Coronavirus disease 2019 (COVID-19) is one of the most destructive pandemics after millennium, forcing the world to tackle a health crisis. Automated classification of lung infections from chest X-ray (CXR) images strengthened traditional healthcare strategy to handle COVID-19. However, classifying COVID-19 from pneumonia cases using CXR image is challenging because of shared spatial characteristics, high feature variation in infections and contrast diversity between cases. Moreover, massive data collection is impractical for a newly emerged disease, which limited the performance of common deep learning models. To address this challenging topic, 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 map from multiscale feature maps. To relieve the shortage of training data, attention guided augmentations along with a soft distance regularization are posed, which requires a few labeled data to generate 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 that it has its unique advantage in...
Wong [13] present a COVID-Net operated on CXR images to classify COVID-19 from pneumonia and normal cases. It improves the classification accuracy of a standard ResNet50V2 model from which raises misdiagnosis rate implicitly [9]. As a result, accurate and robust classification methods are required. This means X-ray diagnosis can cover larger susceptible population [7].

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 and declared as a pandemic by the World Health Organization (WHO) according to its high contagiousity spreading geographically and affecting countries worldwide [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 shows ground-glass opacity, airspace opacities and later consolidation, which mostly observed as bilateral, peripheral, and lower zone predominant distributions [6]. Comparing with CT imaging, CXR diagnosis provides a low-cost and time-saving diagnosis method. Besides, underdeveloped regions can hardly have sufficient CT scanners, making CT based COVID-19 screening impossible. Oppositely, X-rays are the most common diagnostic imaging equipment available even in rural regions, which means X-ray diagnosis can cover larger susceptible population [7].

For patients, diagnosis accuracy of COVID-19 and radiography based infection localization are critical for treatment planning and follow-up evaluations [3]. However, pressure of pandemic forces physicians to judge in limited time, which raises misdiagnosis rate implicitly [9]. As a result, accurate and robust classification methods are required. This is a challenging topic as COVID-19 is a new type of disease which has low amount of data comparing with other datasets, such as image data published by [10] or [11]. In addition, the COVID-19 shares characteristic with other types of pneumonia, which requires the method focus on both global and local features [12]. 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 [13] 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 [14] and [15]. Non-COVID-19 image includes 1591 pneumonia images and 1203 normal images released by National Institutes of Health Clinical Center [16]. The experimental results shown that classification method with residual projection-expansion-projection-extension (PEPX) design pattern achieves a detection accuracy of 0.933, which is better than general deep models such as VGG-19 (0.830 Accuracy) and ResNet-50 (0.906 Accuracy). The authors illustrates the locations focused by their model to visualize its decision making process.

Ghoshal and Tucker [17] presents a Bayesian Convolutional Neural Network to make diagnosis through model uncertainty. The model is trained on 68 COVID-19 cases from [14] and Non-COVID-19 cases from Kaggle’s Chest X-ray Images (Pneumonia) [18]. It 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. [19] design a screening method based on ResNet to detect COVID-19 and find abnormalities from CXR images. Images are evaluated by an anomaly 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 0.960, 0.707, 0.952 in Sensitivity, Specificity and AUC respectively.

Generally, current studies operated on CXR images collect images from the online COVID-19 datasets have 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 in 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, which is a challenging task in 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, may easily attract model’s attention then mislead the predictions. 3) COVID-19 cases share features with non-COVID cases, which require a model to evaluate inputs both globally and locally.

The issues concluded 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. Cases that FGVC focus on are very similar apart from some minor differences and also facing the lack of training data. Coarse-grained Convolutional Neural Networks (CNNs), including VGG [20], ResNet [21] and Inception [22], can hardly reach the state-of-the-

**Keywords** COVID-19 · X-ray Radiology · Multiscale Attention · Convolutional Neural Network

### 1 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 and declared as a pandemic by the World Health Organization (WHO) according to its high contagiousity spreading geographically and affecting countries worldwide [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 shows ground-glass opacity, airspace opacities and later consolidation, which mostly observed as bilateral, peripheral, and lower zone predominant distributions [6]. Comparing with CT imaging, CXR diagnosis provides a low-cost and time-saving diagnosis method. Besides, underdeveloped regions can hardly have sufficient CT scanners, making CT based COVID-19 screening impossible. Oppositely, X-rays are the most common diagnostic imaging equipment available even in rural regions, which means X-ray diagnosis can cover larger susceptible population [7].

For patients, diagnosis accuracy of COVID-19 and radiography based infection localization are critical for treatment planning and follow-up evaluations [3]. However, pressure of pandemic forces physicians to judge in limited time, which raises misdiagnosis rate implicitly [9]. As a result, accurate and robust classification methods are required. This is a challenging topic as COVID-19 is a new type of disease which has low amount of data comparing with other datasets, such as image data published by [10] or [11]. In addition, the COVID-19 shares characteristic with other types of pneumonia, which requires the method focus on both global and local features [12]. 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 [13] 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 [14] and [15]. Non-COVID-19 image includes 1591 pneumonia images and 1203 normal images released by National Institutes of Health Clinical Center [16]. The experimental results shown that classification method with residual projection-expansion-projection-extension (PEPX) design pattern achieves a detection accuracy of 0.933, which is better than general deep models such as VGG-19 (0.830 Accuracy) and ResNet-50 (0.906 Accuracy). The authors illustrates the locations focused by their model to visualize its decision making process.

