IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2020, accepted February 4, 2020, date of publication February 17, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974242

Learning to Recognize Chest-Xray Images Faster and More Efficiently Based on Multi-Kernel Depthwise Convolution

MENGJIE HU, HEZHENG LIN, ZIMENG FAN, WENJIE GAO, LU YANG, CHUN LIU, AND QING SONG

Pattern Recognition and Intelligent Vision Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Qing Song (priv@bupt.edu.cn)

ABSTRACT The development of convolutional neural networks has promoted the progress of computer-aided diagnostic systems. Details in medical image, such as the texture and tissue structure, are crucial features for diagnosis. Therefore, large input images combined with deep convolution neural networks are adopted to boost the performance in recent research of chest X-ray diagnosis. Meanwhile, due to the variable sizes of thoracic diseases, many researchers have worked to introduce additional module to capture multi-scale feature of images in CNN. However, these efforts hardly consider the computational costs of large inputs and introduced additional modules. This paper aims to automatically diagnose diseases on chest X-rays images quickly and effectively. We propose the multi-kernel depthwise convolution(MD-Conv) which contains depthwise convolution kernels with different filter sizes in one depthwise convolution layer. MD-Conv has high calculation efficiency and few parameters. Because its ability to learn multi-scale feature based on the multi-size kernels, it is appropriate for medical images diagnosis tasks in which abnormalities varied in sizes. In addition, larger depthwise convolution kernels are adopted in MD-Conv to obtain a larger receptive field efficiently, which can ensure sufficient receptive field for high resolution inputs. MD-Conv can be easily applied in modern lightweight networks to replace the normal depthwise convolution layer. We conduct experiments on the Chest X-ray 14 Dataset, which is the largest available chest x-ray dataset, and obtain competitive results. We also evaluate the MD-Conv on the new released dataset for pediatric pneumonia diagnosis. We obtain a better performance of 98.3% AUC than original paper (96.8%) for recognize pneumonia versus normal. Meanwhile we compare the FLOPs and Params of different models to show their efficiency for chest X-rays recognition.

INDEX TERMS Chest x-ray recognition, lightweight networks, multi-kernels depthwise convolution.

I. INTRODUCTION

The development of convolutional neural networks(CNN) has made a dramatic breakthrough in a series of computer vision tasks, which has also promoted the computer-aided diagnosis system. Medical images have grown exponentially in hospital, and disease screening is a time consuming task for radiologists. The computer-aided diagnosis system can help to do preliminary screening and reduce the burden of radiologists.

Chest X-ray is one of the most accessible radiology examinations in the world. Research on chest X-ray includes thoracic disease identification and localization [34], lung regional segmentation [19] and diseases report generation [35]. Among all these studies, the details in medical images, such as textures and structure of lung tissue, are crucial features for diagnosis. That is the reason that the Chest X-ray 14 dataset keeps 1024 × 1024 bitmap images [34] to preserve details, which exceeding the 512 × 512 images in OpenI dataset [3]. Similarly, most global image-based CNN methods adopt large images as inputs, 512 × 512 or even 1024 × 1024 in [2], [18], [37]. And some local image-based CNN methods use 224 × 224 inputs [7], because the 224 × 224 inputs are large enough to take local details for local images. On the other hand, the pathologies in chest X-ray images are highly varied in their shapes and sizes. We conduct

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

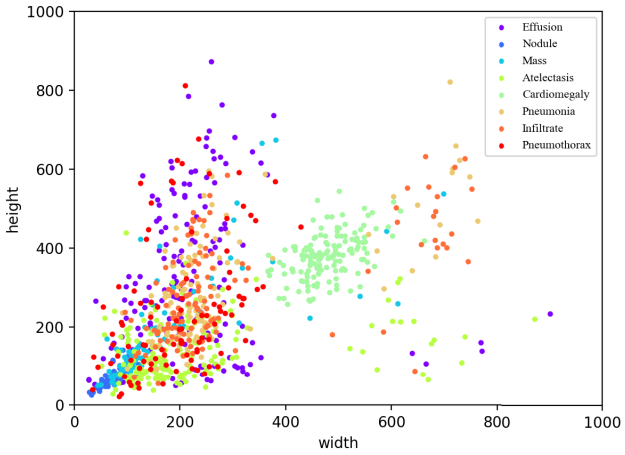


FIGURE 1. Statistics of the different sizes of eight common thoracic pathologies in Chest X-ray 14 dataset.

statistical analysis on the sizes of eight common thoracic pathologies based on the 984 boxes provided by Chest X-ray 14 dataset. As shown in FIGURE. 1, the size of eight common thoracic pathologies varies in a wide range, and even different instances of one thoracic pathology, such as Infiltrate, have different sizes. This requires learning multi-scale convolutional features in CNN. Some recent work attempts to solve this problem by fusing the information from multiple resolutions [39], [41].

Due to the enlarged input as mentioned above, a deeper network is always adopted to ensure the network receptive field is large enough. Many works choose ResNet-50 [13] and DenseNet-121 [5] to extract CNN features [18], [22]. Although this improves the performance, the large inputs combined with deep networks brings quite huge computational costs and parameters, increasing time for network training and optimization. For example, doubled input size will lead to four times training time. And the multiple resolutions feature fusion also cost computation time and store space. Thus this is not conducive to future deployments on mobile and embedded systems.

In this paper, we focus on increasing network receptive field and efficiently learning multi-scale feature. We firstly leverage the lightweight networks. There are many excellent lightweight networks, such as MobileNet [10], ShuffleNet [40], MobileNet-v2 [23], ShuffleNet-v2 [21] and MobileNet-v3 [1]. These networks take the network structures of VGG [28] and ResNet [14] for reference, and replace normal convolution with depthwise convolution to reduce parameters and FLOPs while maintaining the accuracy. By fully considering the balance between computation and accuracy during the design process, these networks have achieved good performance on ImageNet dataset. In addition, we propose the multi-kernel depthwise convolution(MD-Conv), which can capture the multi-scale feature in one convolution layer without introducing extra layers or blocks. Meanwhile the larger depthwise convolution kernels, 5×5 kernels, are adopted in MD-Conv to efficiently obtain a larger receptive field. MD-Conv is appropriate for medical images with abnormalities in various sizes.

