Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(2): 1730–1749.
DOI: 10.3934/mbe.2023079
Received: 03 July 2022
Revised: 17 September 2022
Accepted: 19 October 2022
Published: 04 November 2022

Research article

Lesion detection of chest X-Ray based on scalable attention residual CNN

Cong Lin[1,3], Yiquan Huang[3], Wenling Wang[1], Siling Feng[1,*]and Mengxing Huang[1,2*]

[1] College of Information and Communication Engineering, Hainan University, Haikou 570228, China

[2] State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, Haikou 570228, China

[3] College of Electronic and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China

* **Correspondence:** Email: fengsiling2020@163.com, huangmx09@163.com.

Abstract: Most of the research on disease recognition in chest X-rays is limited to segmentation and classification, but the problem of inaccurate recognition in edges and small parts makes doctors spend more time making judgments. In this paper, we propose a lesion detection method based on a scalable attention residual CNN (SAR-CNN), which uses target detection to identify and locate diseases in chest X-rays and greatly improves work efficiency. We designed a multi-convolution feature fusion block (MFFB), tree-structured aggregation module (TSAM), and scalable channel and spatial attention (SCSA), which can effectively alleviate the difficulties in chest X-ray recognition caused by single resolution, weak communication of features of different layers, and lack of attention fusion, respectively. These three modules are embeddable and can be easily combined with other networks. Through a large number of experiments on the largest public lung chest radiograph detection dataset, VinDr-CXR, the mean average precision (mAP) of the proposed method was improved from 12.83% to 15.75% in the case of the PASCAL VOC 2010 standard, with IoU >0.4, which exceeds the existing mainstream deep learning model. In addition, the proposed model has a lower complexity and faster reasoning speed, which is conducive to the implementation of computer-aided systems and provides referential solutions for relevant communities.

Keywords: chest X-ray; object detection; deep learning; attention mechanism; disease recognition

1. Introduction

The COVID-19 outbreak, which began in 2019, is a viral disease caused by severe acute respiratory syndrome coronavirus type 2 (SARS-COV-2) [1–4]. Most COVID-19 patients have pneumonia, and computed tomography (CT) scans are often used to help doctors diagnose pneumonia in the early stages of COVID-19 outbreaks [5–7]. Compared with CT, chest X-ray (CXR) is more widely used in clinical practice because it is easier, faster, and less expensive to perform. However, the sheer volume of CXR data and limited number of physicians cannot ensure that the system operates with maximum efficiency to save more patients [8–12]. A computer-aided system can play a certain auxiliary role [13,14], but its efficiency and accuracy cannot meet the requirements. Improving the accuracy of CXR image lesion identification is still a key issue that urgently needs to be solved.

Traditional methods usually use the mathematical calculation of regions and feature extraction to recognize and classify CXR images. Jaeger et al. [15] proposed an automated method for tuberculosis detection on posterior-anterior chest radiographs. Lung segmentation was modeled as an optimization problem, integrating lung boundaries, regions, shapes, and other attributes with tight segmentation contours and leakage in some areas. Hogeweg et al. [16] combined a texture anomaly detection system running at the pixel level with the clavicle detection system to suppress false-positive reactions, and the pathological structure changed after segmentation, which was detrimental to the judgment of pathology. Candemir et al. [17] proposed a robust lung segmentation method driven by nonrigid registration using a patient-specific adaptive lung model based on image retrieval to detect lung boundaries, achieving an average accuracy of 95.4% on the public JSRT database. However, opacity caused by fluid in the lung space prevents correct detection of lung boundaries. Although regional segmentation has been valued, there is a lack of corresponding supervision mechanisms. To train classifiers that can effectively monitor, Livieris et al. [18] proposed an SSL algorithm for tuberculosis CXR classification, which combines the individual predictions of three commonly used SSL algorithms applying the CST-voting integration principle and voting method. Statistical accuracy was relatively objective, but the process was too tedious. Faced with the problems of complex mathematical principles and low model robustness existing in traditional methods, more cost-effective methods are required in the CXR recognition field.

In recent years, deep learning models such as convolutional neural networks (CNN) [19–26] have been rapidly developed, and they have become the preferred technical means in the field of computer vision. Experts in the medical imaging field have also noted the rapid growth and impact of CNNs. For example, Irfan et al. [27] developed a hybrid deep neural network (HDNN) that uses computed tomography (CT) and X-ray imaging to predict risks. The classification accuracy reached a very high level by training with a dataset on the web along with a regular dataset. The CoVIRNet method proposed by Almalki et al. [28] can automatically diagnose COVID-19 patient images using chest radiographs and alleviate the overfitting problem owing to the small size of the COVID-19 dataset. Most CXR disease recognition methods based on deep-learning technology can be divided into two types. The first type uses a CNN for image segmentation and classification. Shen et al. [26] extracted the symptom part of an image as a block and inputted several different CNNs, and the features obtained were spliced into vectors as the final result. Discriminant features were extracted from alternately stacked layers to capture the heterogeneity of pulmonary nodules; however, the location of the disease could not be directly represented. Rajpurkar et al. [29] developed a 121-layer CNN named CheXNet, which was tested on the ChestX-Ray14 large pneumonia data set containing 14 diseases and achieved an accuracy of more than 0.7 in the classification of diseases. However, the network stacking method is excessively simple, and the image texture is not used more thoroughly. The second type of methods denoise the data to enhance the recognition effect of the other algorithms. Ucar and Korkmaz [30] proposed an architecture based on SqueezeNet to fine-tune COVID-19 diagnosis through raw data enhancement, Bayes optimization, and validation to achieve high accuracy in categorizing COVID-19, pneumonia, and normality. Jiang et al. [31] proposed a residual CNN for denoising COVID-19 images. The residual connection and attention mechanism were used to make the network pay more attention to the texture details of the CXR images, and the effect of the denoised images was significantly improved in the COVID-19 recognition task.

Waheed et al. [32] used an adversarial network model to synthesize CXR images, which allowed the model to rely on external information to improve the sample quality, thus increasing the number of images of COVID-19 symptoms. However, the recognition task was limited to classification, and the location of the symptoms was not accurately obtained. In a more specialized work, Jaiswal et al. [33] applied the target detection algorithm to the RSNA pneumonia data set, and obtained an optimal accuracy by fusing the prediction boundary boxes of multiple models. Because of its large memory occupation and use of a single category of dataset, the detection results of diversified diseases cannot be determined. This provides the motivation and references for our work.

Although the accuracy of the above methods continues to improve, there are still some obvious shortcomings: 1) most of them are classification and segmentation methods, and lack intuitive target boxes to directly indicate the location of symptoms, so the evaluation efficiency needs to be improved. 2) The ideal accuracy usually requires the fusion of the detection effects of multiple models, which occupies a large memory and is not realistic in practical applications. 3) The CXR data volume of COVID-19 is small, and it is a single category, so generalization results cannot be obtained. To solve the above problems, a lesion detection method based on a scalable attention residual CNN is proposed in this paper. A variety of convolution kernel sizes are used to obtain a variety of resolutions, and adaptive global attention is used to extract the spatial features of each resolution and connect them. A feature fusion method with different tree-structure depths is designed. Finally, the attention mechanism is used to fuse the spatial and channel information. All the effects were tested in a single model using the VinDr-CXR dataset. The main contributions of this study are summarized as follows:

- To improve the ability of the deep learning model in CXR target detection, we propose a CNN model and construct a structure based on the model that can effectively improve the accuracy of CXR location detection and improve the sensitivity of the model to CXR features from the aspects of attention and feature fusion.
- We use zero-based training to customize our CNN model for the data set, and the training effect of a single model can outperform most of the classical deep learning models. The development of CXR data sets containing target location information became the focus, proving the necessity and advance of our work and providing the corresponding references for future work.
- The three modules designed in this study, multi-convolution feature fusion block (MFFB), tree-structured aggregation module (TSAM), and scalable channel and spatial attention (SCSA), can effectively improve the detection effect of the deep learning model in CXR, and the accuracy increases from 12.83 to 15.75% after the addition of the modules, higher than that obtained by the existing mainstream target detection model.