Ghoshal and Tucker [17] presents a Bayesian Convolutional Neural Network to make diagnosis through model uncertainty. The model is trained on 68 COVID-19 cases from [14] and Non-COVID-19 cases from Kaggle’s Chest X-ray Images (Pneumonia) [18]. It 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. [19] design a screening method based on ResNet to detect COVID-19 and find abnormalities from CXR images. Images are evaluated by an anomaly 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 0.960, 0.707, 0.952 in Sensitivity, Specificity and AUC respectively.

Generally, current studies operated on CXR images collect images from the online COVID-19 datasets have 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 in 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, which is a challenging task in 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, may easily attract model’s attention then mislead the predictions. 3) COVID-19 cases share features with non-COVID cases, which require a model to evaluate inputs both globally and locally.

The issues concluded 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. Cases that FGVC focus on are very similar apart from some minor differences and also facing the lack of training data. Coarse-grained Convolutional Neural Networks (CNNs), including VGG [20], ResNet [21] and Inception [22], can hardly reach the state-of-the-
art. 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 introduced, which only needs few amount labeled data to do effective augmentations. Multiscale strategy is applied to attention generator to produce 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 shallower and deeper layers. Comparing with using feature maps from only highest convolution layers, the utilization of lower features may increase the accuracy in incidental findings. Moreover, our weakly-supervised system integrated a soft distance regularization method which refines 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, treating COVID-19 detection as a fine-grained image classification problem. Multiscale attention is introduced into the proposed model to assess attention maps on both high-level target feature and low-level texture feature. Composed attention maps are used as guidance for following 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, as COVID-19 is newly occurred and lack of data comparing with abundant database of known diseases. It includes attention mix-up, attention patching and attention dimming, which focuses on enhancing and searching local feature before generating data. Models are trained on an imbalanced COVID-19 dataset and achieve the state-of-the-art.

3) Without introducing other modules or parameters, we formulate prediction loss using soft distance between predictions. Specifically, a new regularization term, soft distance regularization, is proposed to work together with cross entropy loss. Soft distance regularization works as a constraint between predictions, forcing the classifiers to produce similar output for one target.

The paper is organized as follows. In Section 2, we introduce insightful works which have high relevance with our contribution. Section 3 presents the proposed method. In section 4, database and experimental setup are introduced in detail, then results are presented and discussed individually. The last section concludes this study and highlights the future work.

2 Related Works

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

2.1 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. Clinical research from Katz and Leung[23] demonstrates that typical image pattern for bacterial pneumonia includes opacification 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 [24] shows that the most common pattern on CXR in COVID-19 is 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.

2.2 Fine-Grained Visual Classification

Mass application of CNNs revealed its advantage on solving large scale image classification problem [25], which enlightened using CNN for FGVC tasks, forcing CNN models to explore inconspicuous local features. Some models rely on local annotations to train part-based detectors, localizing certain parts before prediction [26] [27]. However, local feature annotation requires expensive human labor, which limits its reproducibility in real-world application. In recent years, approaches only require labels also emerge whose motivation is to first localize the corresponding parts and then compare their local features [28]. Fu et al. [29] introduce WS-DAN, which is a weakly supervised deep network handling FGVC by posing attention to enhance local feature and guide augmentation. FGVC is also a common problem in medical image as the spatial similarity of infections. Qin et al. [30] propose a fine-grained classification CNN for different types of lung cancer in PET and CT Images.
2.3 Attention for CNNs

For visual task, attention usually indicates a scalar matrix representing the relative importance and inner relevance of local feature \cite{31}. This nonuniform representation is produced by special designed modules \cite{32}. Works have shown that generating the attention map for classification CNN provides a intuitive way to localize the target object, helping to identify visual properties through local representation. An attention guided method demonstrated by Gondal et al. \cite{33} report that attention mechanism is helpful in Diabetic Retinopathy (DR) localization and recognition. Zhang et al. \cite{34} regulate the attention of deep model by training self-attention blocks for skin lesion classification and surpass the models without attention. Generally, attention mechanism force models to analyze global and local feature simultaneously to generate believable classification with localization results.

