Encoder utilizes ResNet50 as backbone, extracting size-different feature maps f_1, f_2, f_3 from image I. *Multiscale attention generator* is used to extract attention map $a1, a2, a3$ and estimate scale-wised interests. Attention maps are resized for a single output A from features. Then, the output of encoder f_3 and attention map A are assessed by *Attention Pooling* to generate prediction vector P.

2) Multiscale Attention Generator: Attention mechanism has been used in natural image topics to guide feedforward process [42], [43]. Recently, tentative efforts have been made on deep models such as image classification [44], person perception [45] and sequential decision tasks [46]. Most of the attention models aim at gathering top level information to decide where to attend for the next learning steps. The proposed attention generating model is operated on multiscale feature maps, aiming at extracting attention from different scale. Layers before downsampling are selected as feature map in order to squeeze information out of single resolution feature. For ResNet50 we used, feature maps with $512 * 28 * 28, 1024 * 14 * 14, 2048 * 7 * 7$ sizes are chosen. The number of attention map is 32.

The architecture of multiscale attention generator is shown in Fig. 3. f_1, f_2 and f_3 are feature maps selected from feature extractor. Each of them are processed by $1 * 1$ convolution to generate corresponding attention. All the attention maps are downsampled to $7 * 7$ and connected residually. The effect of using different number of feature maps is discussed in experiments.

3) Attention Pooling: Attention pooling module mimics the structure proposed by [31], which associates attention output and feature map. Fig. 4 shows the pipeline of the pooling method. Feature map f_3 $(2048 * 7 * 7)$ is extracted from the output of CNN encoder. Multiscale attention map A presented by attention generator is $32 * 7 * 7$. Each attention map focuses on diverse location that may contain valuable fine-grained feature. Attention biased features (*i.e. part feature map* (PF)) are presented by multiplying all the attention maps A, each by each, with

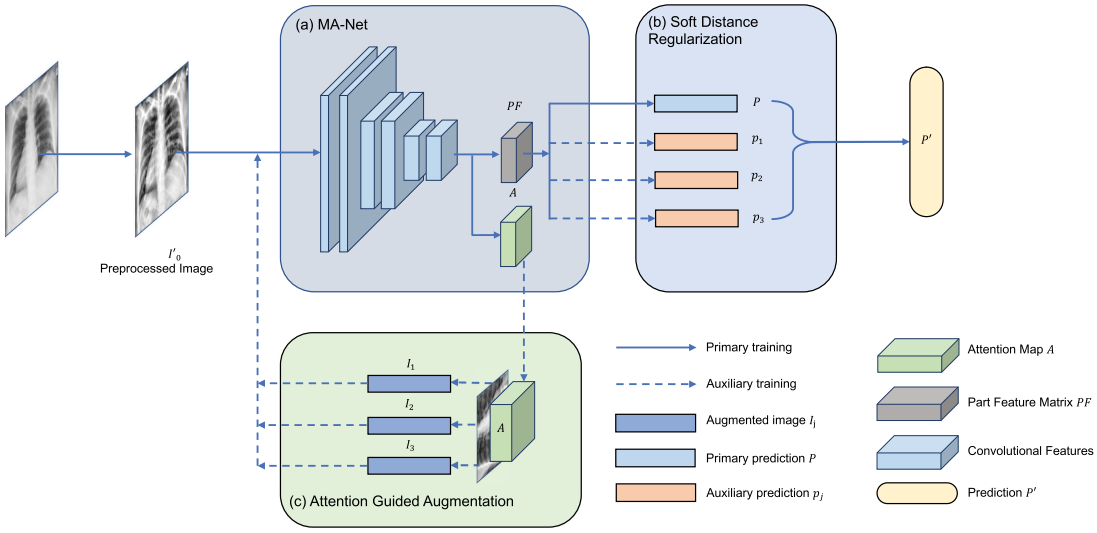

Fig. 1. The architecture of *MAG-SD*. The key components are illustrated in colour-wised blocks. (a): *MA-Net*, which is a CNN model (*e.g. ResNet50*) extracting prediction vectors P, p_1, p_2, p_3 and attention map A; (b): *Soft Distance Regularization* using P, p_1, p_2, p_3 to calculate overall loss; (c): *Attention Guided Augmentation*, which augments preprocessed data I'_0 according to A.

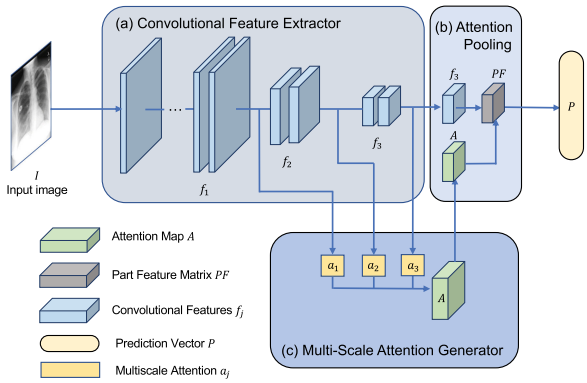

Fig. 2. *MA-Net* illustrated in colour-wised blocks. (a): Convolutional Feature Extractor, which is a pretrained CNN model (e.g. ResNet50) extracting features f_1, f_2, f_3; (b): Attention Pooling (demonstrated in Fig. 4) takes f_3 and attention map A for prediction vector P); (c): Multiscale Attention Generator (demonstrated in Fig. 3) uses f_1, f_2, f_3 to produce A as output.