We replace the normal depthwise convolution with the proposed MD-Conv in popular lightweight networks, and evaluate the modified model on two public datasets: the Chest X-ray14, which is the largest available chest x-ray dataset, and the Chest X-ray2017, which is a recently released chest x-ray dataset for pediatric pneumonia diagnosis. We achieve state-of-the-art results on both datasets. The modified MD-Conv can successfully identify the chest X-rays quickly and effectively.

The contributions of this paper are as follows.

(1) Compared with modern methods adding complex network and additional block to improve performance, we adopt a lightweight network to quickly recognize chest X-rays which requires small model parameters, and is suitable to employ on embedded systems.

(2) The problem of multi-scale feature learning is studied. Based on the various sizes of thorax diseases and enlarged inputs, we propose MD-Conv, which is conducive to learning the multi-scale feature of different thorax diseases and improving network performance.

(3) The modified MobileNet-v2 with MD-Conv achieves competitive results on the Chest X-ray 2017 dataset and the Chest X-ray 14 dataset.

II. RELATED WORK

A. DEEP LEARNING FOR CHEST X-RAY DIAGNOSIS

Wang *et al.* [34] release the largest chest X-ray dataset and utilize different models to recognize and locate thorax diseases. Since then, a series of studies have been explored based on the large dataset, such as image classification, weakly supervised localization, medical report generation for medical image. In [34], four classic CNN models, AlexNet [16], GoogleNet [31], VGG16 [28], ResNet-50 [13] are compared in the proposed DCNN framework for disease localization. In [22], CheXNet is proposed and demonstrate that DenseNet [5] performs much better on chest x-ray images recognition. And then cascade ConvNet [17], global local fusion method [7] and multi-scale feature are proposed to improve recognition performance, and all these works use ResNet and DenseNet as the basic network for feature extraction. Among all these works, ResNet50 and DenseNet121 are most widely used models. Reference [17] also uses DenseNet161, and [38] reduces the Conv-Block number within a DenseBlock to four to get a light model.

In [15], authors try to utilize CNN to diagnose pediatric pneumonia on chest X-ray images. Based on transfer learning algorithm, it reaches a AUC of 96.8% for recognize pneumonia from normal on chest X-rays dataset. It adopts efficient Inception-v3 [32] as the basic network.

B. MULTI-SCALE METHODS IN COMPUTER VISION TASKS

Though CNN is robust to do recognition on images with objects of different sizes, how to obtain a multi-scale feature

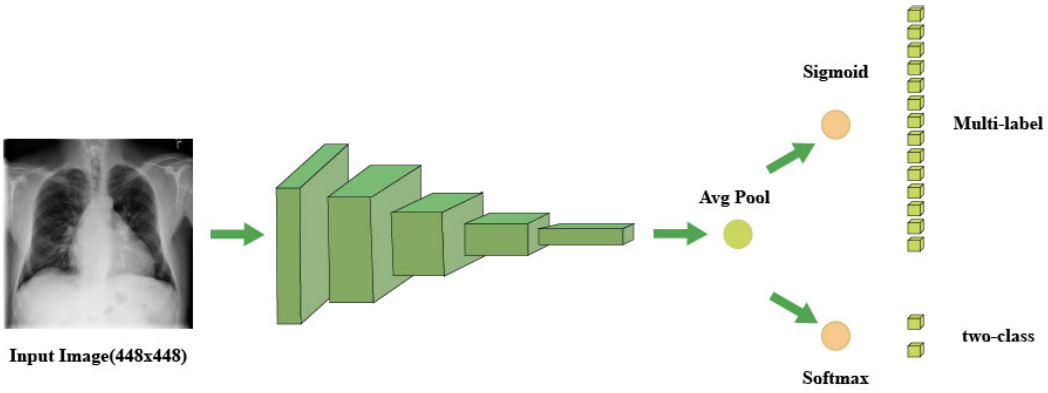


FIGURE 2. The framework of Chest-Xray image recognition.

representation is an important issue in many computer vision tasks.

Traditional approaches use image pyramids to get a more accurate results, for example the multi-scale test in [13]. While most state-of-the-art methods utilize features from different layers to obtain inherent multi-scale in network. FPN [33] uses upsample and latency to generate feature pyramid, and SSD [36] reuses the multi-scale feature maps from different layers. FCN [20] is one of the earliest methods to fuse the multi-scale representations in semantic segmentation. Recently, HRNet [30] for human pose estimation performs repeated multi-scale fusion to achieve state-of-the-art results.

Therefore, in the field of medical deep learning, many multi-scale CNNs have also been proposed to learn multi-scale feature for abnormalities. In [24], [39], additional blocks are added to fuse the multi-scale feature from different layers.

C. EFFICIENT NETWORK DESIGN

With the development of convolution neural networks, researchers have become interested in efficient model design. GoogleNet [31] is one of the earliest networks which is designed for computational efficiency. And since depthwise convolution is proposed [27], depthwise convolution is utilized in modern lightweight network to replace the normal convolution for its efficiency and effectiveness. There are five popular lightweight networks, MobileNet [10], ShuffleNet [40], MobileNet-v2 [23], ShuffleNet-v2 [21] and MobileNet-v3 [1]. Among these five lightweight networks, MobileNet modify VGG structure with depthwise convolution, while others adopt ResNet-like structure. The lightweight network CondenseNet [11] is created based on the extensive and explicit feature reuses structure of DenseNet.

The recent work [29] which is named as HetConv is similar to our work. However, their work focuses on efficient convolution computation, and they design the heterogeneous convolutional filter by 1×1 conv and 3×3 conv. While our work focus on the multi-scale feature learning ability and the large receptive field of the large convolution kernels. Therefore, 3×3 depthwise conv and 5×5 depthwise conv are adopted in our proposed MD-Conv, and 1×1 kernel is not recommended in MD-Conv. The separated channels of depthwise convolution is suitable for implementation of multi-kernel depthwise convolution.