2.4 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 calculating a feature hierarchy layer by layer has an inherent multiscale feature hierarchy in pyramidal shape. Multiscale feature has an advantage in producing semantically strong representations if effective feature fusion is operated. Explorations on multiscale feature fusion includes U-Net \cite{35} and V-Net \cite{36}, which exploit skip connections to associate feature maps across resolutions, FPN \cite{37} which leveraging the prediction of multiscale hierarchy by multiple prediction. For CXR image, Huang et al. \cite{38} present weight concatenation method to cooperate global and local feature. Thriving of spatial attention gives inspiration to extract attention from multi-resolution feature map. Sedai et al. \cite{39} propose A-CNN for chest pathologies localization, which utilize multiscale attention by calculating convex combination from weighted average of the feature maps.

3 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.

3.1 Overview

COVID-19 CXR images are less distinctive comparing with other pneumonia cases, which require a model to extract features for fine-grained feature of input image. We adopt WS-DAN \cite{29} which is competitive on fine-grained image classification topic. The architecture of WS-DAN includes a feature extractor which is ResNet50 in original paper, 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.

3.2 Multiscale Attention Guided Network (**MA-Net**)  

3.2.1 Network Architecture

Fig. 2 presents a demonstration of our proposed **MA-Net**. As observed, a CNN based encoder is operated on augmented images. The encoder utilizes ResNet50 as backbone, extracting size-different feature maps $f_1, f_2, f_3$ from image $I$. **multiscale attention generator** is added to extract attention map $a_1, a_2, a_3$, estimating texture interest and target interest. 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 $P$.

3.2.2 Multiscale Attention Generator

Attention mechanism has been used in natural image contests to guide feedforward process \cite{40}, \cite{41}. Recently, tentative efforts have been made on deep models such as image classification \cite{32}, person perception \cite{43} and sequential decision tasks \cite{44}. Most of the attention models aim at gathering top information, deciding where to attend for the next learning steps. The proposed attention generating model is operated on multiscale feature maps, aiming at extracting attention from both texture level and target level. The last layer before downsampling are selected as
Figure 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 $A$.

Figure 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 (shown in Fig. 4) takes $f_3$ and attention map $A$ for prediction vector $P$; (c): Multiscale Attention Generator (shown in Fig. 3) uses $f_1, f_2, f_3$ to produce $A$ as output.

feature map in order to exploit information of single resolution. For ResNet50 we used, feature maps with size of $512 \times 28 \times 28$, $1024 \times 14 \times 14$, $2048 \times 7 \times 7$ are chosen. The number of attention map to be generated is set to 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 a $1 \times 1$ convolutional layer to generate corresponding attention. All the attention maps are downsampled to $7 \times 7$ and connected residually. The effect of using different number of feature maps is discussed in experiments.

3.2.3 Attention Pooling

Attention pooling module mimic the structure proposed by [29], which associates attention output and feature map. Fig. 4 shows the function pipeline of the pooling method. Feature map $f_3$ ($2048 \times 7 \times 7$) is extracted from the output of CNN encoder. Multiscale attention map $A$ presented by attention generator have the size of $32 \times 7 \times 7$. Each attention map focuses on diverse location which 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 feature map. There are 32 PFs which size equals $2048 \times 7 \times 7$. Global average pooling (GAP) is operated to shrink each $PF$ to $2048 \times 1 \times 1$ to
Figure 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 $1 \times 1$ convolutional layer on each feature map then downsample it to $7 \times 7$. Global attention map $A$ is produced by operating residual connection between resized feature maps. $\oplus$ represents residual connection.

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

describe the activation intensity of attention on feature map. Feature matrix $M$ is produced by concatenating the GAP results, producing a vector of $65536 \times 1 \times 1$. Eq. (1) describes the calculation of $PF_j$.

$$PF_j = A_j \odot f_3 \quad (j = 1, 2, ..., N)$$

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 $2048 \times 1 \times 1$. Feature matrix $M$ is represented by concatenating all condensed $PF_j$, which is represented in Fig. 4.
3.3 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 method shown in Fig. 5 is proposed to present an effective augmentation for original image. For each image, one attention map is randomly chosen for individual augmentation. This attention map is normalized as $A^\ast$.

1) Attention Mixup: Mixup is a augmentation strategy which generates data by mixing overall image and regional data together. As we have a attention map $A_j^\ast$, a detailed region $D_j$ could be extracted by doing threshold.

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

For elements in $A^\ast_j(l,m)$, Eq. (2) set $D_j(l,m)$ to 1 if it is greater than threshold $\theta_m \in [0,1]$, and others to 0. A bounding box surrounding the extracted region could be 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).

$$I_1(p,q) = \gamma I_0(p,q) + (1-\gamma)B(p,q)$$

where $\gamma$ is a parameter range in $[0,1]$ and $B$ stands for the enlarged bounding box. By mixing local feature and global feature together, fine-grained features could be extracted and the model could see target precisely.

2) Attention Patching: Encoder may sensitive to limited part of reception field as valuable spatial feature usually distributes in similar position. To encourage the encoder to exploit feature globally, attention patching is proposed. Specifically, $D$ mentioned in 1) is patched onto the original image $I_0$ to propose patched data $I_2$ as shown in Fig. 5. Attention patching enlarges the model’s interest region, which forces the model to exploit its input globally.

