IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 8, 2021, accepted July 19, 2021, date of publication July 26, 2021, date of current version August 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3099479

An Improved Faster R-CNN for Pulmonary Embolism Detection From CTPA Images

HONGFANG YUAN[1], YAJUN SHAO[1], ZHENHONG LIU[1], AND HUAQING WANG[2]
[1]College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
[2]College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Corresponding authors: Hongfang Yuan (yuanhf@mail.buct.edu.cn) and Huaqing Wang (hqwang@mail.buct.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 81871328, and in part by the Center of Beijing University of Chemical Technology (BUCT)-China-Japan Friendship Hospital (CJFH) under Grant XK2020-04.

ABSTRACT Computer-aided detection of pulmonary embolism is an important technology method for diagnosing pulmonary embolism, which can help doctors diagnose quickly and save a lot of manpower. However, due to the small area of pulmonary embolism in the Computed Tomography Pulmonary Angiography (CTPA) slice images, some previous methods for detecting pulmonary embolism have a high number of false detection and missed detection. This study proposes a detection method of pulmonary embolism based on the improved faster region-based convolutional neural network (Faster R-CNN) named More Accurate Faster R-CNN (MA Faster R-CNN). A new feature fusion network named Multi-scale Fusion Feature Pyramid Network (MF-FPN) is proposed by extending and adding two bottom-up paths on the Feature Pyramid Network (FPN). It enhances the feature extraction capability of the entire network by transmitting low-level accurate location information, and makes up for the original information lost after multiple down-sampling, strengthens the use of detailed information, which is more helpful to the detection of small object. In the prediction module, the residual block is added before the fully-connected layer to deepen the network and enhance the classification accuracy, named residual prediction module (RPM). Compared with the original Faster R-CNN, the proposed MA Faster R-CNN which combines MF-FPN and RPM has a higher detection precision and solves the problems of false detection and missed detection of pulmonary embolism effectively. The average precision (AP) reached 85.88% on the CTPA pulmonary embolism dataset used in this article.

INDEX TERMS Faster R-CNN, multi-scale fusion FPN, residual prediction module, pulmonary embolism.

I. INTRODUCTION

Pulmonary embolism exists in the pulmonary artery and is a thrombus caused by an endogenous or exogenous embolus that blocks central, lobar, segmental, or subsegmental pulmonary arteries, causing an obstruction of pulmonary circulation and influencing pulmonary artery pressure and right heart pressure [1]. About a third of cases got sudden death due to pulmonary embolism which is a common disease with high morbidity and mortality [2], [3]. CTPA is a primary technical means for clinical diagnosis of pulmonary embolism and can diagnose pulmonary embolism quickly and noninvasively, which highlights the pulmonary artery and its branches by injecting high density contrast agent into the vein, pulmonary embolism shows a relatively dark area in the brighter pulmonary artery, because the contrast agent dissolves in the blood and the embolism does not absorb the contrast agent, the clinical features of pulmonary embolism include complete filling defect, partial filling defect, "track shape", and vessel wall defect [1].

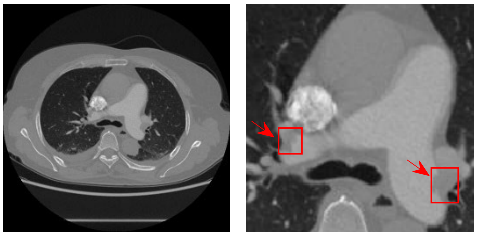


FIGURE 1. Left: The original CTPA slice image of the lung. Right: The enlarged view of the pulmonary artery area.

As shown in Fig. 1, the left image is the CTPA slice image of the lung, and the right image is the enlarged

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

area of the pulmonary artery, the two dark areas marked in the highlighted vessels are pulmonary embolism. It can be observed that pulmonary embolism occupies a very small proportion of the CTPA slice image, which causes a challenge for the detection of pulmonary embolism. The clinical diagnosis method is to manually screen CTPA images, which requires the radiologists to check each branch of the pulmonary artery. The accuracy is influenced by many human factors including the radiologists' professional ability and eye fatigue.

Therefore, it is very necessary to detect pulmonary embolism through computer-aided detection methods. Traditional pulmonary embolism detection methods need to be completed in stages, which is inefficient and inaccurate.

Deep learning has brought great convenience to object detection. Region-based convolutional neural network (R-CNN) is the pioneering work for object detection, which extracts 1k∼2k region proposals by using selective search, solves the location problem of convolutional neural network (CNN) through ''recognition using regions''. But the proposals need to be transformed to the fixed size to adapt to CNNs, which is very time-consuming [4].

Faster R-CNN optimizes the candidate region selection algorithm and introduces the region proposal network (RPN) so that the detection network and the region proposal generator share the same feature map, and object detection can be achieved more effectively and quickly. For RPN which is a fully convolutional network, it takes images of any size as input and object proposals with object score as output [5]. Christian Eggert *et al.* tested the performance of Faster R-CNN for objects of various sizes in the generation of proposals and classification stages, and applied Faster R-CNN to detect company logo in the field of small objects [6]. Mask R-CNN adds a branch to predict the mask on the basis of Faster R-CNN and fixes the quantitative misalignment of Region of Interest (RoI) pooling with RoI align, which is used for object detection and extracting masks [7]. Region-based Fully Convolutional Networks (R-FCN) which is one of the state-of-the-art methods proposed position-sensitive score maps to solve the problem of missing translation-invariance after RoI pooling [8].

However, Faster R-CNN adopts only a single-scale feature, the performance of multi-scale detection is still better for small object detection [9]. Liu *et al.* proposed single shot multi-box detector (SSD) which uses feature maps of different scales for detection and the detection accuracy is higher in a smaller image size compared with other single stage networks [10]. Cai *et al.* proposed the multi-scale CNN (MS-CNN) to detect multi-scale objects, which consists of the proposal generation network with multi-scale output layers, and its performance reaches 15 fps on small object dataset [11]. Setio *et al.* detected pulmonary nodules using multi-view convolutional networks which reduced false positive in pulmonary nodules detection effectively [12]. Li *et al.* applied Faster R-CNN to the detection of the same object, improved the extraction of detail information by merging feature maps, and achieved better performance compared with other methods [13]. FPN adopts ConvNet's multi-level feature structure from low to high semantics to perform multi-scale feature fusion through the lateral connection and top-down pathway [9]. Although FPN introduces a top-down path, it is still limited by one-way information flow.