III. METHOD

We aim to build a CNN to quickly and efficiently recognize chest X-ray images. Compared with the widely used networks, ResNet50 and DenseNet121, we adopt the lightweight network MobileNet-v2. The framework of chest X-ray image recognition is shown in FIGURE. 2. In order to learn the muti-scale feature, we propose the MD-Conv to replace the depthwise convolution in MobileNet-v2.

The MD-Conv compared together with normal convolution and depthwise convolution are illustrated in FIGURE. 3. The MobileNet-v2 block with MD-Conv is shown in FIGURE. 3(e).

A. THE MULTI-KERNEL DEPTHWISE CONVOLUTION

For a standard convolution layer, it takes an input feature map $\mathbf{I} \in \mathbb{R}^{H \times W \times M}$, and outputs a feature map $\mathbf{O} \in \mathbb{R}^{H \times W \times N}$, here we assume that the feature map spatial width(W) and height(H) are constant, and M is the number of input channels, N is the number of output channels. For depthwise convolution layer(DW), each of these filters only connects to one input channel. Then an additional layer pointwise convolution(PW) is added after DW to calculate a linear combination of the output of DW. The feature map operations of standard convolution layer and DW + PW are shown in FIGURE. 3(a)(b) respectively. For a standard convolution layer with kernel size of $K \times K$, the computational cost and parameters are computed as:

$$Cost\_norm = K \times K \times M \times N \times H \times W. \quad (1)$$

$$Parms\_norm = K \times K \times M \times N. \quad (2)$$

For a depthwise convolution of $K \times K$ paired with 1×1 pointwise convolution, the computational costs and parameters are

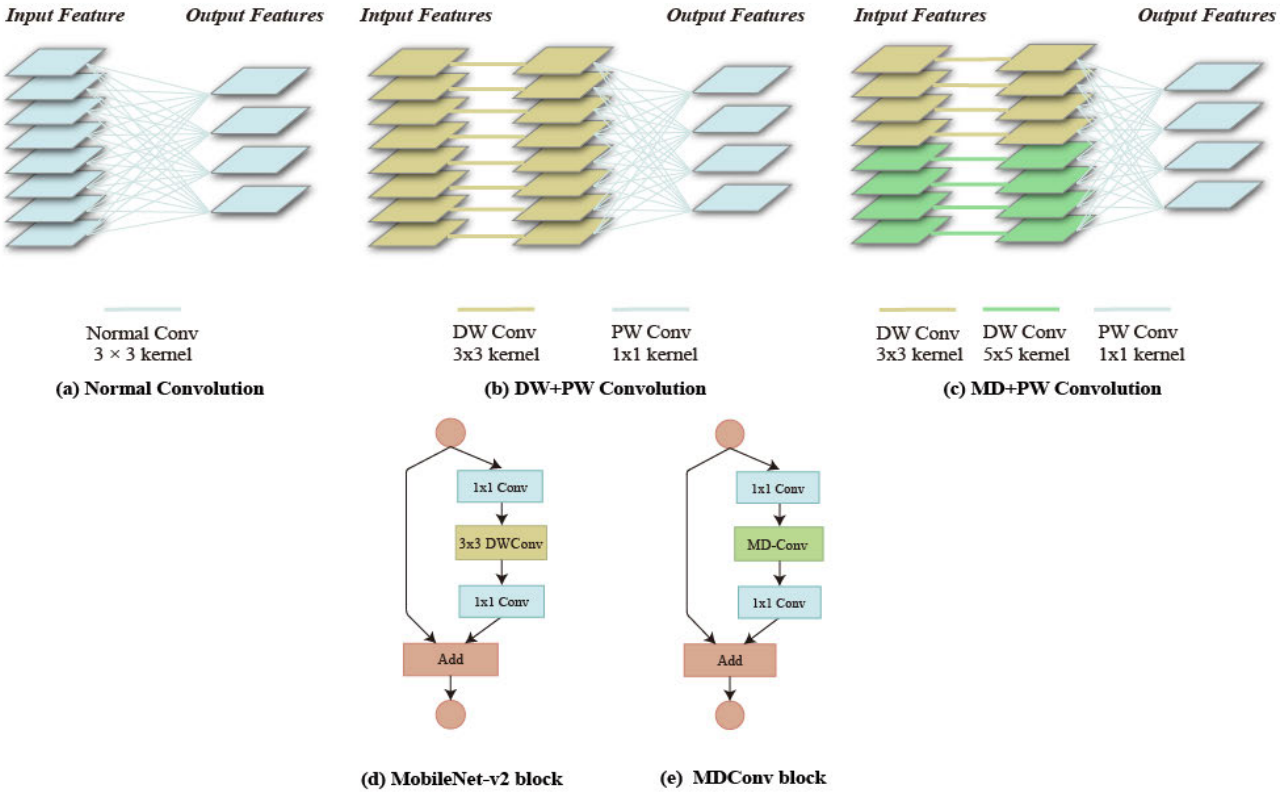


FIGURE 3. (a)(b)(c) show comparison of the different feature operations of normal convolution, depthwise convolution and MD-Conv. (d)(e) show the MobileNet-v2 block with depthwise convolution and MD-Conv.

computed as:

$$Cost\_DW = K \times K \times M \times H \times W + M \times N \times H \times W. \quad (3)$$

$$Parms\_DW = K \times K \times M + M \times N. \quad (4)$$

Because the output channel N is much larger than K^2, the computational cost and parameters of DW + PW is $\frac{1}{N}$ times the normal convolution.