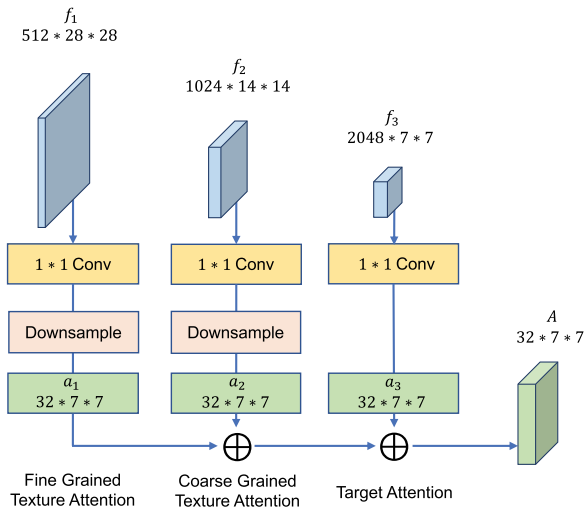

Fig. 3. Demonstration of multiscale attention generator. f_1, f_2, f_3 are three scales of feature maps. The model choose $1, 2$ or 3 feature maps for attention. Attention map is generated by operating $1 * 1$ convolutional layer on each feature map then downsample it to $7 * 7$. Global attention map A is produced by operating residual connection between resized feature maps. $\oplus$ represents residual connection.

feature map. There are 32 PF s which size equals $2048 * 7 * 7$. Global average pooling (GAP) is operated to shrink each PF to $2048 * 1 * 1$ in order to describe the activation intensity of attention on feature map. Feature matrix M is produced by concatenating GAP results, producing a vector of $65\,536 * 1 * 1$. Eq. (1) describes the calculation of PF.

$$PF_j = A_j \odot f_3 \quad (j = 1, 2, \ldots, N) \tag{1}$$

where $\odot$ stands for multiplication of elements between two tensors. f_3 is feature map extracted by CNN. N represents the number of attention maps, which is 32 in our work.

PF_j has to go through a downsampling method such as GAP to get description with compressed size, which is $2048 * 1 * 1$. Feature matrix M is represented by concatenating all condensed PF_j presented in Fig. 4.

C. Attention Guided Augmentation

As mentioned above, attention mechanism emphasizes local feature which affects the classification result. Following the idea, the performance of classification network could be enhanced if attention guided training cases are considered. Weakly supervised methods demonstrated in Fig. 5 present effective augmentations for original image. One normalized attention map (A^*) is randomly chosen for each instance to do individual augmentation.

1) Attention Mixup: Mixup is an augmentation strategy which generates data by mixing overall image and regional feature

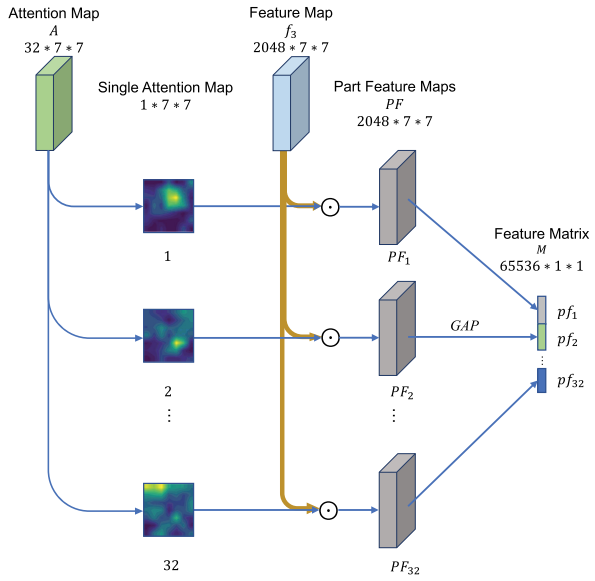

Fig. 4. Attention pooling architecture proposed for feature selecting. Feature map f_3 is extracted from input image and A is generated by the module displayed in Fig. 3. Each individual attention map selected from A multiplies ($\odot$) with f_3 to produce the features with attention bias, known as part feature Maps (PF). After global average pooling (GAP) process, feature matrix M is produced.

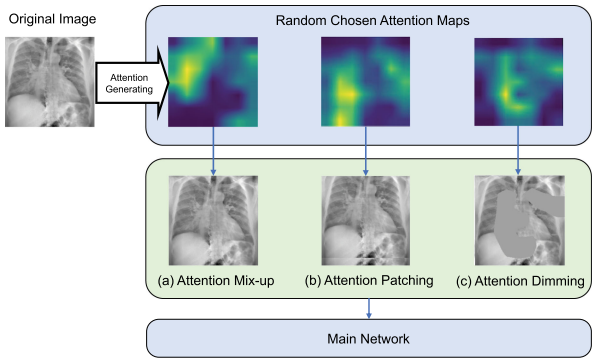

Fig. 5. Demonstration of Attention Guided Augmentation. Multiple attention maps are generated by attention generator, which concentrate on different part of original image. One attention map is chosen randomly for each augmentation method, including: (a) Attention mix-up, (b) Attention Patching and (c) Attention Dimming.

together. As we have a attention map A_j^*, a detailed region D_j could be extracted by doing threshold.

$$D(l,m) = \begin{cases} 1, & \text{if } A^*(l,m) > \theta_m \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

For elements in $A_j^*(l,m)$, Eq. (2) sets $D_j(l,m)$ to 1 if it is greater than threshold $\theta_m \in [0,1]$. If not, it will be set to 0. A bounding box surrounding the extracted region is proposed from the raw region. Region coved by the box is enlarged to the same size as input image then merged together with original input I_0 to get augmented input I_1, which is defined in Eq. (3).

$$I_1(p,q) = \gamma I_0(p,q) + (1-\gamma)B(p,q) \tag{3}$$

where γ is a parameter range in [0,1] and B stands for the enlarged bounding box. Model could see target precisely by learning local and global feature together.

2) Attention Patching: Encoder could be sensitive to limited part of reception field as valuable spatial features usually distribute in similar position. To encourage the encoder to explore feature from varied part, attention patching is proposed. D mentioned in *1)* is patched onto the original image I_0 to propose patched data I_2, which is demonstrated in Fig. 5. Attention patching enlarges the model's interest region by duplicating the interested area, promoting model to global evaluate its input.

3) Attention Dimming: When training attention generating module for feature map, multiple attention maps may be sensitive to similar region. A responsible fine-grain classification model have to focus on different local features of one target. Attention dimming is proposed to stimulate the attention model searching the whole reception field for valuable information. We obtain a Dimming Mask (DM) from A^*, applying threshold $\theta_d \in [0,1]$, as represented in Eq. (4).

$$DM(l,m) = \begin{cases} 00.1, & \text{if } A^*(l,m) > \theta_d \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

Augmented image I_3 is generated by applying the mask onto the input, which is illustrated in Fig. 5(c).