An Yang *et al.* used Faster R-CNN with U-Net to recognize pulmonary tuberculosis in CT images, and introduced residual block to solve the problem of network degradation. Compared with the original model, the sensitivity, specificity and AUC of the improved model were improved by 2.48%, 1.23% and 2.90%, respectively [14]. Deconvolutional single shot detector (DSSD) introduced residual block for each prediction layer to improve the sub-network, the performance of the prediction module with the residual block seems to be significantly better than the original prediction module [15]. Residual block was proposed by He *et al.* and proved to be helpful for increasing the depth of the network and solving the problem of gradient disappearance and gradient explosion [16].

The proportion of pulmonary embolism in the image is very small, multiple down-sampling reduces the resolution of the object and the location information is blurred, which is not good for the detection of pulmonary embolism. The use of precise location information at the bottom layer is very important for enhancing the feature pyramid and improving the positioning ability of the feature extractor. A deep prediction module is also essential to improve the precision of the model. To solve the above problems, an improved Faster R-CNN is proposed in this study.

The main contributions of this paper are as follows:

1) The combination of different level feature map is discussed. MF-FPN is proposed to detect pulmonary embolism more effectively by utilizing the precise location information extracted from the lower layer of the feature extraction network. Experimental results prove that the application of MF-FPN to Faster R-CNN can solve missed detection of pulmonary embolism effectively compared with the original FPN.
2) By adding the residual structure to the prediction module, the depth of the classification network is increased and the average precision of pulmonary embolism detection is improved. Compared with the original prediction module, the proposed residual prediction module (RPM) has a higher confidence score for object detection.
3) MA Faster R-CNN which combines MF-FPN and RPM is proposed for the detection of pulmonary embolism with small areas, which has a higher AP than the current mainstream object detection algorithm. The problems of missed detection and false detection for pulmonary embolism have been solved effectively.

II. THE PROPOSED METHOD

The architecture of MA Faster R-CNN is shown as Fig. 2. Finetuned SE-ResNet-50 + MF-FPN is selected as the

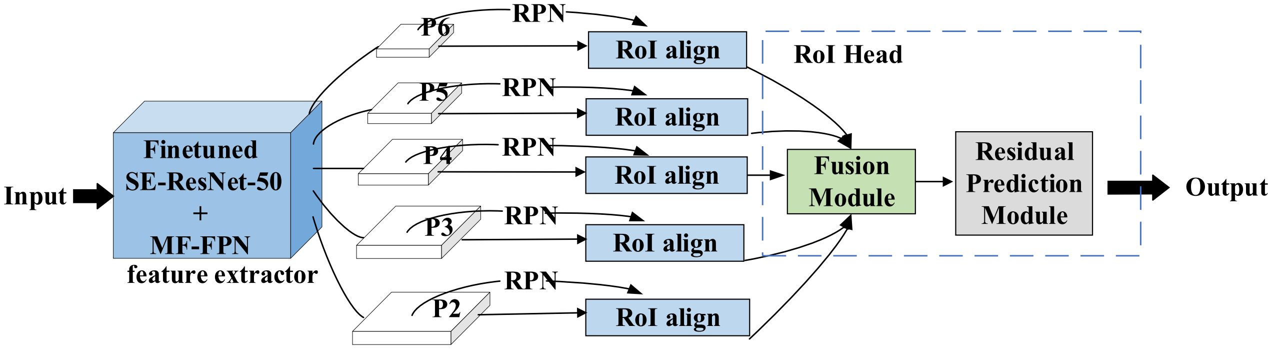


FIGURE 2. The MA faster R-CNN architecture, including MF-FPN applying to feature extractor and RPM in RoI head.

feature extraction network of the model. Finetuned SE-ResNet-50 adds the squeeze-and-excitation (SE) blocks to ResNet-50 to model the interdependence between model channels and recalibrate the feature response related to the channel, brings significant improvements in performance for the model [17]. MF-FPN combines low-level information with high-level information through the top-down and bottom-up paths, which improves the detection accuracy of the whole network. The feature map generated by MF-FPN uses a deep fully convolutional network RPN to generate nearly cost-free region proposals which shares features with the detection network and is helpful to improve the speed and precision of detection. Region proposals use RoI align to generate the fixed-size feature maps for the next fusion module and compress feature to speed up the processing. The pooled feature maps are combined by the fusion module and then predicted by residual prediction module which deepens prediction network and improves the accuracy of the network.

The whole network is synthesized into a unified network through shared convolutional features for detecting objects. In this section, the feature extraction network is first introduced, followed by the improved RoI Head module.

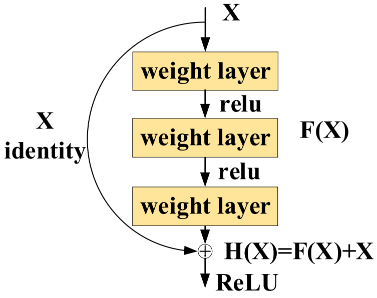


FIGURE 3. The building block of residual learning.

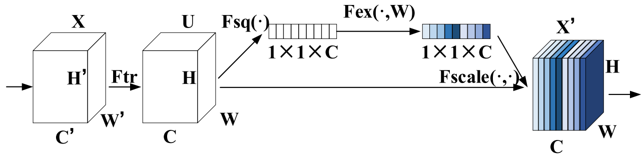


FIGURE 4. The squeeze-and-excitation (SE) block.

A. FEATURE EXTRACTOR

1) BACKBONE NETWORK

A deep feature extractor for training the model with a larger capacity is needed to extract the feature information of pulmonary embolism effectively and improve the accuracy of detection. However, the problems of gradient disappearance and network degradation will occur as the increase of network layers.

In this study, finetuned SE-ResNet-50 was selected as the backbone network of the improved Faster R-CNN for extracting feature. The main contribution of this network is to increase the residual blocks and SE blocks.

Residual blocks are used for solving the problem of network degradation, gradient disappearance and gradient explosion caused by the increase of network layers. The structure of the residual block is shown as Fig. 3. Formally, the underlying expected mapping is expressed as H(X), the stacked residual layer F(X) = H(X) − X, and the original input X becomes F(X) + X through "shortcut connections" [16].