We introduce the multi-kernel depthwise convolution(MD-Conv), which contains both 3 × 3 DWConv and 5 × 5 DWConv in a multi-kernel depthwise convolution layer. The feature map computation of MD-Conv + PW is shown in FIGURE. 3(c). For each depthwise convolution kernel in the MD-Conv layer, it corresponds to only one input channel, thus we can easily implement the MD-Conv by channel split operation. Here, we consider an alternative slicing of the input feature $\mathbf{I} = [\mathbf{i}^{3\times3}, \mathbf{i}^{5\times5}]$, where $\mathbf{i}^{3\times3}$ corresponding to input feature for 3 × 3 DWConv and $\mathbf{i}^{5\times5}$ for 5 × 5 DWConv. As shown in FIGURE. 3(c), the yellow feature maps correspond to $\mathbf{i}^{3\times3}$, and the green feature maps correspond to $\mathbf{i}^{5\times5}$. Therefore, the output of DWConv is

$$\mathbf{O}_{MD-Conv} = [W_{3\times3}\mathbf{i}^{3\times3}, W_{5\times5}\mathbf{i}^{5\times5}]. \quad (5)$$

And then the PW is followed to fuse the separable channels of MD-Conv.

For the MD-Conv with an input feature map of $H \times W \times M$ and same size output feature map, the computation cost and parameters are:

$$Cost\_MD = (9 \times \mathbf{i}^{3\times3} + 25 \times \mathbf{i}^{5\times5}) \times H \times W. \quad (6)$$

$$Parms\_MD = 9 \times \mathbf{i}^{3\times3} + 25 \times \mathbf{i}^{5\times5}. \quad (7)$$

For the 3 × 3 DWConv with the same inputs and outputs, the number of its parameters is $9 \times (\mathbf{i}^{3\times3} + \mathbf{i}^{5\times5}) \times H \times W$. MD-Conv only adds tiny amounts of parameters compared to 3 × 3 DWConv.

1) THE EFFICIENT WAY TO IMPROVE RECEPTIVE FIELD

Receptive field plays an important role in CNN. For visual tasks with low resolution inputs, the receptive field of standard 3 × 3 filter is sufficient. While for medical images with high resolution, increased receptive field is needed to retain more details. The network receptive field is computed as:

$$\mathbf{RF}_n = \mathbf{RF}_{n-1} + (k_n - 1) \times s_n. \quad (8)$$

where n represents the network layer, k_n is the kernel size of layer n, s_n is the stride of layer n.

To increase the receptive field, a simple method is to increase the depth of the network, such as densenet121. The receiving area of the two stacked 3 × 3 convolutions is 5, which is the same as the 5 × 5 convolutions, and has less computation and parameters. In [32], two stacked 3 × 3 convolution are used to replace 5 × 5 convolution for efficiency.

For lightweight networks, an efficient way to increase the receptive field is to adopt a larger kernel depthwise convolution. A 5×5 depthwise convolution has the same receptive field as 5×5 convolution, and it requires much less extra computational costs and parameters. Two stacked 3×3 depthwise convolution is not adopted due to no channel cross talk [40].

Additionally, two stacked 3×3 convolutions have a computation cost of $18 \times M \times N \times H \times W$, while the 5×5 depthwise convolution has a computation cost of $25 \times M \times H \times W$. The 5×5 depthwise convolution has much smaller computation cost while maintaining the same receptive field.

Therefore, the part of 5×5 depthwise convolution can more efficiently obtain a large receptive field in MD-Conv.

2) COMPARED WITH THE MULTI-SCALE FEATURE FUSION METHOD

The image-level classification always requires coarse-scale features with high semantic information and context, while detection and segmentation need fine-scale features to capture detailed appearance information. Therefore, in [39], feature fusion at different resolutions is used. In [41], a multi-resolution CNN is adopted to recognize nodules of different sizes. These methods all require additional convolution blocks and downsample or upsample to integrate information from multiple scales.

The MD-Conv is proposed to replace the normal depthwise convolution. In one MD-Conv block as shown in FIGURE. 3(e), the MD-Conv can extract the multi-scale feature based on the multi-scale depthwise convolution kernels, and then the 1×1 convolution is followed to fuse information of different scales. This introduces no addition convolution layers or operations, and adds only slight parameters of 5×5 depthwise convolution. Moreover, the MD-Conv can be easily integrated into modern networks.

B. HOW AND WHERE TO USE MD-CONV

In this subsection, we will discuss how and where to use MD-Conv, which is the ratio of different kernels($\mathbf{i}^{3\times3}$, $\mathbf{i}^{5\times5}$) in one MD-Conv layer, and where the MD-Conv should be used in a network.

1) BEST $\mathbf{i}^{3\times3}/\mathbf{i}^{5\times5}$ VALUE

As mentioned above, the MD-Conv consists of 3×3 depthwise convolution and 5×5 depthwise convolution. The parameters $\mathbf{i}^{3\times3}$ and $\mathbf{i}^{5\times5}$ control the number of different types of kernels in one MD-Conv layer. To find the best ratio of $\mathbf{i}^{3\times3}/\mathbf{i}^{5\times5}$, we do experiments on different ratios of $\mathbf{i}^{3\times3}/\mathbf{i}^{5\times5}$ in MD-Conv. In these experiments, we replace all the normal depthwise convolution layer in MobileNet-v2 with MD-Conv layer and perform the experiments on the Chest-Xray 14 Dataset. The results are shown in FIGURE. 4. According to it, we choose $\mathbf{i}^{3\times3}/\mathbf{i}^{5\times5}$ ratio of 1:1 to achieve the best cost and accuracy trade-off.

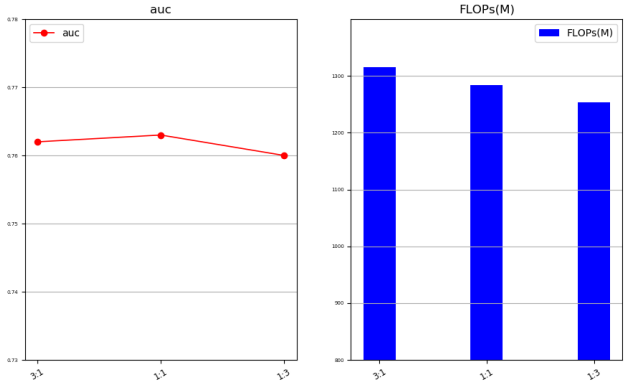


FIGURE 4. Comparison results of different ratios of $\mathbf{i}^{3\times3}/\mathbf{i}^{5\times5}$. The left one shows mean AUC score of 14 diseases, the right one shows the computation cost.