D. Soft Distance Regularization

Disturbances are introduced into the original image by using augmentation. (*e.g. infection area reduced by attention dimming*). To address this problem, we formulate the uncertainty of predictions via the distance between prediction vectors. Intuitively, the distance d could be modeled as Eq. (5).

$$d(x) = |P(I) - p(x)| \tag{5}$$

where x denotes the augmented image, $P(I)$, $p(x)$ represent primary and auxiliary prediction vector respectively. However, the distance between $P(I)$ and $p(x)$ is unstable before the the model well-fitted. Ground truth labels are referenced to stabilize gradients. In Algorithm 1, $P(I)$ is replaced by soft label $P'(I)$, filtering out low confidence inferences. Soft distance $d'(x)$ can be represented in Eq. (6). The value of θ in Algorithm. 1 is 0.7.

$$d'(x) = |P'(I) - p(x)| \tag{6}$$

At last, overall loss is modeled by a combination of cross entropy loss and average soft distance, which is demonstrated in Eq. (7).

$$L_{reg} = L_{ce}^{prim} + \bar{d'} \tag{7}$$

where L_{ce}^{prim} operates between labels and primary prediction. If two vectors have different prediction for one target, L_{reg} will generate a large value, which reflects the uncertainty of the model on one target. It is also notable that L_{reg} punishes soft distance $\bar{d'}$, leading the model to generate similar predictions.

IV. Experiments

In this section, extensive experiments were conducted to comprehensively assess *MAG-SD*. Models were trained on datasets

TABLE I
DATASETS DETAILS

Dataset	Class	Value	Total
Dataset A	COVID-19	90	258
	Non-COVID-19 pneumonia	168	
Dataset B	COVID-19	462	3631
	Non-COVID-19 pneumonia	1567	
	Healthy	1602	
Dataset C	COVID-19	462	6329
	Viral pneumonia	1449	
	Bacterial pneumonia	2816	
	Healthy	1602	
Localization	COVID-19	13	131
	Non-COVID-19 pneumonia	118	

Algorithm 1: Soft Distance Regularization.

Input: $P(I)$: Primary prediction vector
$p(x)$: Auxiliary prediction vector
$G_{lbl}(I)$: Ground truth labels
θ: Confidence threshold

Output: L_{reg}: soft distance regularization term

Cross entropy loss L_{ce}^{prim} is calculated between $P(I)$ and $G_{lbl}(I)$;

$P(I)$, $p(x)$ are fed into *softmax* to extract confidence score over all classes, which are $P^c(I)$, $p^c(x)$;

if $P^c(I) > \theta$ **then**

  Let $P'(I) = P(I)$

else

  Let $P'(I) = G_{lbl}(I)$

end

Predict variance is represented by soft distance between $P'(I)$ and $p(x)$:

$$d'(x) = |P'(I) - p(x)|$$

Overall loss is combined by L_{ce}^{prim} and mean predicting variance:

$$L_{reg} = L_{ce}^{prim} + \bar{d}'$$

return L_{reg}

with different types of pneumonia and the performance of each proposed method was evaluated. Then the models were compared between other baseline methods using several metrics.

A. Dataset and Experimental Settings

The proposed model was trained and tested on several datasets to evaluate its classification performance and ability of fine-grained pneumonia localization. Details of each dataset was shown in Table I. *Dataset A* was a mutated dataset with 90 COVID-19 from [15] and 168 other pneumonia cases from [17], which directly assessed model's fine-grained classification ability. *Dataset B* was selected from [47] and [17], aiming at evaluating the model's performance on larger scale. *Dataset C* was the largest dataset we operated on, which included COVID-19 detection and fine-grained pneumonia classification. Quality of pneumonia localization was evaluated by *Localization* dataset, which had 13 COVID-19 cases with pixel-wise masks from [48] and 118 non-COVID pneumonia cases with bounding boxes annotations from [17]. In experiments, classic ResNet50 has been adopted as feature extractor. Its *layer4* output was chosen as feature map. Attention was extracted from the output of *layer2, layer3* and *layer4* to ensure multiscale attention. Size of the attention maps were $28 * 28$, $14 * 14$ and $7 * 7$. Both training and testing sets were divided roughly in the same class proportions. 5-fold cross validation was applied to get reliable results.

TABLE II
AUGMENTATIONS USED AND FACTOR SETTING

Augmentations	Abstract
Brightness adjustment	Random chosen brightness factor from $[0.5, 1.0]$
Contrast adjustment	Random chosen contrast factor from $[0.7, 1.0]$
Resized cropping	Random cropping then resize to $224 * 224$
Rotation	Random rotation from $[0, 120]$
Horizontal flipping	-
Vertical flipping	-

Models were implemented using Pytorch and trained on two NVIDIA RTX 2080TI GPUs. The optimizer was Stochastic Gradient Descent (SGD) with the momentum of 0.9. For each training, 100 training epochs were deployed, with 10^{-6} weight decay, 32 cases per minibatch and 10^{-3} learning rate at beginning. Images were resized to $224 * 224$ when training and testing.

B. Pre-Processing and Data Augmentation

X-ray images have different appearance according to varied imaging equipment configurations, resolving that the same tissue can be radiologically different. To ensure the intensity distribution of one tissue is similar over the dataset, Z-score normalization was employed when training and testing. Large contrast distribution also introduced extra noise to the dataset, impacting the performance of trained model. Contrast limited adaptive histogram equalization (CLAHE) was proposed to enhance contrast between tissues and restrain noise signal [49].

In image classification, data augmentation has been proved as an effective method to improve robustness and evaluate performance [50]. Augmented data provides more varieties for classification target and remitting the impact of overfitting. Random number of transformations were chosen from a sequence of linear transformation for each training sequence. The list is shown in Table II.