SE blocks are introduced to improve the representational ability of the model by enabling the network to achieve dynamic channel feature recalibration so that the model can use global information to selectively emphasize informative features and suppress less useful features [17]. The structure of the SE block is shown as Fig. 4. The input X is mapped to the feature map U through a convolutional operator F_{tr}. The features U first aggregate the H × W channel feature maps through the squeeze operation F_{sq} to generate a channel descriptor which is used for generating the embedding of the global channel-wise characteristic, and the features U of H × W × C are squeezed to generate features of 1 × 1 × C. Global average pooling is selected as the aggregation technique of squeeze operation to generate channel-wise statistics. To make full use of the feature information, the squeeze operation is followed by the excitation operation, which takes the channel descriptor of the global channel-wise characteristic as input and outputs a set of per-channel modulation weights. The gating mechanism of sigmoid activation is selected as the function for capturing the dependency on the channel. Two FC layers are used to

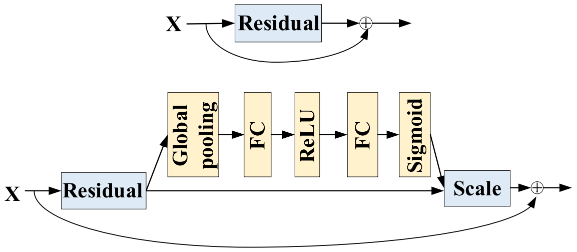


FIGURE 5. (a): The simplified structure of residual block. (b): The structure of SE-ResNet block.

TABLE 1. The architecture of finetuned SE-ResNet-50.

Layers	Output size (Input size 512×512)	Finetuned SE-ResNet-50
C1	128×128	3×3, 32, stride=2 3×3, 32, stride=1 3×3, 64, stride=1 3×3 max pool, stride=2
C2	128×128	$\begin{bmatrix} 1\times1, 64, \text{stride=1} \\ 3\times3, 64, \text{stride=1} \\ 1\times1, 256, \text{stride=1} \\ FC[16, 256] \end{bmatrix}\times3$
C3	64×64	$\begin{bmatrix} 1\times1, 128, \text{stride=1} \\ 3\times3, 128, \text{stride=2} \\ 1\times1, 512, \text{stride=1} \\ FC[32, 512] \end{bmatrix}$ $\begin{bmatrix} 1\times1, 128, \text{stride=1} \\ 3\times3, 128, \text{stride=1} \\ 1\times1, 512, \text{stride=1} \\ FC[32, 512] \end{bmatrix}\times3$
C4	32×32	$\begin{bmatrix} 1\times1, 256, \text{stride=1} \\ 3\times3, 256, \text{stride=2} \\ 1\times1, 1024, \text{stride=1} \\ FC[64, 1024] \end{bmatrix}$ $\begin{bmatrix} 1\times1, 256, \text{stride=1} \\ 3\times3, 256, \text{stride=1} \\ 1\times1, 1024, \text{stride=1} \\ FC[64, 1024] \end{bmatrix}\times5$
C5	16×16	$\begin{bmatrix} 1\times1, 512, \text{stride=1} \\ 3\times3, 512, \text{stride=2} \\ 1\times1, 2048, \text{stride=1} \\ FC[128, 2048] \end{bmatrix}$ $\begin{bmatrix} 1\times1, 512, \text{stride=1} \\ 3\times3, 512, \text{stride=1} \\ 1\times1, 2048, \text{stride=1} \\ FC[128, 2048] \end{bmatrix}\times2$
		Average pool, FC

parameterize the gate mechanism and reduce the complexity of the model. The output weights are mapped to the feature maps U to generate the output of the SE block. The network structure of SE-ResNet module is shown as Fig. 5, which embeds SE blocks into ResNet.

Szegedy *et al.* stated that the computational cost of a 5 × 5 convolution with n filters over a grid with m filters is 25/9 = 2.78 times that of a 3 × 3 convolution with the same number of filters. But a 5 × 5 convolution can be replaced by a multi-layer small convolution kernel with the same input size and output depth, and parameters could be reduced by sharing the weights between adjacent tiles [18]. In this paper, the ideas of Szegedy were used to finetune SE-ResNet-50. Three 3 × 3 convolutions were used to replace 7 × 7 convolution in C1 of SE-ResNet-50. By replacing large convolution kernels with multi-layer small convolution kernels, parameters could be reduced. The input image size is 512 × 512, and the output size after the stem block of finetuned SE-ResNet-50 is 16 × 16. The output size of C2, C3, C4, C5 is 1/4, 1/8, 1/16 and 1/32, corresponding to the original input size respectively. The finetuned architecture of SE-ResNet-50 is shown as Table 1.

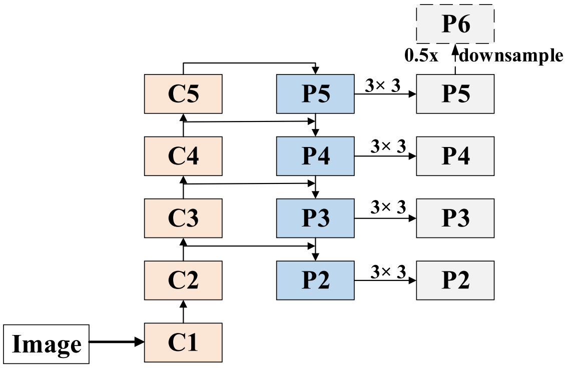


FIGURE 6. FPN with finetuned SE-ResNet-50 as the backbone.

2) MULTI-SCALE FUSION FPN (MF-FPN)

The area of pulmonary embolism extracting from the final feature map is very small after multiple down-sampling operations. Original Faster R-CNN only uses the high-level features from the feature extraction network for prediction, the feature maps extracting from high-level have rich semantic information but the location information from low-level is lost.

FPN performs feature fusion of different layers by the structure of top-down up-sampling and lateral connection [9]. FPN with finetuned SE-ResNet-50 as the backbone is shown in Fig. 6.

Lin *et al.* demonstrated the importance of reusing low-level features in the application of small object detection [9].

In this paper, FPN is improved by adding two bottom-up paths. The improved architecture named MF-FPN is shown in Fig. 7. The backbone network still adopts finetuned SE-ResNet-50. Part (b) continues to adopt the architecture of lateral connections and top-down up-sampling, and performs down-sampling on C5' to generate C6', mainly to fuse down-sampling elements to generate P6' used for the next stage. {C2', C3', C4', C5'} are generated by the combination of the 1 × 1 convolution on the finetuned SE-ResNet-50 feature layers and up-sampling feature maps, and the combination method is elementwise addition. This process is iterative and the calculation of building block is shown in Fig. 8(a).