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

$$DM(l,m) = \begin{cases} 0.001, & \text{if } A^\ast(l,m) > \theta_d \\ 1, & \text{otherwise} \end{cases}$$

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

3.4 Soft Distance Regularization

Comparing with original training images, disturbances are introduced by augmentation. (e.g. infection area reduced by attention dimming). To address this problem, we formulate the uncertainty of model via the distance between prediction vectors. Intuitively, the distance $d$ could be modeled as Eq. (5).

$$d(x) = |P(I) - p(x)|$$
where \( x \) denotes the augmented image, \( P(I), p(x) \) represent primary prediction vector and auxiliary prediction vector respectively. However, the distance between \( P(I) \) and \( p(x) \) is unstable before the model well-fitted. We reference ground truth labels to stabilize gradients. As shown in Algorithm. 1, \( P(I) \) is replaced by soft label \( \hat{P}'(I) \), filtering out low confidence inferences. Soft distance \( d'(x) \) can be represented in Eq. (6). The value of \( \theta \) referenced in Algorithm. 1 is 0.7.

\[
d'(x) = \|P'(I) - p(x)\|
\]  

(6)

**Algorithm 1: Soft Distance Regularization**

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

**Output:** \( L_{reg} \): soft distance regularization term

1. Cross entropy loss \( L_{ce}^{prim} \) is calculated between \( P(I) \) and \( G_{lbl}(I) \);
2. \( P(I), p(x) \) are fed into softmax to extract confidence score over all classes, which are \( P^c(I), p^c(x) \);
3. if \( P^c(I) > \theta \) then
   4. Let \( P'(I) = P(I) \)
   5. else
   6. Let \( P'(I) = G_{lbl}(I) \)
7. end
8. Predict variance is represented by soft distance between \( P'(I) \) and \( p(x) \):
   \[
d'(x) = \|P'(I) - p(x)\|
\]
9. Overall loss is combined by \( L_{ce}^{prim} \) and mean predicting variance:
   \[
   L_{reg} = L_{ce}^{prim} + \bar{d}'
   \]
10. return \( L_{reg} \)

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

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

(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. Besides, it is also notable that \( L_{reg} \) punishes soft distance \( d' \), forcing the model to generate the same prediction vectors.

4 Experiments

In this section, extensive experiments are conducted to show the effectiveness of MAG-SD. The model is trained on datasets which containing different types of pneumonia. Each proposed method is evaluated to prove their effectiveness. Then the model is compared between other baseline methods using several metrics.

4.1 Dataset and Experimental Settings

The proposed model is trained and tested on several datasets to evaluate its classification performance and ability of fine-grained pneumonia localization. Details of each dataset is shown in Tab. 1. **Dataset A** is a mutated dataset with 90 COVID-19 from [14] and 168 other pneumonia cases from [16], which directly assess model’s fine-grained classification ability. **Dataset B** is selected from [45] and [16] aiming at assessing the model’s performance on larger scale. **Dataset C** is the largest dataset we operate on, which includes COVID-19 detection and fine-grained pneumonia classification. Quality of pneumonia localization is evaluated by Localization dataset, which has 13 COVID-19 cases with pixel-wise mask from [46] and 118 non-COVID pneumonia cases with bounding box annotation from [16]. In experiments, classic ResNet50 has been adopted as the convolutional feature extractor and its output of layer4 is chosen as feature map. Attention is extracted from the output of layer2, layer3 and layer4 to ensure multiscale attention. Size of attention are \( 28 \times 28, 14 \times 14 \) and \( 7 \times 7 \) respectively. Both training and test sets are divided roughly in the same class proportions. 5-fold cross validation is applied to get reliable results.
Models are trained on two NVIDIA RTX 2080TI GPUs. The optimizer is Stochastic Gradient Descent (SGD) with the momentum of 0.9. For each training, 100 training epochs are deployed, with $10^{-6}$ weight decay, 32 cases per minibatch and $10^{-3}$ learning rate at beginning. Images are resized to $224 \times 224$ when training and testing.

### 4.2 Pre-Processing and Data Augmentation

X-ray images are affected by varied configurations of imaging equipment resolving that radiology image of the same tissue can be different. To ensure the intensity distribution of one tissue is similar over the dataset, Z-score normalization is employed when doing model training and testing. Large contrast distribution also introduced extra noise to the dataset, impacting the performance of trained deep model. Contrast limited adaptive histogram equalization (CLAHE) is proposed to enhance contrast between tissues and restrain noise signal [47].

In image classification, data augmentation has been proved as an effective method to improve robustness and evaluation performance [48]. Augmented data provides more varieties for classification target, remitting impact of overfitting. Random number of transformations are chosen from a sequence of linear transformation to be applied for each training sequence. Transformations list is shown in Tab. 2.