TABLE 1. The network architecture of modified MobileNet-v2 with MD-Conv. MD-Conv denotes whether MD-Conv is used in bottleneck, n denotes the repeated times of bottleneck in stage, and s denotes stride for the first layer of each stage.

Layer	Operator	n	MD-conv	s
conv	conv2d	-	-	2
stage0	bottleneck	1	-	1
stage1	bottleneck	2	-	2
stage2	bottleneck	3	-	2
stage3	bottleneck	4	-	2
stage4	bottleneck	3	-	1
stage5	bottleneck	3	✓	2
stage6	bottleneck	1	-	1
conv	conv2d	1	-	1
pool	avgpool	-	-	-
conv	conv2d	-	-	-

2) WHERE TO USES MD-CONV

MD-Conv is proposed to efficiently obtain the multi-scale feature and improve receptive field. We find that it is not necessary to use MD-Conv in all layers. Placing MD-Conv in right location can improve performance while saving FLOPs and Params. We conduct experiments to explore which layer the MD-Conv should be used. As shown in TABLE 4, we achieve the best AUC score replacing all the normal depthwise convolution layer in layer2 with MD-Conv. The final network architecture of modified MobileNet-v2 is shown in TABLE 1.

IV. EXPERIMENT

A. CHEST X-RAY DATASETS

Chest X-ray is the most common and efficient technique for screening and diagnosis of lung-related diseases, such as pneumonia, cardiomegaly, lung node. Several chest X-rays datasets have released for study. Early datasets, such as [12], [26], contain only hundreds of chest X-ray images, which are too few for deep learning. OpenI [3] is a publicly available dataset collected by Indiana University, it contains

3955 radiology reports and the corresponding 7,470 chest X-ray images. The original image size is 512 × 512.

Chest X-ray14 is the largest chest X-ray dataset provided by Wang *et al.* [34]. It provides 112,120 frontal-view chest X-ray images of 30,805 unique patients. Each image is annotated with one or more labels of 14 common thorax diseases. The chest X-ray image was originally extracted from DICOM file and resized to 1024 × 1024, which containing more details. Based on the large amount of chest X-ray images in Chest X-ray14, many deep learning methods can be performed on it.

Recently, [15] releases a chest X-rays dataset for pneumonia classification. It contains 5,856 labeled chest X-ray images from children, which also distinguish between bacterial and viral. The sizes of images are various. According to statistics, about 90% images have the aspect ratio between 1.0 and 1.5, and 94.8% images have larger image size than 512 × 512.

In this section, we first perform a series of comparative experiments on the Chest X-ray14 Dataset. For convenience, we perform most experiments from scratch without ImageNet [4] pre-trained. Finally, we compare our methods with others on both the Chest X-ray14 Dataset and the Chest X-ray2017 Dataset, for both chest x-ray dataset, we follow the official patient-wise split.

B. IMPLEMENTATION DETAILS

We use lightweight network architectures and modify the last fully connected (fc) layer to 14 for the Chest X-ray14 dataset, and 2 for the Chest X-ray2017 dataset. For the multi-label classification problem of Chest X-ray14 dataset, a BCE Loss is adopted.

We implement all experiments using PyTorch framework. Adam optimizer is adopted with a initial value of $10e^{-5}$. And cosine learning rate [11] is adopted. We train 100 epochs with a batch size of 64 on 4 GPU. The weight decay is set to $10e^{-5}$ and early stop is used. During the training phase, we perform data augmentation by resizing the original images to 512×512 and randomly cropping to 448×448(for 224×224 input, resize to 256×256), randomly rotating from -10 to 10 degrees, and the probability of random horizontal flips is 0.5. The network weights are initialized using MSRA initialization [9].

C. CHEST X-RAY 14 DATASET

1) CLASSIFICATION PERFORMANCE UNDER INPUTS IN DIFFERENT SCALES

We first run experiments to understand how kernel size and input scale affect the performance of different thoracic pathologies. As shown in FIGURE. 5, (a) For most results, the larger input images improve the performance. As it shows that the green ones has a higher AUC than red ones, expect for Pneumonia and Consolidation which get similar scores. (b) For Atelectasis, Mass, Nodule and Fibrosis, a huge improvement is obtained by enlarged input, and the 3 × 3 kernel size performs the best. And for Mass, Nodule and Fibrosis, on 224 × 224 input, 3 × 3 kernel performs much better than 5 × 5 kernel. This is because the pathologies with small size, as shown in FIGURE. 1, are easier to recognize on large image, and 3 × 3 is more suitable for feature extraction. (c) For diseases like Cardiomegaly, Pneumothorax, Emphysema and Pleural_Thickening, 5 × 5 kernel on 448 × 448 inputs achieve best AUC. As shown in FIGURE. 1, most of Cardiomegaly and Pneumothorax have larger size than 200 × 200. Thus for these diseases, a larger kernel is more suitable for feature extraction on 448 × 448 inputs.

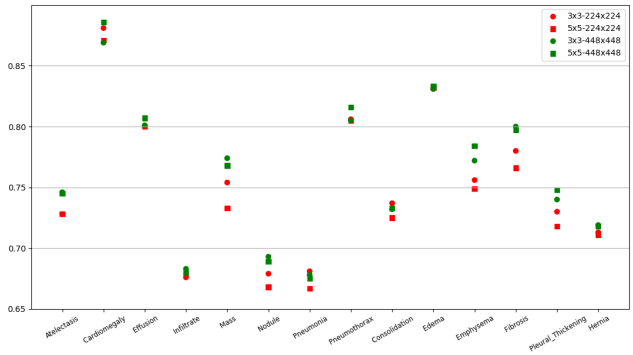


FIGURE 5. Performance of model with different kernels size on different input. 3 × 3 and 5 × 5 represent the depthwise convolution kernel size we adopt in MobileNet-v2. 224 × 224, 448 × 448 represent the input image shape for network.