2. Materials and methods

We summarized our research into three parts: encoder, multiple feature blocks, and decoder, simplifying the complex model structure. The encoder in Figure 1 represents the scalable attention residual CNN (SAR-CNN), whereas the decoder is a simple convolutional layer. As many as four feature blocks, from Features 1–4, are input into the decoder in the form of pyramids, and finally, the positioning of the detection frame and judgment of the lesion category are performed. In the following sections, we introduce the proposed CXR target detection network, SAR-CNN, in detail, including the MFFB, TSAM and SCSA modules. These modules are independent and embeddable and can be migrated to common convolutional networks. The MFFB is designed to interpret CXR image information from a multi-resolution perspective, whereas the TSAM utilizes the characteristics of the tree structure to perform left-right branch and multi-level feature fusion, and finally the SCSA is used for spatial and channel attention integration. Residual modules [34] are used in the rest of the network to improve its learnability. Details of the SAR-CNN training are covered in the next section.

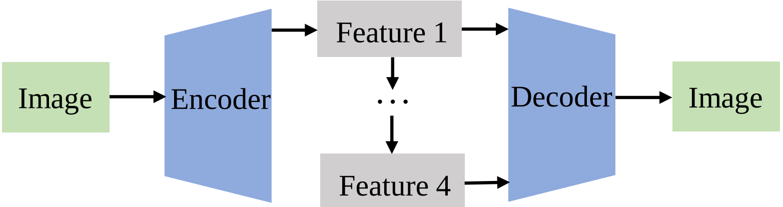


Figure 1. Diagram of the proposed research.

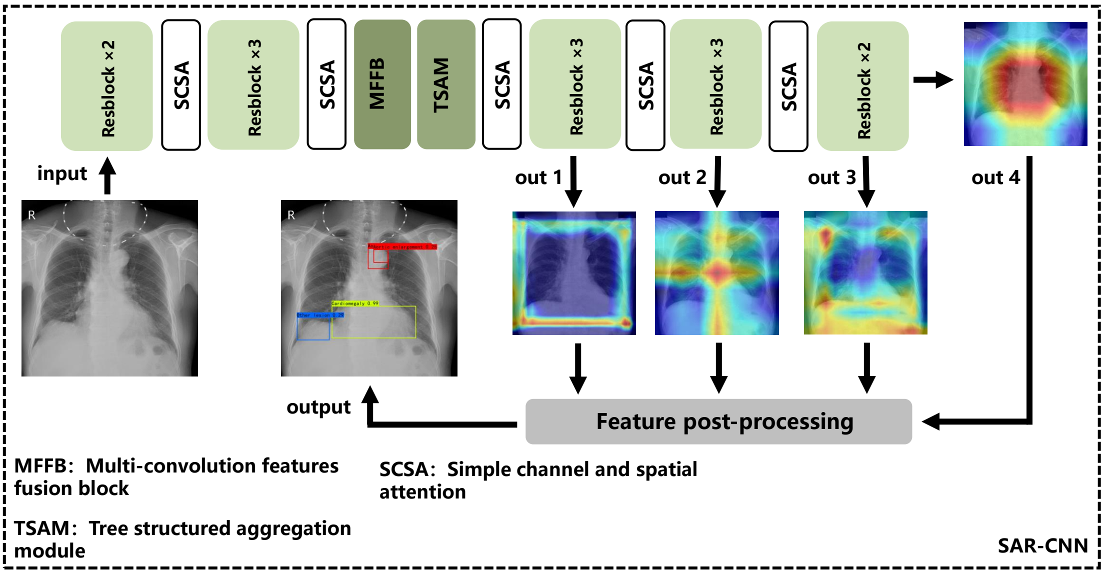


Figure 2. Architecture structure of scalable attention residual CNN (SAR-CNN).

2.1. Network architecture

The proposed SAR-CNN network structure composed of MFFB, TSAM and SCSA is shown in Figure 2. The VGG [35] straight-cylinder module stacking method was adopted to improve the embeddability of modules. The predictive image (input) was input to the CNN in 512 × 512 size. Four feature graphs are output and sent for feature post-processing through the modules mentioned in the following sections. It can be observed from the CAM graph that the four feature maps have different degrees of performance in characterization feature fitting. The neck and head parts in RefineDet [36] have both first-stage and second-stage advantages, so we replace the backbone network of RefineDet with SAR-CNN and finally obtain the final detection result (output) of our predicted image. Such a framework will not only help improve the localization performance, but also, by means of attention, provide a way to explain visually the model decisions, both of which are important for the clinical deployment of deep learning models.

2.2. Multi-convolution features fusion block

As shown in Figure 3, we propose an MFBB that is different from the simple connection and fusion of existing algorithms. M_1, M_2 and M_3 were obtained by 3 × 3, 5 × 5 and 7 × 7 convolution kernel extractions of $M_{in} \in R^{\mathrm{C\times H\times W}}$, where $M_1 \in R^{\mathrm{C\times H\times W}}$ was extracted using the same mode convolution. We used ECA-Net [37], which is a simple and effective attention processing method. The difference is that we used the convolution of more receptive fields to extract richer feature scales and used global average pooling (GAP) to extract features along channel dimensions from M_2 and M_3 with different resolutions. Next, 1D convolution was used for adaptive feature extraction. After the sigmoid, channel attention $S_2 \in R^{1\times1\times\mathrm{C}}$ and $S_3 \in R^{1\times1\times\mathrm{C}}$ were obtained, and the fusion function F(.) defined in Formula (2.1) is adopted. The final output feature of the module $M_{out} \in R^{\mathrm{C\times H\times W}}$ was obtained.