Part (c) is the major improvement, adding the bottom-up architecture. First, this work appends 1 × 1 convolution on

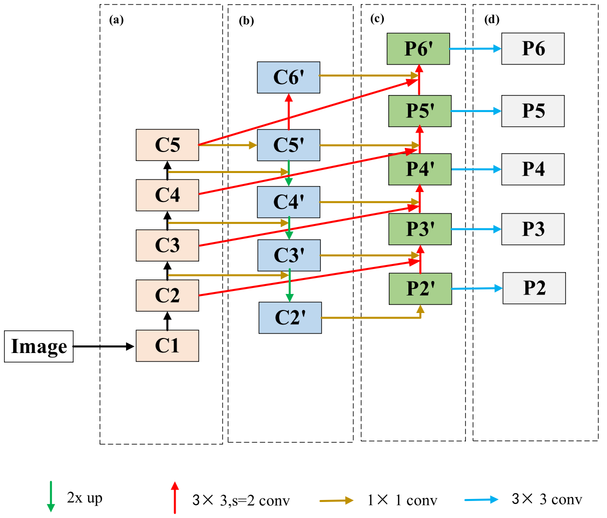


FIGURE 7. The architecture of MF-FPN based on finetuned SE-ResNet-50.

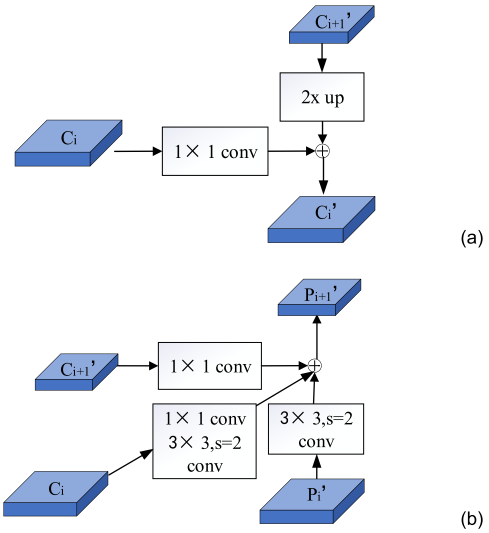


FIGURE 8. (a): Lateral connection and top-down up-sampling are merged by adding. (b): Bottom-up path augmentation. The down-sampled elements consist of two parts, one is generated by bottom-up iteration, and the other is generated by the down-sampling of the near original feature layer.

{C2', C3', C4', C5', C6'} to get the feature maps, and then the feature maps are merged with the down-sampling elements. The down-sampling elements consist of two parts, one is generated by bottom-up iteration, and the other is generated by the down-sampling of the original input feature layers {C2, C3, C4, C5}, which are the lower layers near the current layers. The down-sampling operation is performed by 3 × 3 convolution with a stride of 2. The detailed process of the algorithm is shown in Fig. 8(b). Part (d) is the output of MF-FPN, which generates the final feature map {P2, P3, P4, P5, P6} by performing 3 × 3 convolution on {P2', P3', P4', P5', P6'}, so that objects of different scales can be detected at different feature levels.

The advantage of adding two bottom-up paths is that the location information extraction capability of the entire network can be enhanced by adding a bottom-up path aggregation network to deliver low-level accurate location information, and the original feature hierarchy without fusion has the most characteristic representation in the current level. Mapping the nearest lower layer of the original input feature layer to the output node can make up for the original feature information lost after multiple down-sampling and further strengthen the location information using lower level. In this way, multiple fusions have increased the use of low-level information. All levels have rich semantic information and location information, and the location ability of the entire network is improved, which is more conducive for detecting small objects.

B. ROI HEAD

1) FUSION MODULE (FM)

The original Faster R-CNN architecture adds RPN and RoI pooling on the top layer of the feature extraction network, and can only extract information through a single-scale feature map [5]. The improved MF-FPN can obtain sufficient information by using multi-scale feature map. In this paper, RPN heads are attached to each output level of the MF-FPN, the proposals of each layer shall be processed by RoI pooling layer and the pooled multi-scale feature maps are fused together by FM.

The purpose of RoI pooling layer is to obtain fixed-size outputs for the RoIs with different size using pooling, so as to compress features, reduce parameters and speed up training and testing time. The RoI pooling layer has two modes: RoI pooling and RoI align. RoI pooling divides the RoI with the height and width (h, w) into a H × W grid, the size of each sub-window is (h/W, w/W), and uses max pooling on each sub-window to the corresponding output unit. The pooled feature map has the fixed-size H × W, H and W are the hyperparameters and independent of the size of RoI [19]. However, each quantization operation in the process of mapping the h × w RoI to the H × W feature map will correspond to slight regional feature misalignment. To enhance the positioning information of objects, RoI pooling is replaced by RoI align.

RoI align was proposed by He *et al.* and applied to Mask R-CNN [7]. RoI align traverses RoI and divides the RoI into k × k unit. In this process, the quantization operation is cancelled. The four regularly sampled positions of each unit are calculated by bilinear interpolation, which alleviates the error loss caused by quantization operations and is more beneficial to the detection of small objects, and then max pooling is used to aggregate. Results are insensitive to the exact sampling locations without performing quantization [7].

FM selects two effective combination methods, elementwise sum and elementwise product [15]. Elementwise sum makes fusion by converting the pooled feature map of fixed size to the same channel, then adds the corresponding elements, and elementwise product makes fusion by multiplying the corresponding elements of the feature maps. In the following experiments, the two fusion methods are

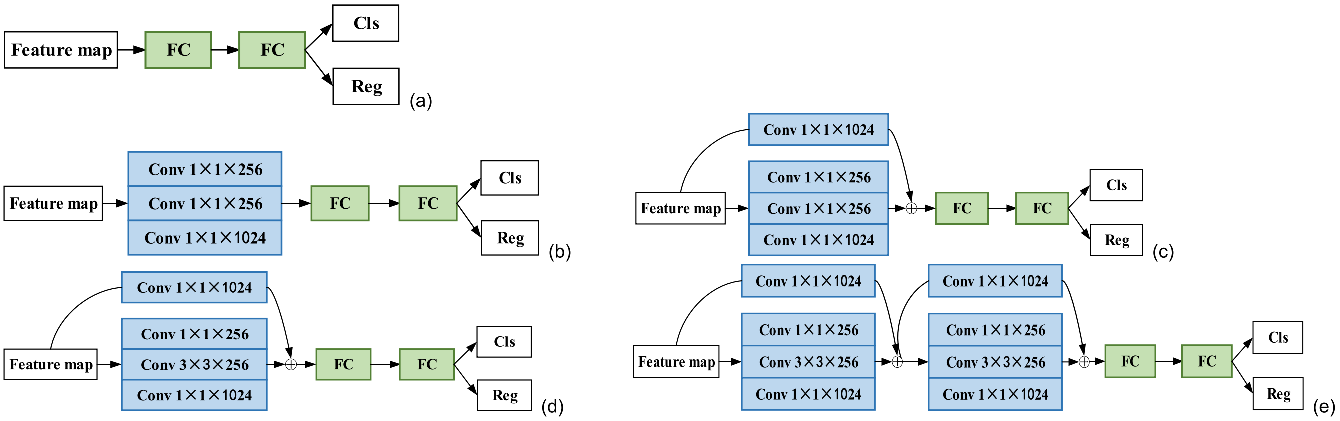


FIGURE 9. The structure of the prediction module (PM).

tested. The results show that elementwise sum is more accurate.