### 4.3 Evaluation Metrics

Several widely adopted metrics are employed. Classification metrics includes Accuracy (ACC), True Positive Rate (TPR), True Negative Rate (TNR) and F1 score. Localization quality is quantified by Intersection over Union (IOU). Accuracy describes the proportion of correctly classified targets, which expressed in Eq. (8).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$  \hspace{1cm} (8)

where TP, TN, FP and FN stand for the number of true positive predictions, true negative predictions, false positive predictions and false negative predictions. TPR, also known as sensitivity, is useful to measure the proportion of true
Table 3: Evaluation of Models

| Model                  | Dataset A | Dataset B | Dataset C |
|------------------------|-----------|-----------|-----------|
|                        | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   |
| VGG16 [20]             | 0.926     | 0.914     | 0.937     | 0.923 | 0.909     | 0.891     | 0.952     | 0.891 | 0.803     | 0.803     | 0.927     | 0.799 |
| ResNet18 [21]          | 0.897     | 0.838     | 0.904     | 0.859 | 0.917     | 0.906     | 0.936     | 0.902 | 0.824     | 0.838     | 0.890     | 0.831 |
| ResNet50 [21]          | 0.928     | 0.932     | 0.915     | 0.921 | 0.922     | 0.921     | 0.963     | 0.926 | 0.831     | 0.844     | 0.934     | 0.832 |
| InceptionV3 [22]       | 0.939     | 0.933     | 0.931     | 0.932 | 0.933     | 0.938     | 0.965     | 0.927 | 0.845     | 0.856     | 0.891     | 0.852 |
| COVID-Net-Large [13]   | 0.937     | 0.932     | 0.947     | 0.936 | 0.939     | 0.941     | 0.966     | 0.936 | 0.836     | 0.842     | 0.938     | 0.829 |
| ResNet [49]            | 0.953     | 0.948     | 0.949     | 0.948 | 0.945     | 0.942     | 0.970     | 0.939 | 0.854     | 0.855     | 0.945     | 0.858 |
| InceptionV3 [49]       | 0.935     | 0.931     | 0.937     | 0.934 | 0.882     | 0.845     | 0.934     | 0.857 | 0.791     | 0.764     | 0.921     | 0.768 |
| BCNN [28]              | 0.961     | 0.941     | 0.972     | 0.954 | 0.948     | 0.953     | 0.971     | 0.948 | 0.844     | 0.831     | 0.942     | 0.832 |
| BCNN(Attention)        | 0.968     | 0.951     | 0.968     | 0.961 | 0.953     | 0.955     | 0.974     | 0.952 | 0.855     | 0.859     | 0.945     | 0.845 |
| FPN [37]               | 0.949     | 0.947     | 0.953     | 0.949 | 0.939     | 0.938     | 0.967     | 0.928 | 0.821     | 0.818     | 0.929     | 0.804 |
| U-Net [35]             | 0.955     | 0.953     | 0.968     | 0.955 | 0.934     | 0.931     | 0.964     | 0.926 | 0.841     | 0.850     | 0.942     | 0.836 |
| MAG-SD(0AUG)           | 0.944     | 0.943     | 0.936     | 0.939 | 0.930     | 0.932     | 0.960     | 0.932 | 0.838     | 0.861     | 0.938     | 0.852 |
| MAG-SD(Proposed)       | 0.966     | 0.957     | 0.970     | 0.962 | 0.961     | 0.955     | 0.966     | 0.954 | 0.878     | 0.884     | 0.959     | 0.877 |

Table 4: Evaluation of CLAHE

| Preprocessing         | Dataset A | Dataset B | Dataset C |
|-----------------------|-----------|-----------|-----------|
|                       | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   |
| without CLAHE         | 0.947     | 0.948     | 0.918     | 0.931 | 0.933     | 0.926     | 0.964     | 0.925 | 0.845     | 0.868     | 0.941     | 0.859 |
| CLAHE                 | **0.966** | **0.957** | **0.970** | **0.962** | **0.961** | **0.955** | **0.966** | **0.954** | **0.878** | **0.884** | **0.959** | **0.877** |

positive predictions over all positive targets, being defined in Eq. (9).

\[ TPR = \frac{TP}{TP + FN} \]  

(9)

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

\[ TNR = \frac{TN}{TN + FP} \]  

(10)

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

\[ F_1 = \frac{2TP}{2TP + FP + FN} \]  

(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 \( A_u \) denotes area of overlap and area of union respectively.

\[ IoU = \frac{A_o}{A_u} \]  

(12)