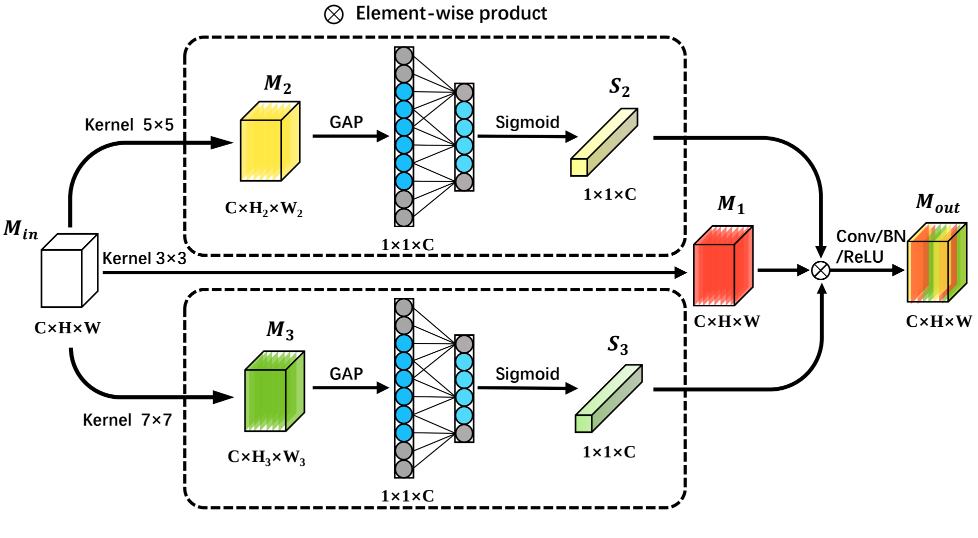


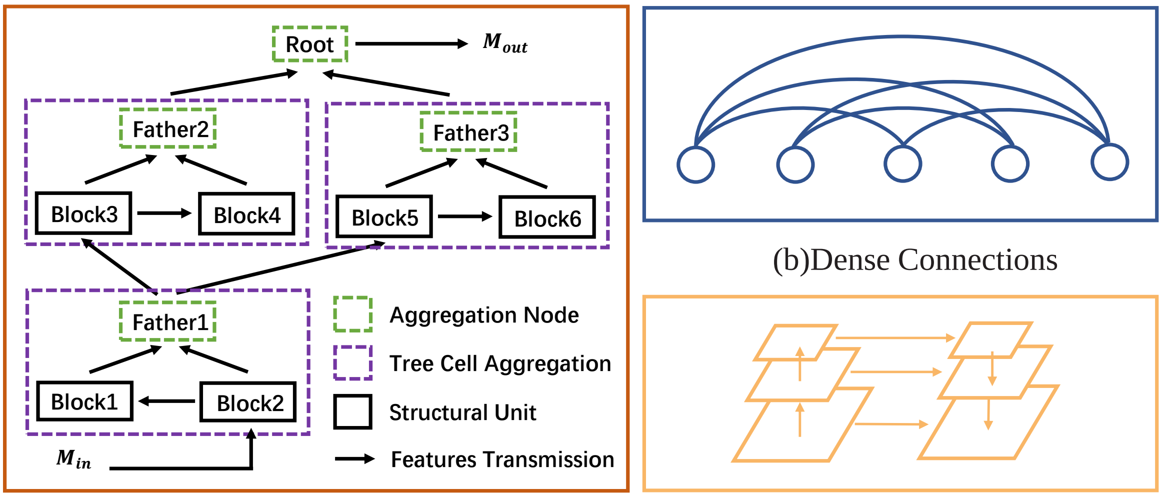
Figure 3. Structure of the multi-convolution feature fusion block.

$$\begin{aligned} M_{out} &= F\left(M_1 \otimes S_2 \otimes S_3\right) \\ &= \sigma\Big(\text{BatchNorm}\,\Big(\text{Conv}^{\,3\times3}\left(M_1 \otimes S_2 \otimes S_3\right)\Big)\Big), \end{aligned} \tag{2.1}$$

where F(.) represents a combination of three layers: the same mode 3 × 3 convolution layer, the BatchNorm layer [38], and the nonlinear activation function ReLU [39]. ⊗ is the element-wise multiplication, σ represents the ReLU layer, and $\text{Conv}^{3\times3}$ is the same-mode convolution layer of 3 × 3 size. The convolution layer is used to fuse feature graphs and channel attention of two different resolutions to avoid the feature mismatch of related images caused by simple multiplication. BatchNorm is a normalization method that solves the phenomenon of inconsistent distribution of input data, highlights the relative difference in distribution between them, and speeds up training. The ReLU layer adds a nonlinear relationship to the feature layer to avoid gradient disappearance and over-fitting, which ensures that our neural network can complete complex tasks.

2.3. Tree structured aggregation module

We designed the TSAM by summarizing the feature fusion methods of DenseNet [40] and FPN [41] and drawing inspiration from the DLA structure [42], as shown in Figure 4.



(b)Dense Connections

(a)Tree structured aggregation module

(c)Feature Pyramids

Figure 4. The internal relationship between TSAM, dense connections and feature pyramids: (a) structure of the TSAM. (b) Dense connections. (c) feature pyramids. The proposed module (a) has the advantages of both (b) and (c), but avoids the problems of (b) over-intensive fusion resulting in excessive use of memory and (c) over-simple fusion.

In contrast to DLA, we adopt a fixed number of layers and use feature layers with more details to avoid the overfitting problems caused by excessively deep iterative networks. Each structural unit corresponds to a residual module. For simplicity and resource saving, this residual module contains only two 3 × 3 convolution and BatchNorm layers. Each aggregation node adopts a 3 × 3 convolution, BatchNorm layer, and ReLU, and the features of the left and right branches are fused to obtain the feature graph:

$$\begin{aligned} A(x_1, x_2) &= \sigma\left(BatchNorm\left(W_1 x_1 + W_2 x_2 + b\right)\right) \\ &= \sigma\left(BatchNorm\left(W_1 x_1 + W_2\left(W_0 x_1 + b_0\right) + b\right)\right), \end{aligned} \tag{2.2}$$

where x_1 and x_2 represent the left- and right-branch features of the binary tree before fusion, respectively. σ represents a nonlinear ReLU. W and b respectively represent the weights and offsets of convolution, and x_2 is obtained by x_1 through a structural unit. The TSAM combines layers of different depths to learn richer combinations that span more feature layers.

2.4. Scalable channel and spatial attention (SCSA)

This module utilizes both the channel and spatial dimensions of attention, and we use SCSA as a transitional stage between the two modules, as shown in Figure 5. In the channel dimension module, M_{input} is input into two paths to obtain $M_S \in R^{1\times H\times W}$ and $M_C \in R^{C\times 1\times 1}$, respectively, and $M_{SCSA} \in R^{C\times H\times W}$ is obtained by combining them. Then, residual fusion is performed between $M_{input} \in R^{C\times H\times W}$ M_{SCSA}, respectively.

$$\begin{aligned} M_{\text{output}} &= M_{\text{input}} + M_{\text{input}} \otimes M_{\text{SCSA}} \\ &= M_{\text{input}} + M_{\text{input}} \otimes \sigma\left(M_S + M_C\right), \end{aligned} \tag{2.3}$$

where σ is the sigmoid function and $\otimes$ is element-wise multiplication. In the channel attention module, the channel information encoded in two different ways is obtained through global average pooling and global maximum pooling. In addition, a full-connection layer is set as Share FC to interact with the information of the two channels, and the feature size obtained is $R^{C/r\times 1\times 1}$. After the BatchNorm layer, the two are added.

$$Mc = BN(\text{ ShareFC (GAP }(M_{\text{input}}))) + BN(\text{ ShareFC (GMP }(M_{\text{input}}))), \tag{2.4}$$

where BN is the batch normalization layer. The spatial-dimension module compresses $M_{input} \in R^{C\times H\times W}$ to $R^{C/r\times H\times W}$ through a layer of 1 × 1 convolution. A layer of 3 × 3 dilated convolution is set to expand the receptive field to utilize more contextual information, and then a layer of the spatial attention map $M_S \in R^{1\times H\times W}$ is obtained through a layer of 1 × 1 convolution. Finally, a batch normalization layer is used to adjust the search space:

$$M_S = BN\left(\text{Conv}^{1\times 1}\left(\text{Conv}^{3\times 3}\left(\text{Conv}^{1\times 1}\left(M_{\text{input}}\right)\right)\right)\right), \tag{2.5}$$

where BN is the batch normalization layer, Conv is the convolution layer, and the superscript is the size of the convolution kernel. In addition, the SCSA module is followed by MaxPooling of a layer with a size of 2 × 2 and a stride of 2 as the lower sampling layer. SCSA has a working principle similar to that of BAM [43], but it uses the different global pooling features of GAP and GMP to extract channel attention and has more types of channel feature maps. Compared to CBAM [44], the original features were added after the fusion of the channel and spatial attention, rather than in sequence. SCSA prefers to combine the advantages of BAM and CBAM, and discard unnecessary parts.