2) RESIDUAL PREDICTION MODULE (RPM)

In the original prediction module (PM) architecture, two fully-connected (FC) layers are directly connected to the pooled feature map, which are shown in Fig. 9(a) and named PM(a). Ren *et al.* pointed out that deep classifier and deep feature extractor are equally important for object detection, deep classifier has a complementary effect on deep features, and the addition of convolutional layer can enhance classification [20].

Inspired by this, in order to increase the network depth of PM, PM(c) adds RPM before the FC layer, as shown in Fig. 9(c). Residual network solves the problem of gradient disappearance of deep network and has a faster convergence speed [20]. Compared with the addition of conventional convolutional network structure PM(b), as shown in Fig. 9(b), the residual network structure is easier to be optimized through the skip connection, which solves the problem of network degradation. PM(d) shown in Fig. 9(d) changes the size of the convolution kernel of the second convolution operation to 3 × 3, and PM(e) shown in Fig. 9(e) adds a residual block structure on the basis of PM(d). Comparative experiments of different prediction modules are shown in Table 4.

III. EXPERIMENTS

A. DATASET

The dataset used in this study is CTPA images of 30 patients from the China-Japan Friendship Hospital marked by professional radiologists, not a public dataset.

The original dataset is a series of DICOM format images corresponding to each patient, and converted to JPG format by preprocessing, which is easy to read. 7771 CTPA slice images containing pulmonary embolism were selected for this experiment, each image has a corresponding pulmonary embolism mask, one group of them is shown as Fig. 10. The dataset is divided into 3736 images for train dataset, 1737 images for validation dataset, and 2298 images for testing dataset.

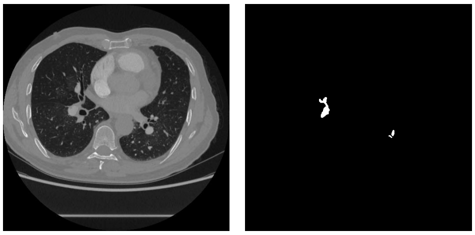


FIGURE 10. Left: The CTPA slice image. Right: The corresponding pulmonary embolism mask of the left CTPA slice image marked by radiologists.

The experiments referred to PASCAL VOC detection benchmark for training Faster R-CNN [5]. The dataset which is segmentation format marked by the radiologists is converted to the PASCAL VOC format which includes three folders to adapt to the training code [22], the folder of Annotations stores the bounding box position information of embolus in XML format, the folder of ImageSets stores file names of training, validation, and test images in TXT format, the folder of JPEGImages stores all CTPA slice images in JPG format.

B. EXPERIMENTAL SETTING AND TRAINING

The hardware environment for experiments is: Nvidia V100 GPU. Deep learning framework MXNet is adopted in the experiments. MXNet supports the parallel operation of CPU and GPU, which enables deep learning with a huge amount of calculation to be completed in a short time.

In this study, original Faster R-CNN is pretrained through ImageNet and the end-to-end training is adopted for the model. Experiments refer to the parameters trained by Ren *et al.* on Faster R-CNN [5]. According to the experimental task, the model parameters are optimized through multiple experiments and set as follow:

Stochastic gradient descent (SGD) was used to optimize with momentum of 0.9 and weight decay of 0.0005. The scale of anchor boxes is set to (2, 4, 8, 16, 32) for small object and 3 aspect ratios of (1:1, 1:2, 2:1) to adapt to this task. Non-maximum suppression (NMS) is adopted to remove the overlapping proposals according to cls scores, the intersection over union (IoU) threshold for NMS is 0.7 and 2000 top proposals are returned after NMS in the training of RPN. Proposal whose IoU larger than 0.5 is regarded as positive samples. The base learning rate used is 0.004 and the batch size used is 8.

The following experiments evaluate the AP of pulmonary embolism detection mainly, which is the current main metric for object detection. AP depends on precision and recall, where precision is based on true positive TP (i.e., the sample is positive and it is classified as positive) and FP (i.e., the sample is negative and it is classified as positive), defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall is based on true positive TP and FN (i.e., the sample is positive and it is classified as negative), defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

AP is defined as follows, where p represents precision and r represents recall:

$$\text{AP} = \int_0^1 \text{p(r)dr} \tag{3}$$

For the problem of high sample imbalance in the detection of pulmonary embolism, sensitivity and specificity are suitable metrics to evaluate the classification performance with imbalanced classes [23].

Sensitivity is equal to recall and represents the proportion of positive correctly predicted samples to the total number of positive samples, which evaluates the performance of the model in predicting actual positive samples [23]. Sensitivity is based on true positive TP and false negative FN, defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Specificity represents the ratio of negative correctly predicted samples to the total number of negative samples, which evaluates the performance of the model in predicting actual negative samples [23]. Specificity is based on true negative TN (i.e., the sample is negative and it is classified as negative), false positive FP, defined as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5}$$

Experiments set the IoU threshold of the proposal box and the ground truth box to exceed 0.5 as TP, otherwise it is FP.

TABLE 2. The comparison of AP and speed among different feature extractors.

Backbone	AP(%)	Train time (sec/img)
ResNet-50	77.02	0.1031
SE-ResNet-50	77.68	0.1037
Finetuned SE-ResNet-50	77.83	0.1075
DenseNet-201	77.79	0.1369
SE-ResNet-50 + FPN	83.87	0.1160
Finetuned SE-ResNet-50 + FPN	84.05	0.1176
SE-ResNet-50 + MF-FPN	84.91	0.1172
Finetuned SE-ResNet-50 + MF-FPN	**85.07**	0.1199