TABLE 2. The AUC score for MD-Conv compared with other kernel sizes on different input size.

Conv type	input 224 × 224	input 448 × 448
3x3	**0.754**	0.772
5x5	0.747	0.763
7x7	0.746	0.756
MD-Conv(all_layer)	**0.754**	**0.777**

We can conclude that for the thoracic pathologies with different sizes, we should try to extract the multi-scale feature for better recognition. Therefore, MD-Conv is right for this problem. We do experiments on the Chest X-ray14 dataset with MobileNet-v2 and TABLE 2 shows the performance of MD-Conv on different input size. MD-Conv performs best on both 224 × 224 and 448 × 448 inputs. And 7 × 7 kernel performs the worst which is too large for both input sizes, thus 7 × 7 kernel is not adopted in MD-Conv.

2) PERFORMANCE OF FIVE LIGHTWEIGHT NETWORKS

Five lightweight networks MobileNet [10], ShuffleNet [40], MobileNet-v2 [23], ShuffleNet-v2 [21] and MobileNet-v3 [1] have different performance on ImageNet dataset which indicates that these networks will also perform the same in chest X-ray images recognition task. In order to determine which network we use, we conduct experiments using these networks on the chest x-ray14 dataset and the results are shown in TABLE 3. According to the experimental result, we choose MobileNet-v2 as the baseline. This is

TABLE 3. Mean auc of five popular lightweight networks on Chest X-ray14.

Network	mean_auc	FLOPs(M)
MobileNet	0.740	2291
MobileNet-v2	**0.772**	1223
MobileNet-v3	0.766	865
ShuffleNet	0.769	521
ShuffleNet-v2	0.758	2369

TABLE 4. Mean auc vs. FLOPs for modified MobileNet-v2 with MD-Conv in different location.

Layer	mean_auc	FLOPs(M)
all	0.777	1278
layer1	0.779	1235
layer2	0.777	1247
layer3	0.775	1236
layer5	**0.782**	**1228**
layer1,2	0.776	1260
layer1,3	0.778	1249
layer1,5	0.781	1241
layer2,3	0.779	1260
layer2,5	0.776	1252
layer3,5	0.781	1241
layer1,2,3	0.770	1273
layer1,2,5	0.776	1265
layer1,3,5	0.775	1254
layer2,3,5	0.774	1265
basic	0.772	1223

because the depthwise convolution layer in MobileNet-v2 has increased channels which are enough for each depthwise convolution group with different kernel size in MD-Conv that ShuffleNet-v1 dose not have. Meanwhile, MobileNet-v2 has the best accuracy in chest X-rays recognition as shown in TABLE 3.

3) ABLATION STUDY ON WHERE AND WHETHER TO USE MD-CONV

To explore where and whether we should use MD-Conv in MobileNet-v2, we conduct ablation experiments. In MobileNet-v2, the stride of layer1, layer2, layer3 and layer5 is 2, and using MD-Conv in these layers can make receptive field increase in multiples. Therefore, we use MD-Conv only in these layers, and the results are shown in TABLE 4. We can find that using MD-Conv in all layers achieve the low AUC score and its FLOPs are the largest. Using MD-Conv in layer5 achieves the highest AUC and the fewest FLOPs. Therefore, we use MD-Conv only in layer5, as shown in TABLE 1. We guess MobileNet-v2 with MD-Conv in layer5 can obtain a suitable receptive field for input size of 448×448. Using MD-Conv in all layers is not necessary.

TABLE 5. The mean auc score and FLOPs(M) with or without MD-Conv in MobileNet-v2 and MobileNet-v3.

Network	mean_auc	FLOPs(M)
MobileNet-v2	0.772	1223
MobileNet-v2-mdconv	**0.777**	1278
MobileNet-v3	0.766	865
MobileNet-v3-mdconv	**0.785**	885

In order to explore the effect of MD-Conv, we also conduct additional ablation experiments using MobileNet-v3, MD-Conv is used in all layers, the results are shown in TABLE 5. As can be seen, MD-Conv improves the performance of both networks.

4) CLASSIFICATION PERFORMANCE COMPARED WITH THE STATE-OF-THE-ART APPROACH

We compare our results with Resnet50 [34], DenseNet + LSTM [38], DenseNet121 [8] based on the official dataset split. The results are shown in TABLE 6. As we can see, our method performs better than [34] which only adopts ResNet-50. While for [38] which introduces LSTM in network, and [8] which adopts DenseNet121 and uses additional data PLCO Dataset [6]. Although our results are not the best, we provide a baseline for latter research studying lightweight networks in chest X-ray images recognition task and the probability to identifying diseases in embedded and mobile devices. Meanwhile, we only provide a basic network for chest X-ray recognition, which can achieve competitive results with less computation costs and parameters and can be easily used in any other networks with depthwise convolution.

The comparison of computational costs of ResNet50, DenseNet121, and our modified MobileNet-v2 is shown in TABLE 7. As the table shows, our modified model is much lighter on FLOPs and Params. Thus we think there is more space to make improvements based on our basic network. And with such small computation cost and parameters, we still outperform the results of [34].

D. CHEST X-RAY 2017 DATASET

To verify the generalization ability of our model, we also do experiments on the ChestX-ray2017 Dataset released by [15]. We use MobileNet-v2 with MD-Conv to recognize pneumonia versus normal, bacterial versus viral pneumonia. The results are shown in Table 8.

For Chest-Xray recognition of pneumonia versus normal, we achieve an accuracy of 93.4%, outperforms [15] by 0.6%. And the sensitivity outperforms [15] by 4.2%. Though the specificity is 2.3% lower, we achieve an AUC of 98.3%, while is higher than 96.8% in [15]. The AUC curve is shown in FIGHRE 6. Besides, our modified MobileNet-v2 with MD-Conv is lighter than Inception-v3 adopted in [15].