2.5. Other modules

We observed that the residual structure plays a key role in medical image detection tasks. Therefore, in addition to the three modules proposed in this study, residual modules were used in the rest of the network to sort out the features obtained by fitting the above three modules and further improve the robustness of the network. The residual module is derived from ResNet, which adds jump connections to the convolutional module to solve the problems of gradient disappearance and gradient explosion during the training of deep neural networks, as shown in Figure 6. In addition, we observed that the residual module could correlate features of different scales. In special cases, the disease regions in CXR will cross and even overlap, and the use of a residual module can help correlate features of overlapping target regions.

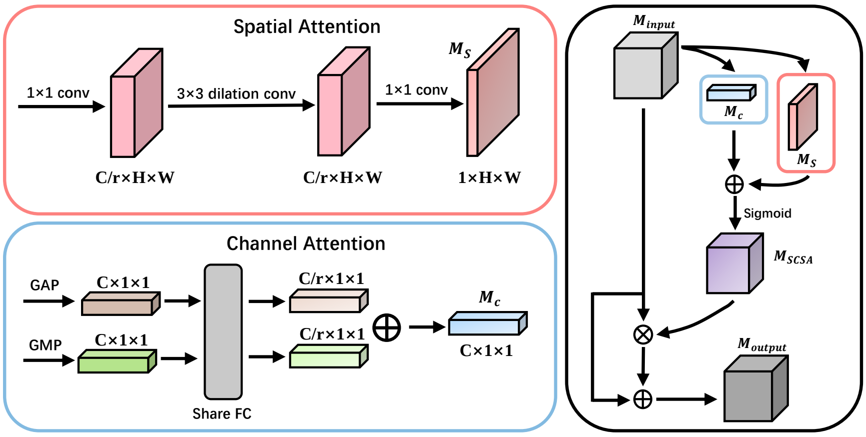


Figure 5. The schematic diagram of SCSA.

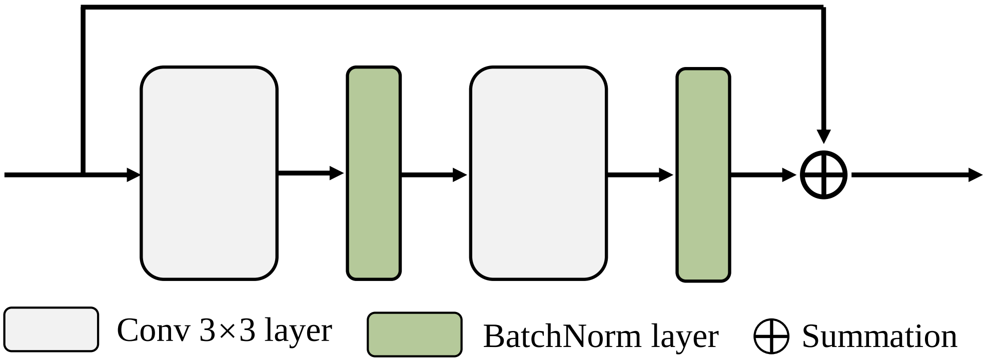


Figure 6. Resblock structure diagram.

3. Experiments and results

3.1. Dataset

As the COVID-19 dataset is not publicly available on a large scale, it is too small and of poor quality, even if it is partially curated and annotated. For the above reasons, we selected the common CXR dataset that meets the requirements, that is, the VinDr-CXR dataset. VinDr-CXR is a chest radiography dataset published by the Vingroup Big Data Institute (VinBigdata) [45] and is currently the largest public CXR dataset with radiologist-generated annotations in both the training and test sets. Collected from the Hospital 108 (H108) and the Hanoi Medical University Hospital (HMUH), 18000 CXR data were manually annotated by 17 professional radiologists. Furthermore, the VinDr-CXR dataset was divided into a 15,000 image training set and a 3000 image test set. The training set was independently annotated by three doctors and the test set was jointly annotated by five doctors. Because of the low number of images in eight of the 22 categories containing local location information, we incorporated these eight minor categories into the other lesions, and thus our task was defined as the target detection problem of 14 lesions. The names and quantity statistics of the different categories in the VinDr-CXR dataset are shown in Figure 7.

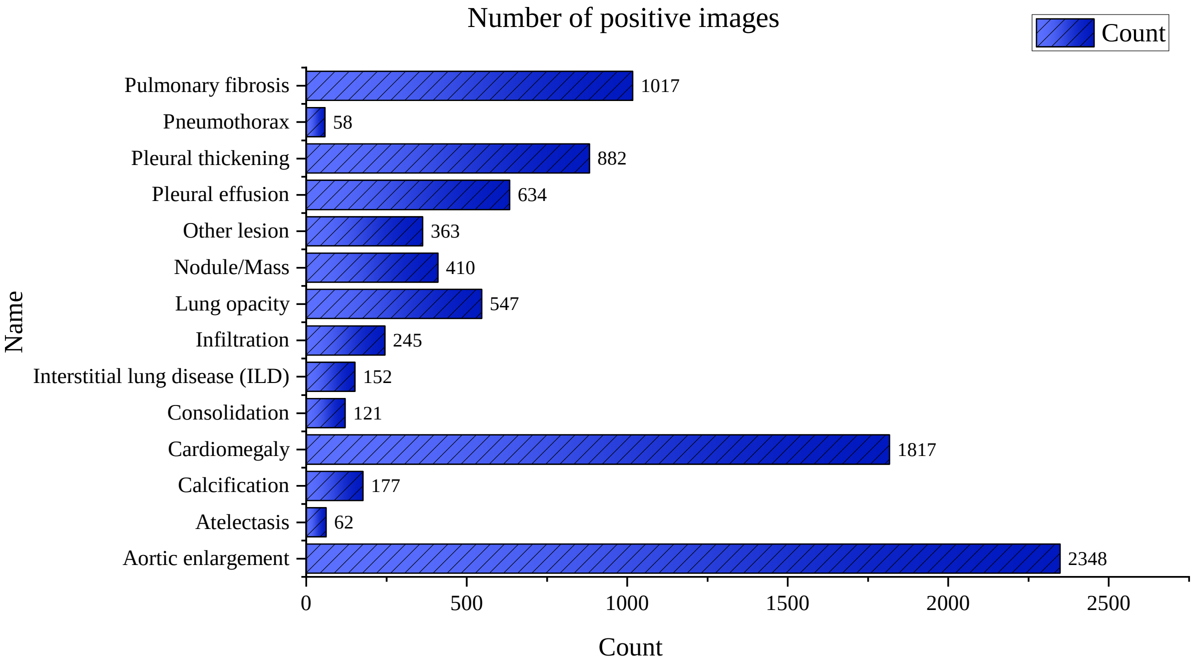


Figure 7. Quantity statistics of different categories in VinDr-CXR dataset.