C. Evaluation Metrics

Experiments were evaluated by several metrics. For Classification, Accuracy (ACC), Sensitivity (SEN), Specificity (SPC) and F1 score were employed. For multi-class datasets, mean value between classes were calculated to represent the final performance score of each model. Plots of receiver operating characteristic (ROC) curve and area under the curve (AUC) value were used to compare model functionality. Localization quality was quantified by intersection over union (IOU) which has been widely used in target detect and semantic segmentation task [41]. Accuracy describes the proportion of correctly classified targets,

TABLE III
EVALUATION OF MODELS

Model	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
VGG16 [21]	92.88±1.35	91.51±1.17	92.77±1.89	92.08±1.44	90.68±1.64	91.05±2.12	94.90±0.82	89.44±1.98	80.23±1.44	77.97±2.02	92.98±0.43	78.82±1.28
ResNet18 [22]	90.94±1.39	91.87±1.55	88.56±1.79	89.83±1.46	92.06±1.14	92.03±1.13	95.62±1.03	91.12±1.34	82.33±1.35	82.61±1.20	93.57±1.22	81.30±1.23
ResNet50 [22]	92.94±1.19	91.16±1.14	94.25±1.53	92.31±1.39	92.56±0.76	92.07±1.68	95.85±0.47	91.74±0.93	82.94±1.02	84.01±1.16	93.61±0.37	82.64±1.39
InceptionV3 [23]	94.13±1.13	92.94±1.02	94.49±1.11	93.62±1.12	93.06±1.19	91.73±1.13	96.23±0.77	92.42±1.51	84.20±1.19	85.19±1.35	94.03±0.99	84.69±1.27
[51](ResNet)	93.88±1.18	93.01±1.32	93.63±1.36	93.30±1.55	93.41±1.14	93.71±1.72	96.26±0.62	93.12±1.49	83.93±1.11	86.40±0.99	93.85±0.42	84.39±1.01
[51](InceptionV3)	94.56±1.75	92.19±1.62	94.91±1.47	93.40±1.47	93.45±1.21	93.00±1.24	96.40±0.50	93.04±1.02	84.93±1.67	85.90±1.01	94.36±0.71	84.92±1.70
COVID-Net [14]	93.25±1.70	91.62±1.88	93.70±1.80	92.51±1.89	88.94±1.28	89.95±2.41	93.75±0.58	87.98±1.56	78.71±1.76	79.26±0.94	92.38±0.47	78.34±1.40
BCNN [30]	96.00±1.52	96.43±1.78	94.52±1.60	95.39±1.42	94.41±1.37	95.26±1.23	96.71±0.81	**96.71±1.60**	84.36±1.84	84.47±0.75	94.15±0.49	84.47±0.85
BCNN(Attention(Ours))	96.43±1.45	96.31±1.55	**96.16±1.74**	96.23±1.30	95.11±1.55	**96.61±2.00**	97.26±0.72	94.12±1.92	85.04±2.36	86.32±1.55	94.33±0.86	84.41±1.52
FPN [39]	94.88±1.61	95.11±1.75	94.30±1.22	94.65±1.72	93.27±1.20	94.05±0.82	96.20±0.78	92.86±1.11	82.17±1.89	83.58±1.42	93.32±0.50	81.85±1.16
U-Net [37]	95.76±1.07	94.92±1.10	95.91±1.62	95.38±1.13	93.00±1.55	92.59±1.39	96.08±0.82	92.33±1.84	84.01±1.36	84.68±1.42	94.14±0.56	83.81±1.24
MAG-SD(0AUG)	94.31±1.28	91.70±1.33	95.89±1.94	93.37±1.38	93.25±0.90	92.01±0.93	96.29±0.45	92.06±1.17	84.13±0.99	85.08±1.27	94.11±0.43	84.30±1.43
MAG-SD(Proposed)	**96.94±1.10**	**97.83±1.53**	94.93±1.26	**96.23±1.02**	**95.85±1.27**	95.74±1.20	**97.73±0.45**	95.54±1.59	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**

TABLE IV
EVALUATION OF CLAHE

Preprocessing	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
w/o CLAHE	95.56±1.14	93.12±1.59	96.23±1.03	94.50±0.90	93.45±1.58	92.75±2.17	96.37±1.40	92.64±1.23	85.47±1.20	86.64±1.96	94.59±0.85	85.96±1.64
CLAHE	**96.94±1.10**	**97.83±1.53**	**94.93±1.26**	**96.23±1.02**	**95.85±1.27**	**95.74±1.20**	**97.73±0.45**	**95.54±1.59**	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**

which is expressed in Eq. (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP, TN, FP and FN stand for the number of true positive, true negative, false positive and false negative predictions. Sensitivity, also known as true positive rate (TPR), is useful to measure the proportion of true positive predictions over all positive targets, which is defined in Eq. (9).

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

Specificity, or true negative rate (TNR), is a ratio between the amount of true negative (TN) and false positive (FP) predictions, defined in Eq. (10).

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

F1 Score considers the performance from both precision and recall, defined in Eq. (11).

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

IoU represents a value calculated by dividing the overlap of prediction and ground truth by their union. It could be defined straightforward in Eq. (12), where A_o and A_u denote area of overlap and area of union respectively.

$$IoU = \frac{A_o}{A_u} \tag{12}$$

D. Components Validation and Discussion

The methods composed could be concluded into attention modules, attention guided data augmentation and soft distance regularization. Each component was studied by evaluating its improvement in classification performance, which has been quantified by metrics mentioned above. Performance gain was obtained by the following method: the proposed model was first trained on specific dataset with metrics, then, single component was changed or removed and reevaluate on the same dataset. For all the tested models, Mean value and standard deviation of ACC, SEN, SPC, F1 were recorded. Components validations were reported in Tables IV, V, VI, VII and VIII. Inter-model comparisons could be found in Table III and Fig. 6. Regions interested the attention module were presented in Fig. 7. Parameters in all the experiments were maintained unchanged as possible for condition control. The model were trained on the same size of training set then evaluated on the same size of testing set.