TABLE 6. Comparison results of different methods. we list fourteen abnormalities and their AUCs, we also list the method adopted in every paper.

Abnormality	Wang et al. [34]	Yao et al. [38]	Gündel et al. [8]	Ours
Atelectasis	0.716	0.772	0.767	0.767
Cardiomegaly	0.807	0.904	0.883	0.896
Effusion	0.784	0.859	0.828	0.828
Infiltration	0.609	0.695	0.709	0.696
Mass	0.706	0.792	0.821	0.752
Nodule	0.671	0.717	0.758	0.733
Pneumonia	0.633	0.713	0.731	0.725
Pneumothorax	0.806	0.841	0.846	0.815
Consolidation	0.708	0.788	0.745	0.742
Edema	0.835	0.882	0.835	0.857
Emphysema	0.815	0.829	0.895	0.829
Fibrosis	0.769	0.767	0. 818	0.816
PT	0.708	0.765	0.761	0.749
Hernia	0.767	0.914	0.896	0.735
mean auc	0.738	0.803	0.804	0.782
method	Resnet50	DenseNet+LSTM	DenseNet121	modified MobileNet-v2

TABLE 7. The computation cost and parameters of some basic networks.

Method	FLOPs(M)	Params(M)
ResNet50	16395	23.5
DenseNet121	11405	7.0
ours	**1241**	**2.3**

TABLE 8. Comparison results between ours and [15]. pne represents pneumonia and bac represents bacterial.

Class	Method	accuracy(%)	sensitivity(%)	specificity(%)
pne vs. norm	[15]	92.8	93.2	**90.1**
	ours	**93.4**	**97.4**	86.8
bac vs. viral	[15]	90.7	**88.6**	90.9
	ours	**91.0**	88.5	**92.6**

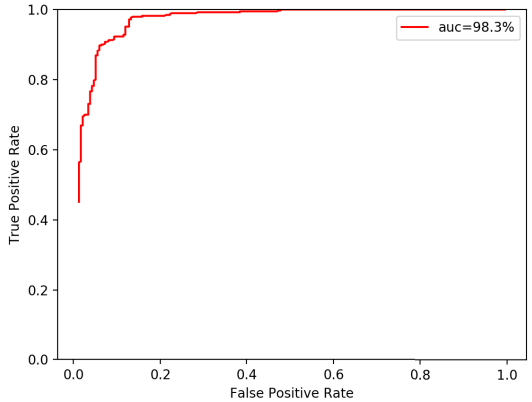


FIGURE 6. The ROC curve of our method for detecting pneumonia versus normal. The area under the ROC curve(AUC) is 98.3%.

E. QUALITATIVE ANALYSIS

We obtain the discriminative regions for each disease by Grad-CAM to show the visual explanation of how CNN recognize Chest-Xray images. As shown in FIGURE 7, different

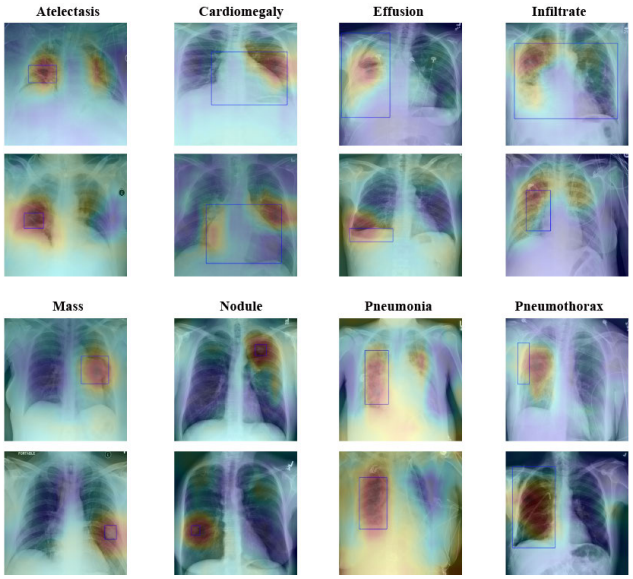


FIGURE 7. Examples of some network attention visualization on test image in Chest X-ray14 dataset. The attention map is generated by Grad-CAM [25], the red region represents the place where disease appears probably. The blue boxes is the ground-truth bounding box provided in the dataset.

diseases have different sizes. The Mass and Nodule are small and have equal length and width, concluded in size distribution in FIGURE 1 and the visualization results in FIGURE 7, the location heatmaps concentrate on an approximate circular area. And for Pneumonia, the location heatmaps focus on most chest region, and the area with most attention is in the blue ground-truth box. For Cardiomegaly, the ground-truth boxes include all the heart region, while the heatmaps pay attention to the enlarged heart margin which can also recognize Cardiomegaly effectively. Since that, the model can locate the disease region no matter the shapes and sizes of the diseases.

V. CONCLUSION

In this paper, we explore the various sizes of thoracic pathologies and multi-scale feature learning for thoracic pathologies recognition. We propose MD-Conv, which contains both 3 × 3 depthwise convolution and 5 × 5 depthwise convolution to quickly and efficiently recognize Chest X-ray images. We conduct experiments on both the Chest X-ray14 Dataset and the Chest X-ray2017 Dataset, and our modified MobileNet-v2 with MD-Conv outperforms the basic network adopted in other methods. The MobileNet-v3 with MD-Conv is better, and have much less computational costs and parameters. We believe that the results will be better than the current version, and may even exceed other existing methods with carefully adjusting hyperparameters in training phase. And there will be more work can be done on efficient model in the future.

REFERENCES

[1] A. Howard, M. Sandler, G. Sandler, L.-C. Chu, B. Chen, M. Chen, W. Tan, Y. Wang, R. Zhu, V. Pang, Q. Vasudevan, H. V. Le, and Adam, "Searching for mobilenetv3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[2] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative attention mining for weakly supervised thoracic disease pattern localization in chest X-rays," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, Cham, Switzerland: Springer, Sep. 2018, pp. 589–598.

[3] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Jul. 2015.