3.2. Implementation details

The proposed method was trained and tested on the Pytorch framework [46], and the relative hyperparameters of the network were based on ScratchDet [47], with a learning rate of 0.05, using SGD with 0.0005 weight decay and 0.9 momentum. The batch size was set to eight, and the training wheel number adaptive adjustment mechanism was adopted. The experiment was interrupted when the accuracy of more than 150 epochs did not exceed the highest accuracy for five consecutive epochs. The number of decreased rounds was 50, 100 and 150, respectively, and the reduction at each time was 1/10. The resolution of the image input was 512 × 512, and the SSD was used for data enhancement (random expansion, clipping, inversion, random photometric distortion, etc.). All the convolution layers were initialized using the Xavier uniform method. The comparative experimental algorithms were tested using pre-training weights in the Pytorch framework, and the number of training rounds was uniformly set to 300. The first 20 rounds frozen the trunk network. In addition, each round was verified and the weight was saved. Every 50 rounds, a weight was selected for the accuracy test, and the highest accuracy was used as the comparative experimental data.

3.3. Loss function

Network training involves loss functions and optimizers using RefineDet parameter requirements. To imitate the prediction process of the two-stage target detection algorithm, the loss function L_{SAR} consists of L_A and L_B, where L_A corresponds to the stage of the position and size of the target in the returned image and L_B determines the category of the target according to the returned target. N_{ARM} and N_{ODM} in the formula are the numbers of positive anchors in ARM and ODM, respectively. In particular, i is the index of each anchor box and the smooth *L1* loss is used as L_S, and s_i is used to judge whether the predicted category is consistent with the ground truth label; the match is 1; otherwise, it is 0, and the ground truth is represented by $g{*}i$.

In Formula (3.1), p_i and x_i are the probability and corresponding position coordinates of the target in ARM anchor boxes i, respectively, and L_b uses the cross-entropy loss over two classes as a dichotomous loss function.

$$L_A = \frac{1}{N_{ARM}} \left(\sum L_b \left(p_i, S_i\right) + \sum s_i L_s \left(x_i, g^*\right) \right). \tag{3.1}$$

In Formula (3.2), c_i and t_i are the categories of the ODM anchor boxes i and the corresponding coordinates of the bounding box, respectively, whereas l_i is the ground truth class label of the anchor. L_m uses the softmax loss over multiple class confidences as a multiclass classification loss.

$$L_B = \frac{1}{N_{ODM}} \left(\sum L_m \left(c_i, l_i\right) + \sum S_i L_s \left(t_i, g^*{}_i\right) \right). \tag{3.2}$$

The final value of the entire loss function can be obtained by adding the values of the two aforementioned loss functions.

$$L_{SAR} = L_A + L_B. \tag{3.3}$$

3.4. Ablation study

3.4.1. Backbone architecture

To prove the validity of the trunk network we designed, we performed several comparative experiments on the trunk network: 1) VGG-16 in RefineDet source code; 2) on the basis of 1), a BN layer was added to each convolutional layer as a combination; and 3) the original VGG-16 was replaced with other common backbone networks. The experimental results are listed in Table 1. The precision of the trunk network designed by us was higher than that of other experiments. Although the number of parameters was reduced after the replacement of some trunk networks, the corresponding precision was too low and did not have the functional effect required by the task. After the analysis, we believe that although VGG, ResNet, DLA, and other trunk networks are frequently used as a means to improve the task effect, they are essentially designed for classification, which is different from our CXR target detection task, resulting in poor performance. The trunk network we designed increased the number of parameters within the allowed range, thus obtaining a 15.75% mAP, which is higher than the accuracy of the other versions, proving the effectiveness of our network in handling this task.

Table 1. Performance comparison of different backbone networks.

Backbone	mAP@0.5(%)	Inference speed (fps)	Params (M)
VGG-16	12.83	9.96	34.27
VGG-16+BN	13.67	8.50	34.28
ResNet-18	8.05	4.62	22.75
ResNet-34	8.25	3.32	32.85
DLA-60	8.54	7.98	33.90
DLA-102	9.00	5.86	45.30
SAR-CNN (Ours)	**15.75**	**2.60**	**56.48**

3.4.2. Module contribution

According to the modules proposed in Section 2, we conducted the corresponding combined experiments to show the contribution and role of each module to the whole. A tick indicates that the network applies to the module. The experimental results are listed in Table 2. For single-module embedding, two-module combination, or three-module embedding, the accuracy can be further improved. To fit the information of the finishing module, ResBlock was added, and the accuracy of the network was finally improved to 15.75%. We believe that the three specially designed modules improve the robustness of the overall network for the following reasons: First, features of multiple resolutions are conducive to the formation of relatively rich image information, especially for sites with subtle lesions. Second, a simple and effective topological structure is needed to fuse medical image features with insufficient information. Third, medical images require the network to apply attention mechanisms from different angles.

Table 2. Impact of the different components.

Component	SAR-CNN								
MFFB	✓	✓	✓	✓		✓			
TSAM	✓	✓	✓		✓		✓		
SCSA	✓	✓		✓	✓			✓	
ResBlock	✓								
mAP@0.5(%)	**15.75**	**14.20**	13.18	**13.49**	13.15	13.11	13.12	13.08	12.83

3.5. Model comparisons

To prove the superiority of our method, we used the PASCAL VOC 2010 standard [48] and IoU >0.4 (0.5, 0.6, 0.7 and 0.8), and compared the mAP index with the mainstream target detection model. The experimental results are presented in Table 3. In CenterNet, owing to the special setting of non-maximum suppression (NMS), the value of IoU does not affect the accuracy of the algorithm; therefore, it is uniformly 0.5. It can be observed that the performance of mainstream target detection algorithms on CXR image datasets is not significant, with most attaining approximately 11% and EfficientDet even less than 8%. The special design of RefineDet allows it to perform better than most models, 12.83%. Yolov3 also shows excellent detection ability in this task, but it is still lower than that of our algorithm under different IoU standards. The accuracy of SAR-CNN is improved by 2.92% compared with the RefineDet benchmark, which exceeds most mainstream algorithms and is crucial for assisting physicians in detection, proving the effectiveness of our module in the field of medical image target detection.

Table 3. Detection results of different methods on the VinDr-CXR test set.