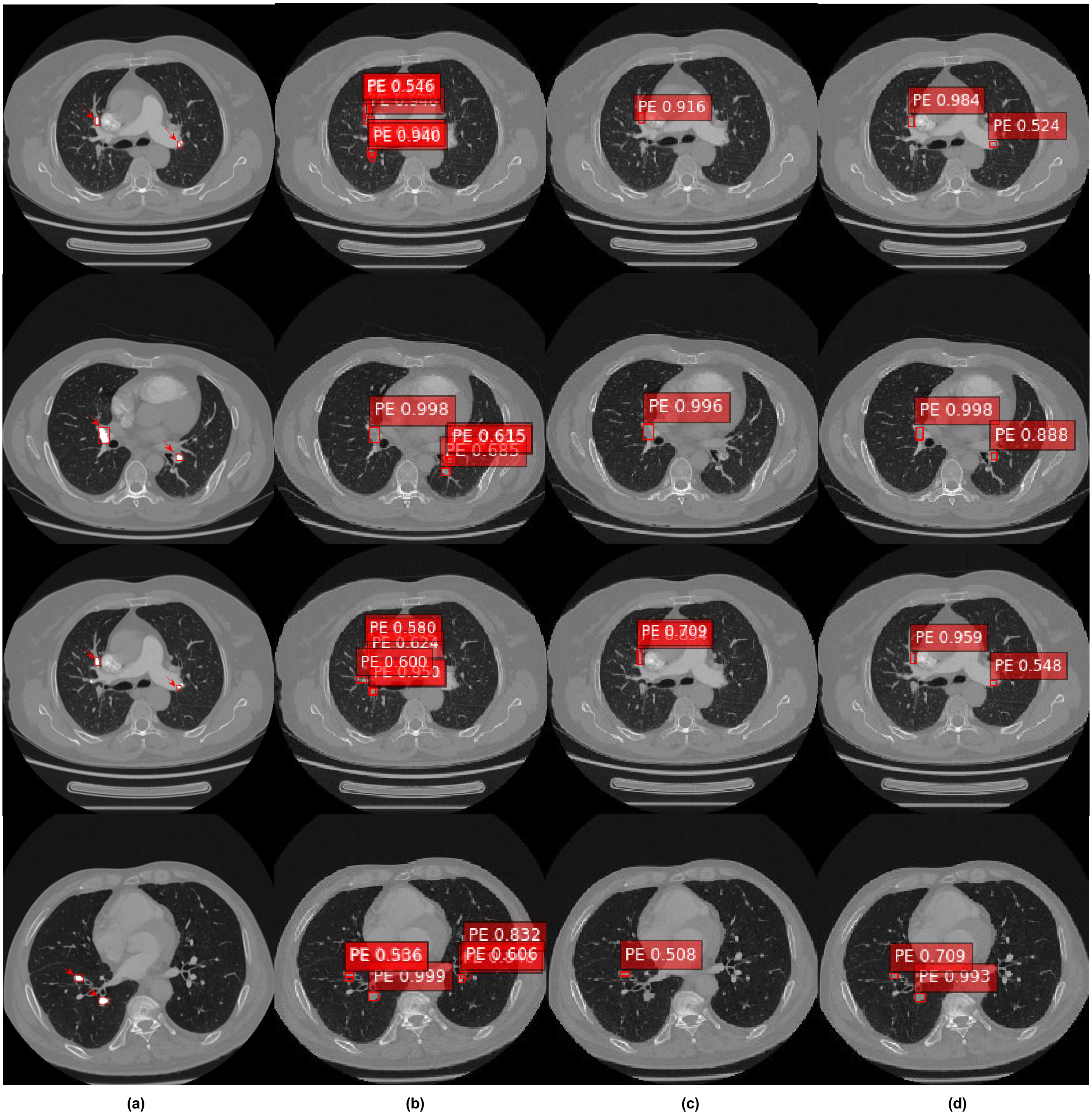
C. PERFORMANCE OF FEATURE EXTRACTORS

This part verifies the effectiveness of feature extractors discussed in Section II-A. Different backbone networks including ResNet-50, SE-ResNet-50, finetuned SE-ResNet-50 and DenseNet-201 [24] were tested. The precision of the model before and after adding FPN and MF-FPN was compared to evaluate the influence of FPN and MF-FPN on accuracy more comprehensively, as shown in Table 2.

As the experimental results show, the accuracy of finetuned SE-ResNet-50 is improved by more than 0.8% compared to ResNet-50, and is similar to that of DenseNet-201. The channel attention mechanism and dense connection are proved to be helpful to improve the detection accuracy of the whole model. But the speed of DenseNet-201 decreases as the complexity of the model increases. Considering the detection speed and accuracy comprehensively, finetuned SE-ResNet-50 was selected as the backbone network of the model for the following experiments.

The AP of the whole model is improved by 6.2% after adding FPN, which proves the importance of multi-level feature fusion. Using the MF-FPN, the AP of the model is improved by more than 1.0% compared with the FPN. It is proved that adding bottom-up path on the basis of adding top-down path and strengthening the transmission of low-level location information can improve the precision.

Four groups of visualization results were selected as shown in Fig. 11. From (a) to (d) are the pulmonary embolism marked by radiologists, the detection results of Faster R-CNN based on finetuned SE-ResNet-50, finetuned SE-ResNet-50 + FPN, finetuned SE-ResNet-50 + MF-FPN. The pixel-level feature of pulmonary embolism is a relatively dark block located in the bright pulmonary artery and the detected pulmonary embolisms are marked by the bounding boxes with confidence score. As the visualization results show, the detection result of Faster R-CNN with finetuned SE-ResNet-50 has many problems such as false detection, high false positive (FP). Some false detection objects are eliminated using the model adding FPN, but only part of the embolisms has been detected, there are still problems of missed detections and high false negative (FN). The test results of Faster R-CNN based on finetuned SE-ResNet-50+MF-FPN have corrected many missed detections and are the same as the actual marked pulmonary embolism even for small embolism, which are more accurate.



(a) (b) (c) (d)

FIGURE 11. Visualization of detection results. (a): The pulmonary embolism marked by radiologists. (b)~(d): The detection results of pulmonary embolism using the model based on finetuned SE-ResNet-50, finetuned SE-ResNet-50 + FPN and finetuned SE-ResNet-50 + MF-FPN.

D. THE COMBINATION STRATEGY OF MF-FPN

The idea of this experiment is that the original feature layers {C2, C3, C4, C5} retain the most original feature information of the current level without multiple down-sampling and fusion processing, and mapping the original feature layer to the output layer can make up for the information loss caused by multiple up-sampling and down-sampling operations. The hierarchy selection of the original feature layer is a factor affecting the performance of the model. This part discusses the precision of the strategy without combining the original feature layer, combining the original feature layer of the current level and combining the original feature layer of the adjacent lower level. As a comparative experiment, Strategy1 does not combine the original feature layer. Strategy2 and Strategy3 respectively represent mapping the original feature layer of the current level and the original feature layer of the adjacent lower level to generate the {P2, P3, P4, P5, P6} of MF-FPN, in addition