1) Architecture Comparing: Advantages of architecture design has been deeply explored. It has been performed by evaluating classic coarse-grained deep neural networks (*i.e. VGG16, ResNet18, ResNet50* and *InceptionV3*), COVID-19 oriented architectures (*i.e. [51] (ResNet), [51] (InceptionV3), COVID-Net-Large*), high performance fine-grained image classification structure (*i.e. BCNN, BCNN(Attention)*) and multiscale feature fusion models (*i.e. FPN, U-Net*). Statistics analysis between these deep structures helped to explain our advantages in fine-grained feature extraction. It can be observed in Table III and Fig. 6 that proposed model had noticeably better performance over others. For our model, accuracy on dataset A, B and C reached 96.94% ± 1.10%, 95.85% ± 1.27%, 87.12% ± 1.55% respectively and performance assessed by AUC are 99.94%, 98.72%, and 95.11%.

Comparing with classic models, our model was specialized for COVID-19 image classification and attention guided training phase had advantage in fine-grained visual classification task. Most of the other COVID-19 oriented models presented better performance than classic models, however, none of them applied attention mechanism or considered fine-grained features, which impacted their accuracy on large scale, multi-class dataset such as Dataset B and C. Comparison between FPN, U-Net and classic models demonstrated that FPN presented results over InceptionV3 in Dataset A and B. In Dataset C, U-Net had higher accuracy than FPN, which exceeded ResNet50. Results indicated that multiscale feature fusion models reached competitive results using relatively simple structures comparing with classic deep models, which left us a hint that multiscale attention might be a possible route to improve.

TABLE V
EVALUATION OF MULTISIZE ATTENTION MAPS

Attention Maps	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
7*7	95.31±1.10	96.70±1.49	91.09±1.04	93.48±1.07	94.75±1.54	94.93±1.44	96.98±1.06	94.11±1.57	85.44±1.47	86.56±0.96	94.56±0.36	86.36±0.92
7*7 + 14*14	**96.94±1.10**	**97.83±1.53**	**94.93±1.26**	**96.23±1.02**	**95.85±1.27**	**95.74±1.20**	**97.73±0.45**	**95.54±1.59**	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**
7*7 + 14*14 + 28*28	95.88±1.84	94.76±2.11	96.63±2.23	95.54±1.91	94.74±1.79	95.80±2.49	96.92±1.17	94.57±2.16	85.01±1.89	86.65±2.29	94.29±0.97	85.73±2.25

TABLE VI
COMPARISON OF POOLING METHODS

Pooling	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
GMP	94.31±2.21	90.98±2.15	95.85±1.84	92.97±2.02	94.47±1.44	93.38±1.47	96.94±0.81	93.80±1.60	84.79±1.53	86.14±2.13	94.36±0.48	85.15±1.55
GAP	95.06±1.14	95.59±1.11	93.37±1.61	94.35±1.24	94.91±1.64	95.01±1.76	97.16±0.91	94.39±1.68	85.09±1.64	87.25±2.02	94.36±0.83	85.60±1.81
Attention Pooling	**96.94±1.10**	**97.83±1.53**	**94.93±1.26**	**96.23±1.02**	**95.85±1.27**	**95.74±1.20**	**97.73±0.45**	**95.54±1.59**	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**

TABLE VII
CONTRIBUTION OF ATTENTION GUIDED AUGMENTATION

Augmentation	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
A^{M*}	94.69±1.38	96.06±1.68	92.82±1.50	94.10±1.08	93.56±1.51	92.79±1.60	96.57±0.68	92.65±1.43	85.47±1.40	85.48±1.44	94.64±0.54	85.10±1.13
$A^M + A^{D**}$	95.81±1.23	96.70±1.52	94.86±2.23	95.59±1.34	94.97±1.57	95.10±1.34	97.25±0.73	94.84±1.38	86.31±1.78	87.12±1.82	94.83±0.85	86.60±1.64
$A^M + A^D + A^{P***}$	**96.94±1.10**	**97.83±1.53**	**94.93±1.26**	**96.23±1.02**	**95.85±1.27**	**95.74±1.20**	**97.73±0.45**	**95.54±1.59**	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**

$*A^M$: Attention Mix-up; $**A^D$: Attention Dimming; $**A^P$: Attention Patching.

TABLE VIII
COMPARISON OF L2 AND SOFT DISTANCE REGULARIZATION

Loss	Dataset A				Dataset B				Dataset C			
	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)	ACC(%)	SEN(%)	SPC(%)	F1(%)
L2	95.93±0.61	96.36±0.85	95.48±0.71	95.83±0.79	94.62±0.86	93.75±0.91	97.05±0.68	93.75±0.93	83.99±0.94	85.33±1.34	94.11±0.69	84.80±0.83
Soft Distance	**96.94±1.10**	**97.83±1.53**	**94.93±1.26**	**96.23±1.02**	**95.85±1.27**	**95.74±1.20**	**97.73±0.45**	**95.54±1.59**	**87.12±1.55**	**87.20±1.64**	**95.20±0.64**	**86.98±1.27**