Methods	Backbone	mAP@0.5 (%)	mAP@0.6 (%)	mAP@0.7 (%)	mAP@0.8 (%)
SSD	VGG-16	10.01	9.47	8.80	7.66
Faster RCNN	VGG-16	11.27	10.34	8.74	6.49
	ResNet-50	10.47	9.58	8.10	5.93
EfficientDet	EfficientNet-b5	7.37	7.08	6.57	5.96
RetinaNet	ResNet-50	11.24	10.55	9.28	7.16
Yolov3	DarkNet-53	15.21	14.56	13.17	11.25
CenterNet	Hourglass	11.76	–	–	–
	ResNet-50	11.38	–	–	–
RefineDet	VGG-16	12.83	12.33	10.66	7.54
Ours	SAR-CNN	**15.75**	**15.34**	**14.57**	**11.87**

3.6. Other comparisons

3.6.1. Resolution differences

As can be observed in Table 4, the SAR-CNN maintains an accuracy of more than 10% for each size. In images with a resolution of 320, the SAR-CNN is slightly less accurate than the benchmark algorithm, but the accuracy increases from 448. We believe this is because images with a resolution that is too low struggle to provide sufficient lesion features for network fitting, which results in weak performance. The benefit of our approach progressively became apparent as the resolution increased, peaking at 16.25% when the image resolution was 768 × 768.

Table 4. Comparison of algorithm accuracy for different resolution images.

Resolution	Base (RefineDet)	SAR-CNN (Ours)
320 × 320	12.02%	11.54%
448 × 448	12.59%	15.12%
512 × 512	12.83%	15.75%
640 × 640	13.87%	15.32%
768 × 768	13.59%	16.25%

3.6.2. Effect of separate categories

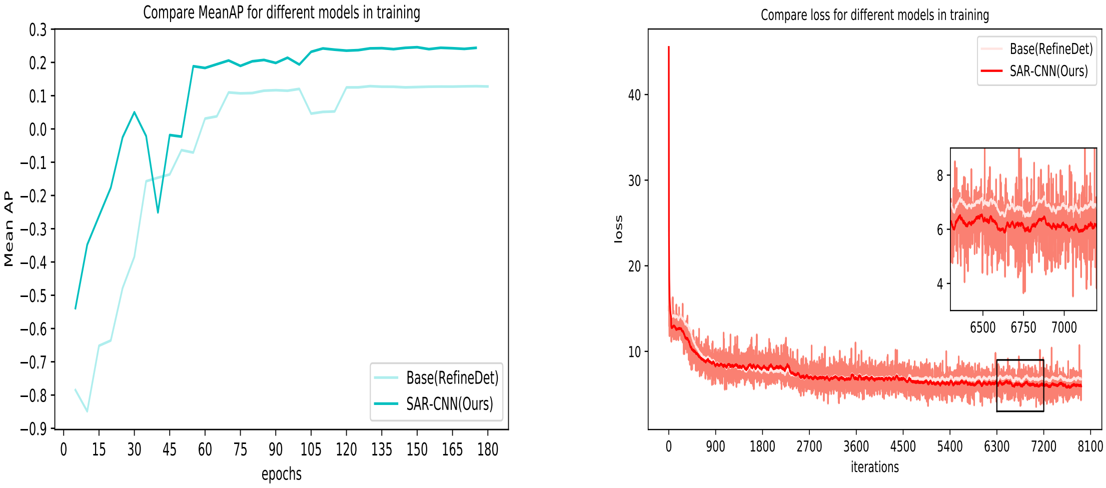
We used the PASCAL VOC 2010 dataset to evaluate the criterion, that is, the highest per-category accuracy (AP value) obtained during training, set the IoU value to 0.4, and compared it with the benchmark. As listed in Table 5, the benchmark performs slightly higher than our algorithm in detecting the categories of aortic enlargement, cardiomegaly, pulmonary fibrosis and infiltration, but all other categories exceeded the benchmark, and the benchmark value in pneumothorax was −100 (i.e., failed to detect this category). This illustrates the effectiveness of our targeted design network structure in enhancing the fitting of lung X-ray images.

Table 5. AP values for each category under the criteria of using PASCAL VOC 2010.

Method	Aortic Enlargement	Atelectasis	Calcification	Cardiomegaly	Consolidation	ILD	Pleural Effusion
Base	56.56%	10.20%	0.35%	44.18%	15.48%	16.50%	38.72%
SAR-CNN	54.46%	16.20%	9.13%	42.66%	25.97%	24.93%	39.39%
Method	Infiltration	Lung Opacity	Nodule Mass	Other Lesion	Pleural Thickening	Pneumothorax	Pulmonary Fibrosis
Base	28.21%	17.36%	12.18%	0.24%	16.03%	–	24.06%
SAR-CNN	22.92%	22.89%	13.13%	9.63%	19.44%	19.95%	23.32%

3.6.3. Training set performance

We present the performance effects of the benchmark algorithm and SAR-CNN on the training set in the form of mean AP and loss. The accuracy of the algorithm using the PASCAL VOC 2010 dataset evaluation criterion was tested every five epochs, the APs of each category were obtained, and the final mean AP was obtained. Figure 8(a) shows that the overall mAP of our algorithm was higher than that of the benchmark algorithm during the training process, and the fit of the features was always better. In Figure 8(b), the loss of the benchmark algorithm decreases slowly at the beginning of the training phase, and the loss is always higher than that of our algorithm at a later stage of training (e.g., iteration is in the range of 6500–7000), indicating that our algorithm converges faster than the benchmark algorithm.



(a)

(b)

Figure 8. (a) and (b) represent the comparison of the trend of mean AP and loss when SAR-CNN is trained with the benchmark algorithm using images with a resolution with 512.

3.7. Analysis of detection results

3.7.1. Detection box effects

Figure 9 shows a comparison between the detection effect of the benchmark model RefineDet and that of SAR-CNN (image resolution 512 × 512), where Figure 9(a) is the position of the real frame on the image in the test set, Figure 9(b) is the performance effect of benchmark model RefineDet in this task, and Figure 9(c) is the performance effect of our proposed model. It can be accurately observed that RefineDet is unable to detect the diseased areas at the margins of the lungs in the results from columns 1, 2 and 5 on the left. RefineDet also has the problem of error verification. For example, the error results of lung opacity, ILD and calcification were detected in the second and fourth columns on the left. However, the number, category, and position of the predicted boxes of the SAR-CNN are close to those of the label, and the confidence of the label box is as high as 0.99. In addition, the SAR-CNN can still maintain its detection accuracy under the complex intersection and superposition of multiple lesion areas, such as the detection results from the fourth and sixth columns on the left. In the case of high confidence, the SAR-CNN labeling frame is even more concise and intuitive than the real frame, for example, in the detection result from the fifth column on the left.