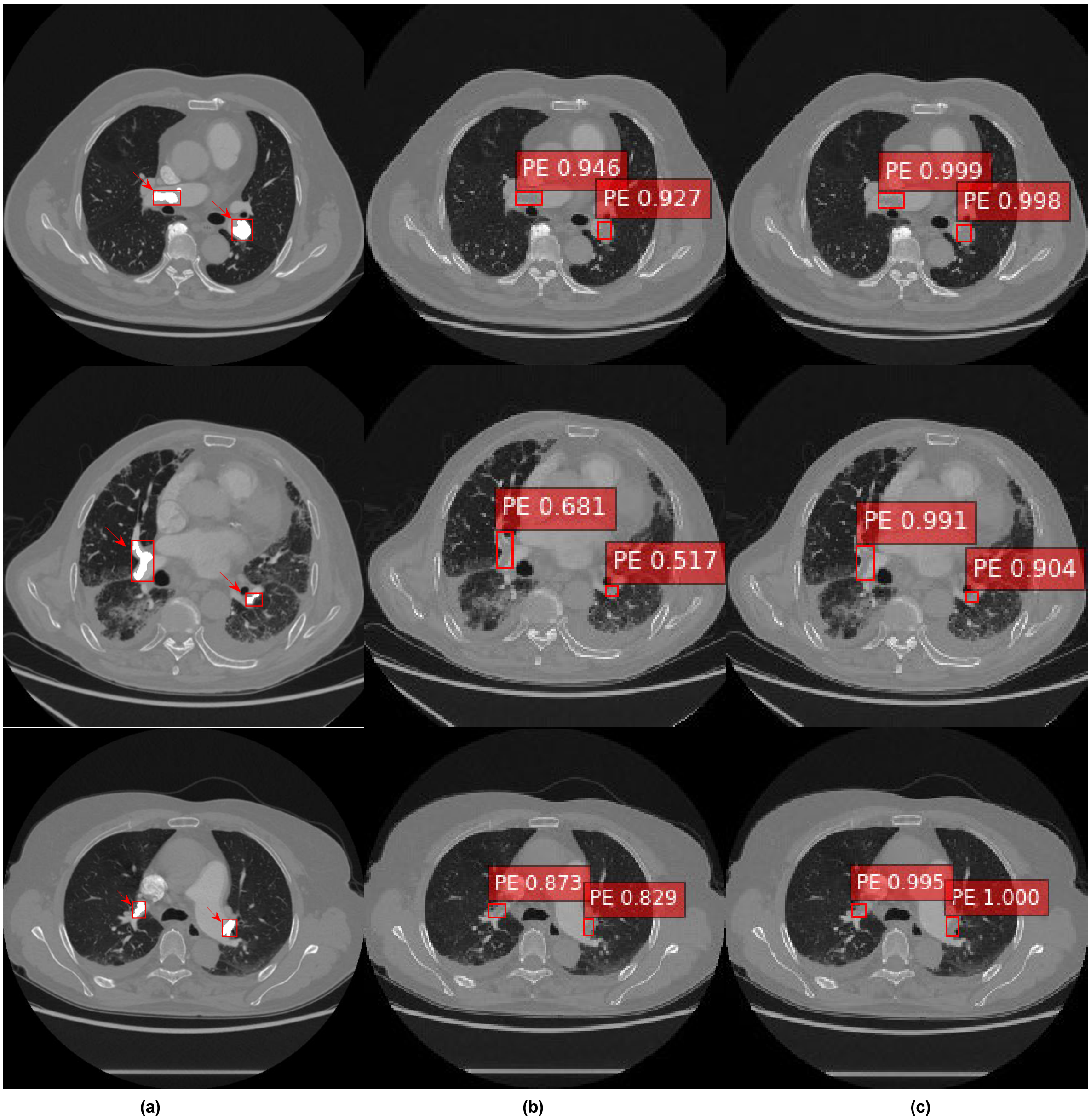


FIGURE 12. Visualization of detection results. (a): The pulmonary embolism marked by radiologists. (b): The detection results of Faster R-CNN based on finetuned SE-ResNet-50 + MF-FPN. (c): The detection results of MA Faster R-CNN.

to adding {P2’, P3’, P4’, P5’} generated from the bottom-up path and lateral connection {C2’, C3’, C4’, C5’, C6’}. For example, P3 in Strategy2 is generated by the fusion of C3’, P2’ and the current layer C3, P3 in Strategy3 is generated by the fusion of C3’, P2’ and the adjacent lower layer C2, and so on. The accuracy of different combination strategies is shown in Table 3.

TABLE 3. The average precision of various combination strategy for generating P2, P3, P4, P5, P6.

Output layer	Strategy1	Strategy2	Strategy3
P2	C2’	C2’+C2	C2’
P3	C3’+P2’	C3’+P2’+C3	C3’+P2’+C2
P4	C4’+P3’	C4’+P3’+C4	C4’+P3’+C3
P5	C5’+P4’	C5’+P4’+C5	C5’+P4’+C4
P6	C6’+P5’	C6’+P5’	C6’+P5’+C5
AP(%)	84.26	84.35	**85.07**

As the experimental results show that the strategy of mapping the original feature layer to the output layer can improve the AP of detection. The detection accuracy of the method combining the original feature layer of the adjacent layer is improved by 0.7% compared with the method combining the original feature layer of the current layer. This is due to the low-level features are lack of abundant semantic information but contain obvious location information, for the single kind of small object to be detected in this subject, the positioning information acquisition ability of the whole model is of great importance. The method of mapping the

TABLE 4. The performance of various FM and PM.

Method	AP(%)	Sensitivity (%)	Specificity (%)
FM(sum)+PM(a)	85.07	86.36	85.32
FM(product)+PM(a)	84.91	85.87	85.74
FM(sum)+PM(b)	85.15	86.54	85.92
FM(sum)+PM(c)	85.21	86.41	86.04
FM(sum)+PM(d)	**85.88**	**87.30**	86.29
FM(sum)+PM(e)	85.46	86.85	86.31

original feature layer of adjacent lower layer to the output layer achieves a balance between semantic information and location information which is suitable for this subject.

E. PERFORMANCE OF DIFFERENT FUSION MODULE (FM) AND PREDICTION MODULE (PM)

In order to understand the influence of the pooling fusion mode and the prediction module discussed in Section II-B more effectively, the comparative experiments on the structure of different settings are performed in this part. Finetuned SE-ResNet-50 + MF-FPN, the backbone network with the best effect in the previous experiment, is selected as the feature extractor. Two pooling fusion methods, elementwise-sum and elementwise-product are combined with the original PM(a) and the improved PM(b), PM(c), PM(d) and PM(e) respectively, and the experimental results are shown in Table 4.