Table 5: Contribution of Attention Guided Augmentation

| Augmentation          | Dataset A | Dataset B | Dataset C |
|-----------------------|-----------|-----------|-----------|
|                       | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   | ACC       | TPR       | TNR       | F1   |
| \( A^M \)              | 0.954     | 0.952     | 0.947     | 0.949 | 0.937     | 0.934     | 0.968     | 0.931 | 0.852     | 0.856     | 0.943     | 0.849 |
| \( A^M + A^D \)        | 0.959     | 0.951     | 0.952     | 0.942 | 0.947     | 0.944     | 0.950     | 0.941 | 0.867     | **0.885** | 0.947     | **0.884** |
| \( A^M + A^D + A^P \)  | **0.966** | **0.957** | **0.970** | **0.962** | **0.961** | **0.955** | **0.966** | **0.954** | **0.878** | **0.884** | **0.959** | **0.877** |

*\( A^M \): Attention Mix-up; **\( A^D \): Attention Dimming; ***\( A^P \): Attention Patching.*
Table 6: Evaluation of Multisize Attention Maps

| Attention Maps               | Dataset A | Dataset B | Dataset C |
|-----------------------------|-----------|-----------|-----------|
|                             | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   |
| 7*7                         | 0.957 | 0.949 | 0.942 | 0.946 | 0.947 | 0.950 | 0.973 | 0.942 | 0.859 | 0.869 | 0.938 | 0.847 |
| 7*7 + 14*14                 | 0.966 | 0.957 | 0.970 | 0.962 | 0.961 | 0.955 | 0.966 | 0.954 | 0.878 | 0.884 | 0.959 | 0.877 |
| 7*7 + 14*14 + 28*28         | 0.960 | 0.957 | 0.962 | 0.959 | 0.956 | 0.959 | 0.978 | 0.955 | 0.855 | 0.865 | 0.950 | 0.868 |

Table 7: Comparison of Pooling Methods

| Pooling        | Dataset A | Dataset B | Dataset C |
|----------------|-----------|-----------|-----------|
|                | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   |
| GMP            | 0.934 | 0.936 | 0.931 | 0.933 | 0.945 | 0.932 | 0.971 | 0.938 | 0.838 | 0.843 | 0.940 | 0.837 |
| GAP            | 0.941 | 0.940 | 0.943 | 0.938 | 0.944 | 0.939 | 0.969 | 0.938 | 0.845 | 0.855 | 0.947 | 0.844 |
| Attention Pooling | 0.966 | 0.957 | 0.970 | 0.962 | 0.961 | 0.955 | 0.966 | 0.954 | 0.878 | 0.884 | 0.959 | 0.877 |

4.4 Components Validation and Discussion

The method composed for this context could be concluded into attention modules, attention guided data augmentation and soft distance regularization. Each component is studied by evaluating its improvement in classification performance, which is quantified by metrics mentioned above (i.e., Accuracy, TPR, TNR and F1 Score). Performance gain is obtained by the following method: the proposed method is trained on the dataset with all the metrics; then components of the method are been removed or substituted then evaluated on the same dataset. For all the tested models, mean value of each metric is calculated as the final result. Experiments on our model are reported in Tab. 4, 5, 6, 7 and 8 with all metrics. Inter-model comparison could be found in Tab. 3. Fig. 6 shows the interest area proposed by the model. In all the experiments, parameters are maintained unchanged as possible for condition control. The model are trained on the same size of training set then evaluated on the same size of testing set.

4.4.1 Architecture Comparing

Deep explorations into architecture design are proposed by designing experiments using varied methods. To perform this analysis, we evaluated classic coarse-grained deep neural networks (i.e. VGG16, ResNet18, ResNet50 and InceptionV3), COVID-19 oriented architectures (i.e. [49](ResNet), [49](InceptionV3), COVID-Net-Large), high performance fine-grained structure (i.e. BCNN, BCNN(Attention)) and multiscale feature fusion models (i.e. FPN, U-Net). Choosing these deep structures helps to explain our advantages in fine-grained feature extraction. It can be observed in Tab. 3 that our model has noticeably higher performance over other models. Accuracy on dataset B and C reaches 0.961, 0.878 respectively. BCNN with attention get the best performance on TPR, TNR in Dataset B.

Comparing with classic models, our model is specialized for COVID-19 image classification and attention guided architecture has its advantage in fine-grained visual classification task. Most of the other models for COVID-19 show better performance than classic models, however, none of them applies attention mechanism or considers fine-grained features, which impacted their accuracy on large scale, multi-class dataset such as Dataset B and C. Comparing FPN, U-Net with classic models, it can be shown that models considering multiscale feature are highly over InceptionV3 in Dataset A. Performances of multiscale models trained on dataset B are similar to Inception V3, which reaches 0.938 on TPR and overrun U-net in TPR, TNR and F1. In Dataset C, FPN has higher accuracy than VGG16 and U-Net exceeds ResNet50. Results show that multiscale feature fusion models reaches high performance in a relatively simple structure comparing with classic deep models, which leave us a hint that multiscale attention is a promising route to improve the model.