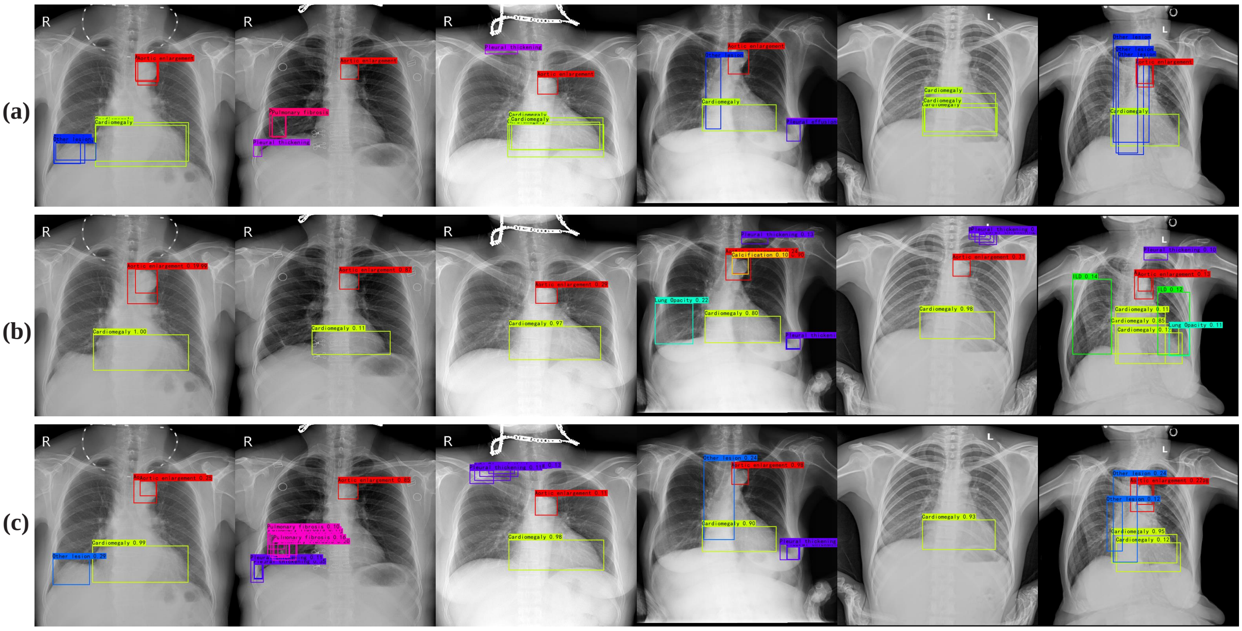


Figure 9. Comparison of detection effects between RefineDet and SAR-CNN: (a) Ground truth boxes position of test set, (b) RefineDet post-training testing effect, (c) SAR-CNN post-training testing effect.

3.7.2. Application of algorithms

As shown in Figure 10, we produced comparative maps of the lesion areas for the four sets of images based on the labels of the dataset and the recommendations of the physician. The four sets of images contained the original image, the focus area judged by the doctor, and the algorithm detection effect. The area judged by the doctor overlaps to a high degree with the results of our algorithm, and our detection frame does not obscure the original area, which is conducive to secondary analysis and examination.

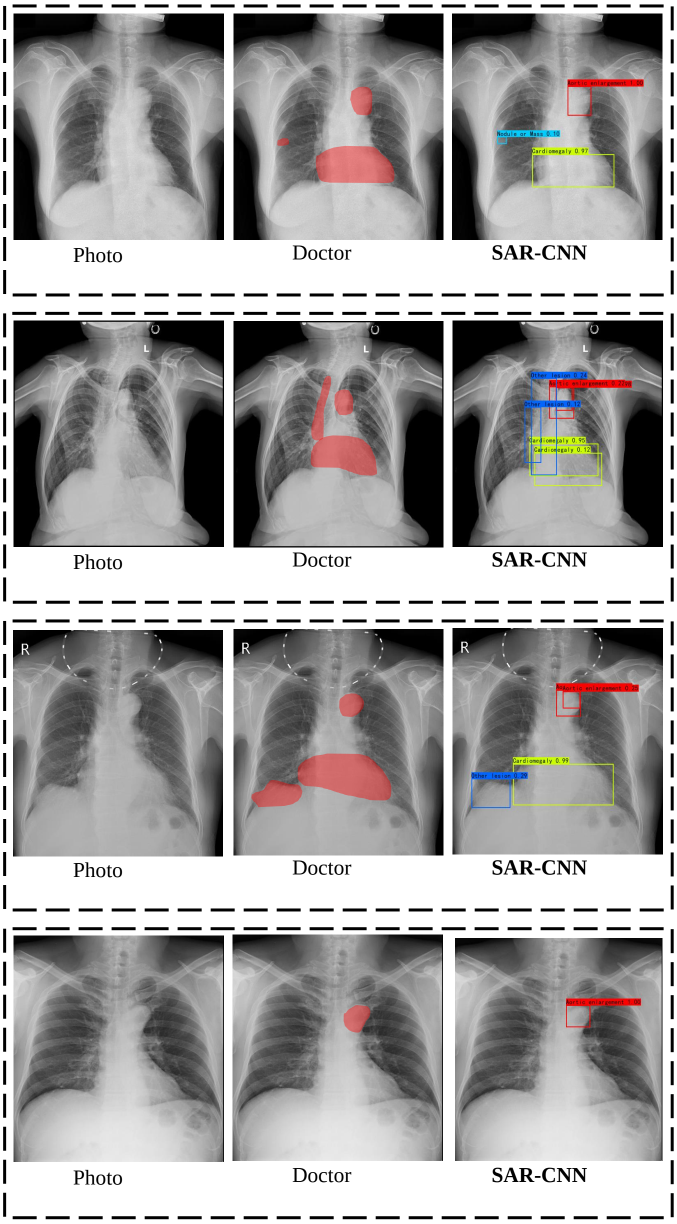


Figure 10. Comparison of the original image with the results of physician detection and SAR-CNN results.

4. Conclusions

In this paper, we propose a new SAR-CNN algorithm for disease localization detection in CXR images to improve the efficiency of physicians in diagnosing chest image. Three unique modules were proposed to help the CNN improve the sensitivity of the model to CXR features in terms of attention and feature fusion, and a training strategy from scratch was used to make the network more targeted. We tested the mAP, AP per category, and loss of the training set for different IoU values, and concluded that our algorithm is superior to mainstream target detection models. Using target detection technology to carry out AI medical research on CXR medical images can promote not only the application of deep learning technology and computer-aided diagnosis systems in the field of imaging examination but also the innovative intersection of information fields and biomedical research work. More importantly, it can reduce the workload of doctors and help promote the implementation of a national plan for the prevention and treatment of COVID-19, which has important theoretical and practical significance.

During our experiments, we found that lung X-ray images of COVID-19 for target detection were far less mature than those in the large CXR dataset and the corresponding target detection labels had less content. Most of these labels only indicate that the image has a certain disease classification or segmentation area. A mapping needs to be established between the CXR dataset from the previous CXR dataset and the COVID-19 dataset, similar to the approach of pre-training using a specific algorithm. However, we performed a targeted design in the network structure of feature processing and combined it with our proposed method for a comprehensive evaluation.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2021ZD0111000), Hainan Provincial Natural Science Foundation of China (621MS019), Major Science and Technology Project of Haikou (Grant: 2020-009), Innovative Research Project of Postgraduates in Hainan Province (Qhyb2021-10), Guangdong University Student Science and Technology Innovation Cultivation Special Fund Support Project (pdjh2023 a0243) and Key R&D Project of Hainan province (Grant: ZDYF2021SHFZ243).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. K. Singh, R. Pandey, R. N. Babu, COVIDScreen: Explainable deep learning framework for differential diagnosis of COVID-19 using Chest X-Rays, *Neural Comput. Appl.*, **33** (2021), 8871–8892. https://doi.org/10.1007/s00521-020-05636-6
2. C. Sohrabi, Z. Alsafi, N. Oneill, M. Khan, A. Kerwan, A. Al-jabir, et al., World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), *Int. J. Surg.*, **76** (2020), 71–76. https://doi.org/10.1016/j.ijsu.2020.02.034
3. J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, et al., Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation, *Med. Phys.*, **48** (2021), 1197–1201. https://doi.org/10.1002/mp.14676