As shown in Table 4, elementwise-sum fusion method shows the better detection accuracy based on PM(a). The comparative experiments of PM(b), PM(c), PM(d) and PM(e) are performed based on the elementwise-sum fusion method. The result shows that improving the prediction module of the model can improve the accuracy, and the residual structure of 3 × 3 convolution kernel in PM(d) is the best, the accuracy does not continue to improve when the residual structure of 3 × 3 convolution kernel is continued to be added. The method of adding residual structure in the prediction module can increase the depth of detection module and avoid the problem of gradient disappearance. Compared with the original prediction module, the residual prediction module of adding the residual structure of 3 × 3 convolution kernel improves the detection accuracy of the whole model by 0.8%.

Fig. 12 shows the pulmonary embolism marked by radiologists, the prediction results of Faster R-CNN based on finetuned SE-ResNet-50+MF-FPN and the results of MA Faster R-CNN. Although both of Faster R-CNN based on MF-FPN and MA Faster R-CNN which combines MF-FPN and RPM detect pulmonary embolism accurately, the confidence scores of prediction results produced by MA Faster R-CNN are higher than that of Faster R-CNN based on MF-FPN, MA Faster R-CNN has more reliable prediction results.

F. COMPARISON WITH OTHER FRAMEWORK

Table 5 shows the comparison of pulmonary embolism detection results by using other models and the proposed MA Faster R-CNN. The results show that MA Faster R-CNN has the higher AP, which is superior to the current mainstream object detection model. Compared with SSD [10], Mask R-CNN [7], RFCN [8] and original Faster R-CNN [5], the AP of MA Faster R-CNN is increased by 7.65%, 4.85%, 5.98% and 8.79%, MA Faster R-CNN also has a good effect in sensitivity and specificity.

TABLE 5. The comparison of AP, sensitivity and specificity with other models.

Model	AP (%)	Sensitivity (%)	Specificity (%)
SSD [10]	78.23	81.42	79.62
Mask R-CNN [7]	81.03	84.84	81.47
RFCN [8]	79.90	83.13	81.64
Original Faster R-CNN [5]	77.09	82.25	80.52
MA Faster R-CNN	**85.88**	**87.30**	**86.29**

IV. CONCLUSION

In this study, the architecture of Faster R-CNN is systematically studied. In order to improve the performance of detecting small objects, the architecture of FPN was improved and a new MF-FPN was proposed. The bottom-up fusion and nearest layer feature fusion were added to improve the detection precision. And RPM is introduced into the prediction module to increase the depth of the classifier and enhance the confidence scores of prediction results. The proposed MA Faster R-CNN which combines MF-FPN and RPM can not only be used for the detection of pulmonary embolism, but also has reference value for detecting small objects in other fields. It improves the problems of poor detection efficiency, missed detection and false detection of pulmonary embolism effectively, which can save a large part of manpower and assist doctors in diagnosis. In the following research, the location information of pulmonary embolism detected will be used to determine the pulmonary artery branch where pulmonary embolism is located, and the risk assessment of pulmonary embolism will be performed.

REFERENCES

[1] X. Yang, Y. Lin, J. Su, X. Wang, X. Li, J. Lin, and K.-T. Cheng, "A two-stage convolutional neural network for pulmonary embolism detection from CTPA images," *IEEE Access*, vol. 7, pp. 84849–84857, 2019, doi: 10.1109/ACCESS.2019.2925210.

[2] M. L. Domingo, L. Martí-Bonmatí, R. Dosdá, and Y. Pallardó, "Interobserver agreement in the diagnosis of pulmonary embolism with helical CT," *Eur. J. Radiol.*, vol. 34, no. 2, pp. 136–140, May 2000.

[3] J. Liang and J. Bi, "Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.* Kerkrade, The Netherlands: Springer-Verlag, 2007, pp. 630–641.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[6] C. Eggert, S. Brehm, A. Winschel, D. Zecha, and R. Lienhart, "A closer look: Small object detection in faster R-CNN," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 421–426, doi: 10.1109/ICME.2017.8019550.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[8] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2016.

[9] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[11] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 354–370.

[12] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016, doi: 10.1109/TMI.2016.2536809.

[13] H. Li, Y. Huang, and Z. Zhang, "An improved faster R-CNN for same object retrieval," *IEEE Access*, vol. 5, pp. 13665–13676, 2017, doi: 10.1109/ACCESS.2017.2729943.

[14] A. Yang, X. Jin, and L. Li, "CT images recognition of pulmonary tuberculosis based on improved faster RCNN and U-Net," in *Proc. 10th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Aug. 2019, pp. 93–97, doi: 10.1109/ITME.2019.00032.

[15] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[16] 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, doi: 10.1109/CVPR.2016.90.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze- and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[18] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[20] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1476–1481, Jul. 2017, doi: 10.1109/TPAMI.2016.2601099.

[21] H. Yuan, Z. Liu, Y. Shao, and M. Liu, "ResD-Unet research and application for pulmonary artery segmentation," *IEEE Access*, vol. 9, pp. 67504–67511, 2021, doi: 10.1109/ACCESS.2021.3073051.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[23] T. Alaa, "Classification assessment methods: A detailed tutorial," *Appl. Comput. Inform.*, p. S2210-8327(18)30154-6, Sep. 2018.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[25] J. Li, H. Wang, and L. Song, "A novel sparse feature extraction method based on sparse signal via dual-channel self-adaptive TQWT," *Chin. J. Aeronaut.*, vol. 34, no. 4, pp. 157–169, Jul. 2020, doi: 10.1016/j.cja.2020.06.013.

[26] H. Xue, D. Ding, Z. Zhang, M. Wu, and H. Wang, "A fuzzy system of operation safety assessment using multi-model linkage and multi-stage collaboration for in-wheel motor," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 15, 2021, doi: 10.1109/TFUZZ.2021.3052092.

HONGFANG YUAN received the Ph.D. degree in mechatronics from Beijing Institute of Technology, Beijing, China, in 2001. She is currently an Associate Professor with Beijing University of Chemical Technology, with a focus on intelligent diagnosis of equipment failure, data processing, and information fusion technology.

YAJUN SHAO received the B.E. degree in electronic information engineering from Beijing University of Chemical Technology, Beijing, China, in 2019, where she is currently pursuing the M.E. degree in information and communication engineering. Her research interests include medical image identification and computer-aided diagnosis.

ZHENHONG LIU received the B.E. degree in communication engineering from Beijing University of Chemical Technology, Beijing, China, in 2018, where she is currently pursuing the M.E. degree in control engineering. Her research interests include medical image processing and computer-aided diagnosis.

HUAQING WANG received the Ph.D. degree from Mie University, Japan, in 2009. He is currently a Professor with Beijing University of Chemical Technology. His research interests include intelligent diagnosis of equipment failure, machinery dynamics, and information fusion technology.