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

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

[6] J. K. Gohagan, P. C. Prorok, R. B. Hayes, and B.-S. Kramer, "The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the national cancer institute: History, organization, and status," *Controlled Clin. Trials*, vol. 21, no. 6, pp. 251S–272S, Dec. 2000.

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

[8] S. Gundel, S. Grbic, B. Georgescu, S. Liu, A. K. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *Proc. Iberoamerican Congr. Pattern Recognit.*, 2018, pp. 757–765.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[10] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[11] G. Huang, S. Liu, L. V. D. Maaten, and K. Q. Weinberger, "CondenseNet: An efficient DenseNet using learned group convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2752–2761.

[12] S. Jaeger, S. Candemir, S. K. Antani, Y. J. Wang, P. Lu, and G. R. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imag. Med. Surg.*, vol. 4, no. 6, pp. 475–477, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 770–778.

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

[15] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[17] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *Proc. Int. Conf. Image Anal. Recognit.*, 2018, pp. 546–552.

[18] Z. Li, C. Wang, M. Han, Y. Xue, and F. F. Li, "Thoracic disease identification and localization with limited supervision," in *Proc. CVPR*, 2018, pp. 8290–8299.

[19] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, and J. Pu, "SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images," *Computerized Med. Imag. Graph.*, vol. 75, pp. 66–73, Jul. 2019.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[21] N. Ma, X. Zhang, H.-T. Zheng, and S. Jian, "ShuffleNet V2: Practical guidelines for efficient cnn architecture design," in *Proc. ECCV*, 2018, pp. 116–131.

[22] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, M. P. Lungren, A. Y. Ng, and K. Shpanskaya, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: https://arxiv.org/abs/1711.05225

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," 2018, *arXiv:1801.04381*. [Online]. Available: http://arxiv.org/abs/1801.04381

[24] 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," 2018, *arXiv:1808.08280*. [Online]. Available: http://arxiv.org/abs/1808.08280

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

[26] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of Radiologists' detection of pulmonary nodules," *Amer. J. Roentgenol.*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[27] L. Sifre, "Rigid-motion scattering for texture classication," Ph.D. dissertation, Ecole Polytechnique, Palaiseau, France, 2014.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[29] P. Singh, V. Kumar Verma, P. Rai, and V. P. Namboodiri, "HetConv: Heterogeneous kernel-based convolutions for deep CNNs," 2019, *arXiv:1903.04120*. [Online]. Available: http://arxiv.org/abs/1903.04120

[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*. [Online]. Available: http://arxiv.org/abs/1902.09212

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

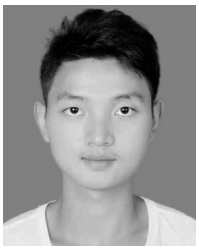
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[33] 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. (CVPR)*, Jul. 2017, pp. 2117–2125.

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

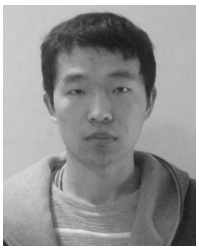
[35] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.
[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[37] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*. [Online]. Available: https://arxiv.org/abs/1710.10501
[38] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*. [Online]. Available: https://arxiv.org/abs/1710.10501
[39] L. Yao, J. Prosky, E. Problez, B. Covington, and K. Lyman, "Weakly supervised medical diagnosis and localization from multiple resolutions," 2018, *arXiv:1803.07703*. [Online]. Available: https://arxiv.org/abs/1803.07703
[40] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," 2017, *arXiv:1707.01083*. [Online]. Available: http://arxiv.org/abs/1707.01083
[41] W. Zuo, F. Zhou, Z. Li, and L. Wang, "Multi-resolution CNN and knowledge transfer for candidate classification in lung nodule detection," *IEEE Access*, vol. 7, pp. 32510–32521, 2019.

MENGJIE HU received the Ph.D. degree from Beihang University, Beijing, China, in 2017. She is currently a Lecturer with the Beijing University of Posts and Telecommunications. Her current research interests are primarily in computer vision and machine learning, especially object detection, visual tracking, and visual geometry.



HEZHENG LIN received the B.Sc. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2019, where he is currently pursuing the master's degree with the Pattern Recognition and Intelligent Vision Laboratory (PRIV). His research interests include lightweight networks, video classification, and object detection.

ZIMENG FAN received the B.Sc. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2017, where she is currently pursuing the master's degree with the Pattern Recognition and Intelligent Vision Laboratory (PRIV). Her research interests cover lightweight networks, image classification, and human pose estimation.



WENJIE GAO was born in Henan, China, in 1994. He received the B.Sc. degree from the Nanyang Institute of Technology. He is currently pursuing the M.Sc. degree with the Beijing University of Posts and Telecommunications. His research interests include image classification, 3-D image detection, and human pose estimation in computer vision.

LU YANG received the bachelor's degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2012, where he is currently pursuing the Ph.D. degree with the Automated Institute. He has participated in the MSCOCO 2018 Challenge, one of the top competitions in the field of computer vision, and won the DensePose champion. He has been involved in research work with the Pattern Recognition and Intelligent Vision Laboratory, since 2012. His research interests include the fields of artificial intelligence, computer vision, and machine learning.

CHUN LIU received the Ph.D. degree from the University of Kassel, Germany, in 2014. She is currently a Lecturer with the Beijing University of Posts and Telecommunications (BUPT). Her main research interests are primarily in intelligent computation and optimization, especially evolutionary algorithms in solving optimization problems, e.g., planning and scheduling.

QING SONG received the Ph.D. degree from Tianjin University, Tianjin, China, in 2006. She is currently a Scientific Researcher with the Beijing University of Posts and Telecommunications (BUPT), where she is also involved in computer vision technology study. She is also the Founder of the Pattern Recognition and Intelligent Vision Laboratory (PRIV) and led the PRIV Team to the Championship of COCO2018-DensePose Challenge. She is also in charge of many national, provincial and ministerial projects, and enterprise cooperation projects. She has published more than 70 academic articles in international journals and conferences.