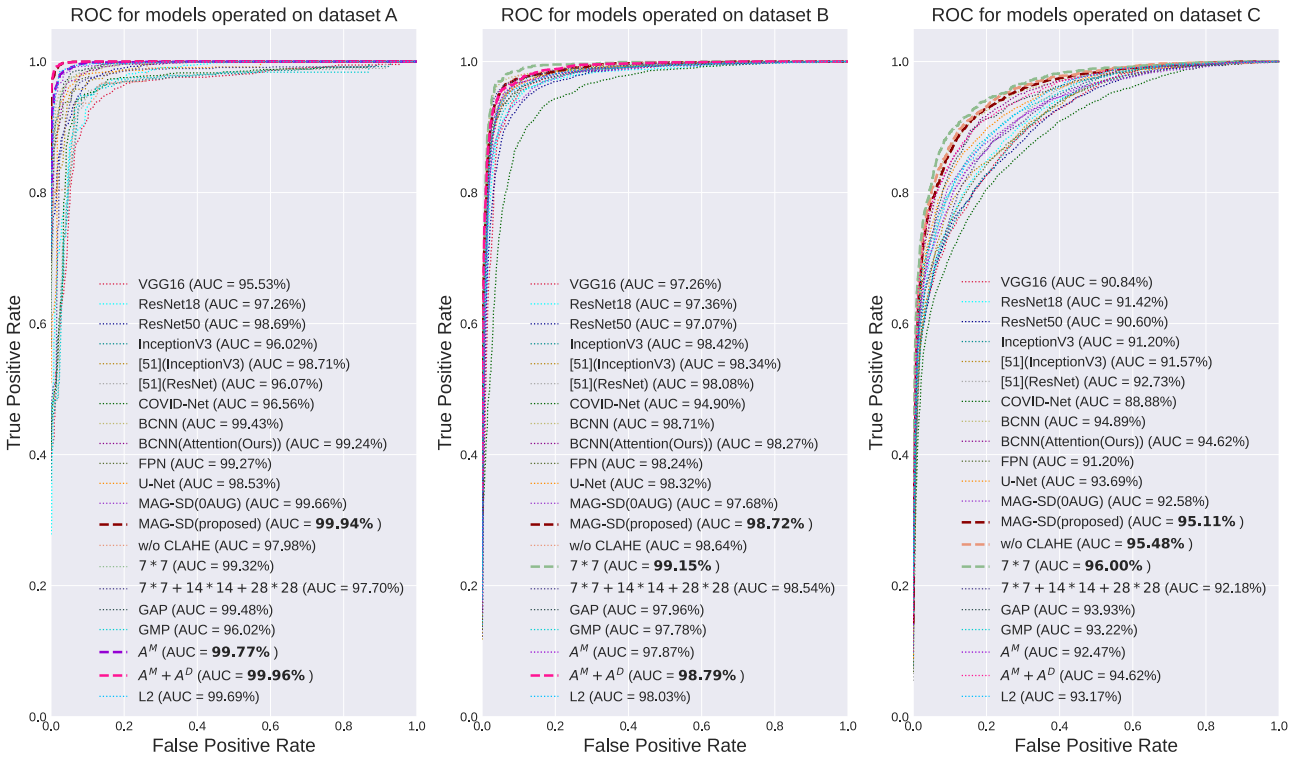

Fig. 6. Demonstration of ROC curves and AUC values. Three charts, from left to right, show the performance of all the trained models operating using datasets A,B and C respectively. Top-3 highest AUC values and their ROC curves are emphasized. Results demonstrate that comparing with baselines, results generated by our model has advantage in AUC value, which is over 0.5% in dataset A, B and C. Architecture differences of our proposed method also influence the performance over datasets. Generally, MAG-SD(proposed) is the most stable model which stays in top-3 in all the datasets, which is a method given consideration to both generalizability and robustness.

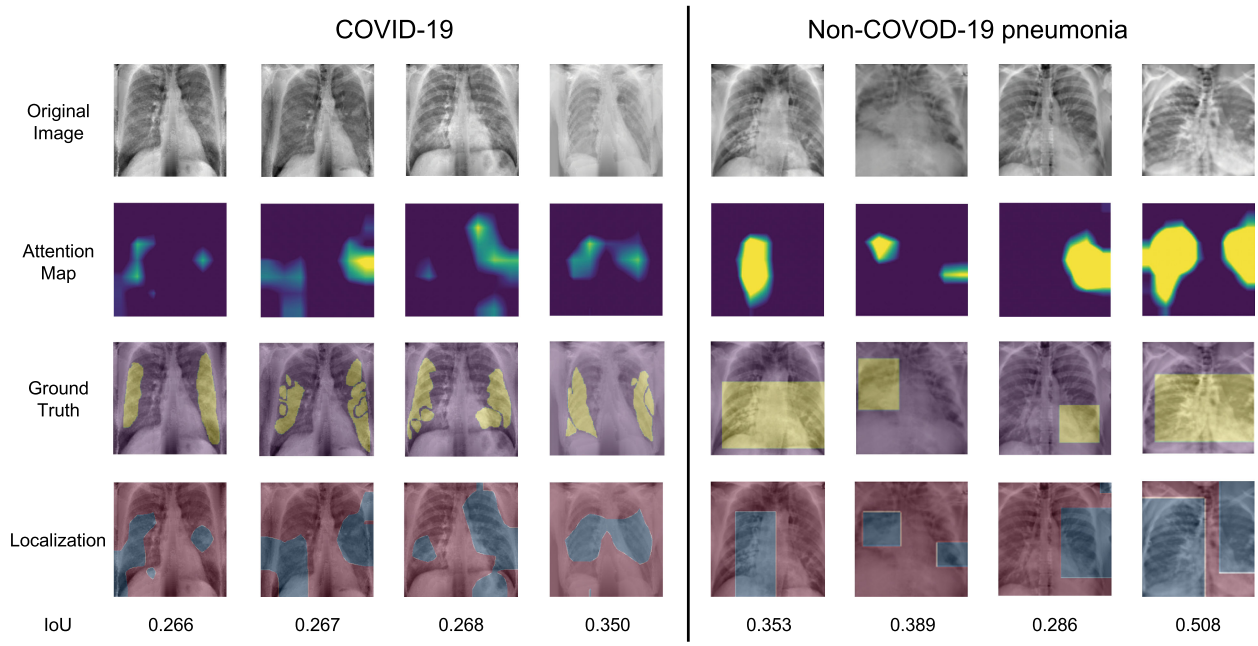

Fig. 7. Demonstration of pneumonia localization. Images were selected from *Localization* dataset. COVID-19 cases has pixel-wise mask while bounding boxes were provided for other pneumonia. IoU was calculated for each prediction. Localization result was provided by apply threshold onto the attention map of each case. Results illustrated that attention focus on different area when detecting various classes.

BCNN was a FGVC model with stable performance on multiple datasets. In order to evaluate the generalization ability of our attention module, multiscale attention and attention pooling were transported to BCNN to train BCNN(Attention). Statistically, BCNN reached $96.00\% \pm 1.52\%$, $94.41\% \pm 1.37\%$, $84.36\% \pm 1.84\%$ in Accuracy, which was competitive in all the evaluated models. Attention modules remarkably boosted the performance of BCNN, exceeded our proposed method in Dataset A (SPC) and Dataset B (SEN), which were $96.16\% \pm 1.74\%$, $96.61\% \pm 2.00\%$ respectively.