Table 8: Comparison of L2 and Soft Distance Regularization

| Loss          | Dataset A | Dataset B | Dataset C |
|---------------|-----------|-----------|-----------|
|               | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   | ACC  | TPR  | TNR  | F1   |
| L2            | 0.944 | 0.938 | 0.934 | 0.935 | 0.941 | 0.935 | 0.966 | 0.933 | 0.836 | 0.863 | 0.937 | 0.848 |
| Soft Distance | 0.966 | 0.957 | 0.970 | 0.962 | 0.961 | 0.955 | 0.966 | 0.954 | 0.878 | 0.884 | 0.959 | 0.877 |
Figure 6: Demonstration of pneumonia localization. Images are selected from Localization dataset. COVID-19 cases has pixel-wise mask while bounding boxes are provided for other pneumonia. IoU is calculated for each prediction. Localization result is provided by apply threshold onto the attention map of each case. Results show that attention focus on different area when detecting various classes.

BCNN is a FGVC oriented classic model which obtains a stable performance on multiple datasets. In order to evaluate the generalization ability of our attention module, multiscale attention and attention pooling are transported to BCNN to train BCNN(Attention). Statistically, BCNN reaches 0.961, 0.948, 0.844 in Accuracy, which is compatible in all the evaluated models. Attention modules remarkably boosted the performance of BCNN, reaching 0.968 and exceeding our proposed method on Accuracy in Dataset A. In dataset B, BCNN(Attention) acquires 0.974 TNR.

4.4.2 CLAHE Preprocessing

Images collected by different devices probably be distinct in contrast due to configuration variety. CLAHE is employed to relieve the noise brought by contrast distribution. Tab. 4 shows the result that CLAHE obviously improve the performance of proposed model, raising over 0.02 Accuracy on average. Larger datasets such as Dataset B and C are reported to have more performance gain. Model trained without CLAHE is worse on all the metrics in three datasets.

4.4.3 Multiscale Attention Generator and Attention Pooling

Normally, state-of-the-art coarse-grained CNN models suffer from similar global features between classes when operating on FGVC, meaning local feature is the key to improve. local features are effectively localized by our multiscale attention method, which performance is evaluated on COVID-19 datasets. Models trained with attention module (i.e. MAG-SD(0AUG)) and baseline model (i.e. ResNet50) are compared in Tab. 3. Networks with attention achieves better performance than ResNet50 baseline. Metrics show that proposed model surpasses baseline on dataset A and C using all the benchmarks, and gets higher score in ACC, TPR and F1 on dataset B. Our model reaches 0.944, 0.930 and 0.838 on ACC in dataset A, B and C with an over 0.01 average gain comparing with baseline. Furthermore, attention module includes two parts, attention generating and attention pooling. Investigations of these two parts are concluded in two steps. In the first, models are compiled for assessing effectiveness of multiscale attention, with 1, 2 or 3 size of attention maps considered. Results are presented in Tab. 6, which shows that the model achieves the best accuracy in all three datasets when considering 2 feature maps. Model with 3 attention maps have better performance on TPR, TNR and F1 in Dataset B. Low-level texture feature may be ignored using single attention map while to many scale of maps introduce noise and surpass high-level target information. Secondly, we evaluate 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 are presented in Tab. 7 which shows that our proposed attention pooling method surpass GAP and GMP in all three datasets.

4.4.4 Attention Guided Augmentation

Attention emphasis local feature that interested the model. With attention, data could be augmented effectively. Attention guided augmentation is shown in Fig. 5. Models are trained with 0, 1, 2 or 3 augmentation to discuss its effect in COVID-19 CXR image classification task. In the case of 1 augmentation, attention mixup is selected. 2 augmentations model includes attention mixup and attention patching. The results obtained is presented in Tab. 5. In all the datasets, model with all three augmentations has the best Accuracy. In dataset C, Model with 2 augmentations is slightly better in TPR and F1. The proposed augmentations emphasis data according to attention map, minimizing negative effect caused by random augmentations.

4.4.5 Soft Distance Regularization

Soft distance regularization is presented to relieve augmentation variance. To effectively verify the effectiveness of soft distance regularization, we compare it with $L^2$ distance regularization. Tab. 8 illustrated that our proposed regularization method surpasses $L^2$ on all the metrics. Constraint between auxiliary vector and primary vector screen the false prediction introduced by attention guided augmentations. Regularization is calculated between ground truth and auxiliary vector when primary vector cannot provide reliable prediction, keeping the final result away from local minima.

4.4.6 Attention Based Infection Localization

Technically, attention improve the models by roughly localize the part with high activation intensity. This characteristic of attention inspires us to try MAG-SD on localization topics. The models are trained on the Dataset B we proposed, then test on Localization. Fig. 6 demonstrates several cases from Localization dataset. COVID-19 cases has pixel-wise segmentation and non-COVID-19 cases has bounding box for pneumonia infection. Attention maps are $A$ are upsampled from $7 \times 7$ to $224 \times 224$. Localization masks for COVID-19 cases are extracted by applying threshold to the attention maps. Bounding boxes for other pneumonia are produced by simply enclosing the localization masks with rectangles. IoU is calculated to evaluate the quality of localization. Image shown that the attention module we proposed could roughly indicate the position of different type of pneumonia with over 0.25 IoU score. Attention map emphasis the influential part from the input image effectively to improve the model’s performance.