2) CLAHE Preprocessing: Images collected by different devices were probably distinct in contrast due to configuration variety. CLAHE was employed to relieve the noise brought by contrast distribution. Table IV showed the result that CLAHE obviously improved the performance of proposed model and raised over 1.5% Accuracy on average. Model trained without CLAHE had notable higher standard deviation value. Larger datasets such as Dataset B and C were reported to have more performance gain.

3) Multiscale Attention Generator and Attention Pooling: Normally, state-of-the-art coarse-grained CNN models suffer from similar global features when dealing FGVC. Under this circumstance, models have to depend on local features, which could be effectively localized by our multiscale attention module. Models trained with attention module (*i.e. MAG-SD(0AUG)*) and baseline model (*i.e. ResNet50*) were compared in Table III, and Fig. 6. Results revealed that proposed model surpasses the baseline on dataset A, B and C using most of the benchmarks. In dataset B (SPC), ResNet50 has slight advantage. Comparing with AUC, MAG-SD(0AUG) was 2% over ResNet50. Furthermore, two parts of attention module, attention generating and attention pooling has been investigated separately. Firstly, models were compiled to assess multiscale attention, with 1,2 or 3 size of attention maps considered. Results presented in Table V and Fig. 6 showed the model considering two feature maps achieved the best performance in all three datasets. Possible explanation was that the proposed attention module was too simple to locate valuable fine grained feature on low-level feature maps. Instead of importing meaningful location information, noise was brought into the proposed model. Secondly, we evaluated attention pooling module with models trained with other commonly used pooling methods such as global average pooling (GAP) or global max pooling (GMP). Results on pooling methods were presented in Table VI and Fig. 6, demonstrating that attention pooling surpassed GAP and GMP in all three datasets.

4) Attention Guided Augmentation: The generated attention maps emphasized local feature that interested the model, which could be used to effectively guide data augmentation in Fig. 5. Models trained with 0, 1, 2 or 3 augmentation were discussed in experiment. In the case of 1 augmentation, attention mixup was selected. 2 augmentations model included attention mixup and attention patching. The results were presented in Table VII and Fig. 6. The table reflects that model with all three augmentations had better performance, however, AUC value showed that in dataset A and B, two augmentations was advantageous. The proposed augmentations emphasized data according to attention map, minimizing negative effect caused by random augmentations.

5) Soft Distance Regularization: Soft distance regularization was presented to relieve augmentation variance. Experiments have been composed to compare it with $L2$ distance regularization. Table VIII and Fig. 6 illustrated that it surpassed $L2$ in mean value, but, inferior in standard deviation. Constraint between auxiliary vector and primary vector screen the false prediction

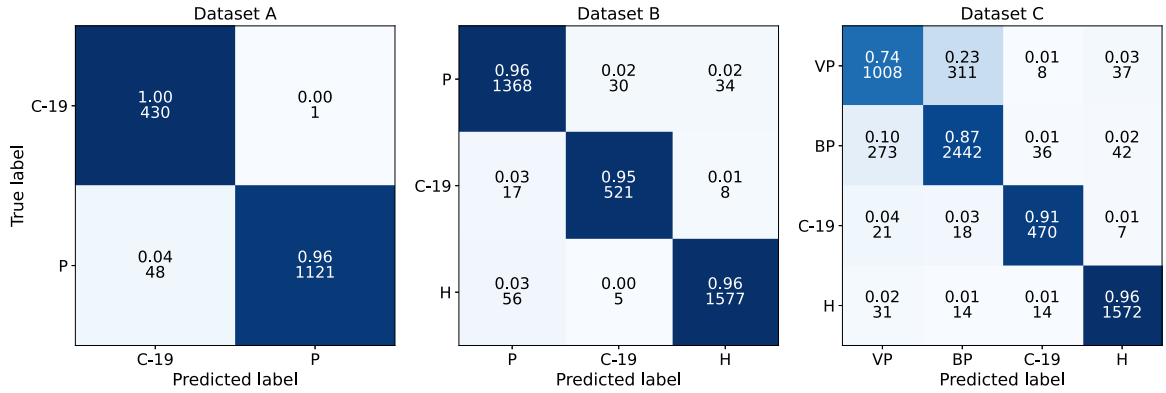

Fig. 8. Three charts of confusion matrices generated by proposed MAG-SD, demonstrating the distribution of predictions. The color of confusion matrices depend on the normalized values of predictions for a better visualization, which are placed at the top of each grid. The number of predictions are placed below the normalized values. Symbols used in figure are denoted as: P: Non-COVID19 Pneumonia, VP: Viral Pneumonia, BP: Bacterial Pneumonia, C-19: COVID-19, and H: Healthy.

introduced by attention guided augmentations. Regularization was calculated between ground truth and auxiliary vector when primary vector cannot provide reliable prediction, keeping the final result away from local minima. $L2$ compared all predictions directly with ground truth, which has higher stability and sidestepped the disturbances introduced by primary prediction.

6) Attention Based Infection Localization: Technically, attention improved the models by roughly localize the part with high activation intensity. This characteristic of attention inspired us to try *MAG-SD* on localization topics. The models were trained on the *Dataset B* we proposed, then test on *Localization* dataset demonstrated in Fig. 7. It included COVID-19 cases with pixel-wise segmentation and non-COVID-19 cases with bounding box for pneumonia infection. Attention maps A were upsampled from $7 * 7$ to $224 * 224$. Localization masks for COVID-19 cases were extracted by applying threshold to the attention maps. Bounding boxes for other pneumonia were produced by simply enclosing the localization masks with rectangles. IoU was calculated to evaluate the quality of localization. Image showed that the attention module we proposed could roughly indicate the position of different type of pneumonia with over 0.25 IoU score. Attention map emphasized the influential part from the input image effectively.

E. Distribution Analysis

As we imported multiple fine-grained classes into this topic, it was necessary to report the distribution of our prediction result, which has been shown using confusion matrix in Fig. 8. MAG-SD has been selected to generate the charts to represent the classification result of deep learning models. It could be inferred that the model was suitable for searching definitive features from cases showed in dataset A and B as most of the cases were located on the diagonal line of matrices. In dataset C, classification result between viral pneumonia and bacterial pneumonia was significantly inferior than others, which impacted the global performance of the classification model. These results proved the arguments reviewed in Section II-A, indicated that the CXR visual appearance between viral and bacterial pneumonia was insufficient to make accurate diagnosis.

V. Conclusion

We have presented *MAG-SD* for automatic COVID-19 CXR image classification that reached the state-of-the-art on our dataset. The proposed novel method treated this topic as a fine-grained image classification task, utilizing local features efficiently under the guidance of attention mechanism. Attention maps were generated using multiscale features then used as a reference to data augmentation, helping the model to overcome the lack of COVID-19 cases. The proposed network learned to weight the predictions from both primary and auxiliary training pathways by calculating soft distances between vectors, gaining improvements by screening noise generated by augmentations.

Findings of our exploration were demonstrated and discussed in Section IV. The results indicated the great potential of applying advanced pattern recognition model to clinical diagnosis and epidemic screening. Trained on the clinical knowledge acquired by physicians, our model was capable to extract fine-grained spatial features for COVID-19. Attention was applied in both feature extraction and augmentation stage, which helped to localize pneumonia infection and accrete the data effectively as part of weakly supervised method. Attention module also shows its capability in different models. It could be interesting to design more auxiliary training strategies to guide the model to an optimal solution. Positive feedback on soft distance regularization proved that our method considered auxiliary predictions and eliminated label noise simultaneously, however, hard threshold may limit its adaptability in complicated data.

Although deep learning methods seem promising in clinical diagnosis and pandemic screening, lacking of prior knowledge is always the Achilles' Heel. Supervised learning method, such as *MAG-SD* we proposed, have to be trained on expensive labeled data. Newly occurred or rare diseases without available data may not be classified properly. Abnormal detecting and clustering model could be proposed as a guidance for supervised models to alleviate the limitations, which is part of our future work.

Acknowledgment

The authors would like to thank the institutes who generously open-sourced image database.

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. H. Organization *et al.*, "Coronavirus disease 2019 (COVID-19): Situation report, 72," Apr. 1, 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/331685

[3] I. Apostolopoulos, S. Aznaouridis, and M. Tzani, "Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases," *J. Med. Bio. Eng.*, vol. 40, pp. 462–469, 2020.

[4] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, 2020, Art. no. 200642.

[5] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[6] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021.

[7] Y. Wang, L. L. Sun, and Q. Jin, "Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2911947.

[8] T. Franquet, "Imaging of pneumonia: Trends and algorithms," *Eur. Respir. J.*, vol. 18, no. 1, pp. 196–208, 2001.

[9] A. Torres and C. Cillóniz, *Clinical Management of Bacterial Pneumonia*. Berlin, Germany: Springer, 2015.

[10] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Trop. Med. Int. Health*, vol. 25, no. 3, pp. 278–280, 2020.

[11] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest X-ray image dataset with multi-label annotated reports," *Med. image Anal.*, vol. 66, p. 101797, 2020.

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

[13] N. Chen *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study," *Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[14] L. Wang, Z. Q. Lin and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Rep.*, vol. 10, no. 1, 2020, Art. no. 19549.

[15] J. P. Cohen *et al.*, "Covid-19 image data collection: Prospective predictions are the future," 2020, *arXiv:2006.11988*.

[16] L. Wang, "Figure 1 covid-19 chest x-ray dataset initiative," Accessed: May 9, 2020. [Online]. Available: https://github.com/agchung/Figure1-COVID-chestxray-dataset

[17] 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.

[18] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," 2020, *arXiv:2003.10769*.

[19] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[20] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "COVID-19 screening on chest X-ray images using deep learning based anomaly detection," 2020, *arXiv:2003.12338*.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[24] D. S. Katz and A. N. Leung, "Radiology of pneumonia," *Clin. Chest Med.*, vol. 20, no. 3, pp. 549–562, 1999.

[25] A. J. Rodriguez-Morales *et al.*, "Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis," *Travel Med. Infect. Dis.*, vol. 34, 2020, Art. no. 101623.

[26] S. Rajaraman and S. Antani, "Weakly labeled data augmentation for deep learning: A study on COVID-19 detection in chest X-rays," *Diagnostics* (Basel, Switzerland), vol. 10, no. 6, 2020, Art. no. 358.

[27] X.-S. Wei, J. Wu, and Q. Cui, "Deep learning for fine-grained image analysis: A survey," 2019, *arXiv:1907.03069*.

[28] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[29] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, 2018.

[30] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[31] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4438–4446.

[32] R. Qin *et al.*, "Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism," *Complexity*, vol. 2020, pp. 1–12, 2020.

[33] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," 2018, *arXiv:1804.02391*.

[34] J. Wang *et al.*, "Prior-attention residual learning for more discriminative COVID-19 screening in CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2572–2583, Aug. 2020.

[35] W. M. Gondal, J. M. Kohler, R. Grzeszick, G. A. Fink, and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 2069–2073.

[36] 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.

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

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

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[40] Z. Huang *et al.*, "Fusion high-resolution network for diagnosing chestx-ray images," *Electronics*, vol. 9, no. 1, 2020, Art. no. 190.

[41] S. Sedai, D. Mahapatra, Z. Ge, R. Chakravorty, and R. Garnavi, "Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in X-ray images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*. 2018, pp. 267–275.

[42] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[43] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[44] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.

[45] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[46] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[47] S. Edwardsson, "COVID-19 X-ray dataset," Accessed: Sep. 23, 2020. [Online]. Available: https://github.com/v7labs/covid-19-xray-dataset

[48] M. d. l. *et al.*, "BIMCV COVID-19: A large annotated dataset of RX and CT images from COVID-19 patients," 2020, *arXiv:2006.01174*.

[49] E. D. Pisano *et al.*, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *J. Digit. Imag.*, vol. 11, no. 4, pp. 193–200, 1998.

[50] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, 2017.

[51] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," 2020, *arXiv:2003.10849*.