5 Conclusions

We have presented MAG-SD for automatic COVID-19 CXR image classification that reaches the state-of-the-art on our dataset. The proposed novel method achieved splendid performance by treating 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 in Section. 4. 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 is 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 labeled data. Which means new diseases or rare diseases without data available cannot be classified properly. To alleviate this limitation, abnormal detecting and clustering model could be proposed as a guidance for supervised models, which is part of our topic in future work.
Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.KYZ043718114) and Biomedical Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University (Grant No.YG2020YQ17). The authors would like to thank the institutes who generously open-sourced image database.

References

[1] Chen Wang, Peter W Horby, Frederick G Hayden, and George F Gao. A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223):470–473, 2020.
[2] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 72. 2020.
[3] Ioannis Apostolopoulos, Sokratis Aznaouridis, and Mpesiana Tzani. Extracting possibly representative covid-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. *arXiv preprint arXiv:2004.00338*, 2020.
[4] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-PCR testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642, 2020.
[5] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.
[6] Feng Shi, Jun Wang, Jun Shi, Ziyun Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Reviews in Biomedical Engineering*, 2020.
[7] T Franquet. Imaging of pneumonia: trends and algorithms. *European Respiratory Journal*, 18(1):196–208, 2001.
[8] Antoni Torres and Catia Cillóniz. *Clinical Management of Bacterial Pneumonia*. Springer, 2015.
[9] Thirumalaisamy P Velavan and Christian G Meyer. The covid-19 epidemic. *Tropical medicine & international health*, 25(3):278, 2020.
[10] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.
[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilicus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.
[12] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet*, 395(10223):507–513, 2020.
[13] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871*, 2020.
[14] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.
[15] Linda Wang. Figure 1 covid-19 chest x-ray data initiative. [https://github.com/agchung/Figure1-COVID-chestxray-dataset](https://github.com/agchung/Figure1-COVID-chestxray-dataset). Accessed May 29, 2020.
[16] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chest-xray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
[17] Biraja Ghoshal and Allan Tucker. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769*, 2020.
[18] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
[19] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv preprint arXiv:2003.12338*, 2020.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[23] Douglas S Katz and Ann N Leung. Radiology of pneumonia. *Clinics in chest medicine*, 20(3):549–562, 1999.

[24] Alfonso J Rodriguez-Morales, Jaime A Cardona-Ospina, Estefanía Gutiérrez-Ocampo, Rhuvi Villamizar-Peña, Yeimer Holguín-Rivera, Juan Pablo Escalera-Antezena, Lucia Elena Alvarado-Arnez, D Katterine Bonilla-Aldana, Carlos Franco-Paredes, André F Henao-Martínez, et al. Clinical, laboratory and imaging features of covid-19: A systematic review and meta-analysis. *Travel medicine and infectious disease*, page 101623, 2020.

[25] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*, 2019.

[26] Ning Zhang, Jeff Donahue, Ross B Girshick, and Trevor Darrell. Part-Based R-CNNs for Fine-Grained Category Detection. In *European Conference on Computer Vision*, pages 834–849, 2014.

[27] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN Models for Fine-Grained Visual Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015.

[28] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[29] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip Torr. Learn to Pay Attention. In *ICLR 2018 : International Conference on Learning Representations 2018*, 2018.

[30] Jun Wang, Yiming Bao, Yaofeng Wen, Hongbing Lu, Hu Luo, Yunfei Xiang, Xiaoming Li, Chen Liu, and Dahong Qian. Prior-attention residual learning for more discriminative covid-19 screening in ct images. *IEEE Transactions on Medical Imaging*, 39(8):2572–2583, 2020.

[31] Zhiwei Huang, Jinzhao Lin, Liming Xu, Huiqian Wang, Tong Bai, Yu Pang, and Teen-Hang Meen. Fusion high-resolution network for diagnosing chestx-ray images. *Electronics*, 9(1):190, 2020.

[32] Suman Sedai, Dwarikanath Mahapatra, Zongyuan Ge, Rajib Chakravorty, and Rahil Garnavi. Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in x-ray images. In *International Workshop on Machine Learning in Medical Imaging*, pages 267–275. Springer, 2018.
[40] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[41] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[42] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[43] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[44] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[45] Simon Edwardsson. Covid-19 xray dataset. [https://github.com/v7labs/covid-19-xray-dataset](https://github.com/v7labs/covid-19-xray-dataset). Accessed September 23, 2020.

[46] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

[47] Etta D Pisano, Shuquan Zong, Bradley M Hemminger, Marla DeLuca, R Eugene Johnston, Keith Muller, M Patricia Braeuning, and Stephen M Pizer. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital imaging*, 11(4):193, 1998.

[48] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[49] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.