4. C. Mattiuzzi, G. Lippi, COVID-19 vaccination is highly effective to prevent SARS-CoV-2 circulation, *J. Infect. Public Health*, **15** (2022), 395–396. https://doi.org/10.1016%2Fj.jiph.2022.03.006

5. G. M. Feuchtner, F. Barbieri, A. Luger, E. Skalla, J. Kountchev, G. Widmann, et al., Myocardial injury in COVID-19: The role of coronary computed tomography angiography (CTA), *J. Cardiovasc. Comput. Tomogr.*, **15** (2021). https://doi.org/10.1016/j.jcct.2020.07.002

6. H. Okano, R. Furuya, S. Mishima, K. Shimada, S. Umeda, T. Michishita, et al., DUAL-energy computed tomography findings in a case of COVID-19, *Acute Med. Surg.*, **8** (2021), e677. https://doi.org/10.1002/ams2.677

7. D. C. Rotzinger, C. Beigelman-Aubry, C. Von Garnier, S. D. Qanadli, Pulmonary embolism in patients with COVID-19: time to change the paradigm of computed tomography, *Thromb. Res.*, **190** (2020). https://doi.org/10.1016/j.thromres.2020.04.011

8. Y. Oh, S. Park, J. C. Ye, Deep learning COVID-19 features on cxr using limited training data sets, *IEEE Trans. Med. Imaging*, **39** (2020), 2688–2700. https://doi.org/10.1109/TMI.2020.2993291

9. Y. Peng, Y. Tang, S. Lee, Y. Zhu, R. M. Summers, Z. Lu, COVID-19-CT-CXR: a freely accessible and weakly labeled Chest X-Ray and CT image collection on COVID-19 from biomedical literature, *IEEE Trans. Big Data*, **7** (2020), 3–12. https://doi.org/10.1109/TBDATA.2020.3035935

10. W. Y. Chan, M. T. R. Hamid, N. F. M. Gowdh, K. Rahmat, N. A. Yaakup, C. Chai, Chest radiograph (CXR) manifestations of the novel coronavirus disease 2019 (COVID-19): A mini-review, *Curr. Med. Imaging*, **17** (2021), 677–685. https://doi.org/10.2174/1573405616666201231103312

11. E. J. Hwang, H. Kim, S. H. Yoon, J. M. Goo, C. M. Park, Implementation of a deep learning-based computer-aided detection system for the interpretation of chest radiographs in patients suspected for COVID-19, *Korean J. Radiol.*, **21** (2020), 1150. https://doi.org/10.3348%2Fkjr.2020.0536

12. M. Igi, M. Lieux, J. Park, C. Batte, B. Spieler, Coronavirus disease (COVID-19): The value of chest radiography for patients greater than age 50 years at an earlier timepoint of symptoms compared with younger patients, *Ochsner J.*, **21** (2021), 126–132. https://doi.org/10.31486/toj.20.0102

13. L. Cong, W. Feng, Z. Yao, X. Zhou, W. Xiao, Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer, *J. Cancer*, **11** (2020), 3615. https://doi.org/10.7150%2Fjca.43268

14. S. A. Agnes, J. Anitha, Appraisal of deep-learning techniques on computer-aided lung cancer diagnosis with computed tomography screening, *J. Med. Phys.*, **45** (2020), 98. https://doi.org/10.4103%2Fjmp.JMP_101_19

15. S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, et al., Automatic tuberculosis screening using chest radiographs, *IEEE Trans. Med. Imaging*, **32** (2013), 233–245. https://doi.org/10.1109/TMI.2013.2284099

16. L. Hogeweg, C. Mol, P. A. Jong, R. Dawson, H. Ayles, B. v. Ginneken, Fusion of local and global detection systems to detect tuberculosis in chest radiographs, in *International conference on medical image computing and computer-assisted intervention Springer*, (2010), 250–257. https://doi.org/10.1007/978-3-642-15711-0_81

17. S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, et al., Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration, *IEEE Trans. Med. Imaging*, **32** (2013), 577–590. https://doi.org/10.1109/TMI.2013.2290491

18. I. E. Livieris, A. Kanavos, V. Tampakas, P. Pintelas, An ensemble SSL algorithm for efficient Chest X-Ray image classification, *J. Imaging*, **47** (2018), 95. https://doi.org/10.3390/jimaging4070095

19. I. D. Apostolopoulos, T. A. Mpesiana, Covid-19: automatic detection from X-Ray images utilizing transfer learning with convolutional neural networks, *Phys. Eng. Sci. Med.*, **43** (2020), 635–640. https://doi.org/10.1007/s13246-020-00865-4

20. A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using X-Ray images and deep convolutional neural networks, *Pattern Anal. Appl.*, **24** (2021), 1207–1220. https://doi.org/10.1007/s10044-021-00984-y

21. L. Wang, Z. Q. Lin, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from Chest X-Ray images, *Sci. Rep.*, **10** (2020), 19549. https://doi.org/10.1038/s41598-020-76550-z

22. A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, A.Mohammadi, Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks, *Comput. Biol. Med.*, **121** (2020), 103795. https://doi.org/10.1016/j.compbiomed.2020.103795

23. D. Singh, V. Kumar, M. Kaur, Classification of COVID-19 patients from chest CT images using multi-objective differential evolution based convolutional neural networks, *Eur. J. Clin. Microbiol. Infect. Dis.*, **39** (2020), 1379–1389. https://doi.org/10.1007/s10096-020-03901-z

24. Y. Sun, B. Xue, M. Zhang, G. G. Yen, Evolving deep convolutional neural networks for image classification, *IEEE Trans. Evol. Comput.*, **24** (2019), 394–407. https://doi.org/10.1109/TEVC.2019.2916183

25. T. Mahmud, M. A. Rahman, S. A. Fattah, CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from Chest X-Ray images with transferable multi-receptive feature optimization, *Comput. Biol. Med.*, **122** (2020), 163869. https://doi.org/10.1016/j.compbiomed.2020.103869

26. W. Shen, M. Zhou, F. Yang, C. Yang, J. Tian, Multi-scale convolutional neural networks for lung nodule classification, in *International conference on information processing in medical imaging Springer*, (2015), 588–599. https://doi.org/10.1007/978-3-319-19992-4_46

27. M. Irfan, M. A. Iftikhar, S. Yasin, U. Draz, T. Ali, S. Husaain, et al., Role of hybrid deep neural networks (HDNNs), computed tomography, and Chest X-Rays for the detection of COVID-19, *Int. J. Environ. Res. Public Health*, **18** (2021), 3056. https://doi.org/10.3390/ijerph18063056

28. Y. E. Almalki, A. Qayyum, M. Irfan, N. Haider, A. Glowacz, F. M. Alshehri, et al., A novel method for COVID-19 diagnosis using artificial intelligence in Chest X-Ray images, *Healthcare*, **9** (2021), 522. https://doi.org/10.3390/healthcare9050522

29. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., Chexnet: Radiologist-level pneumonia detection on Chest X-Rays with deep learning, preprint, arXiv: 1711.05225.

30. F. Ucar, D. Korkmaz, COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-Ray images, *Med. Hypotheses*, **140** (2020), 1207–109761. https://doi.org/10.1016/j.mehy.2020.109761

31. X. Jiang, Y. Zhu, B. Zheng, D. Yang, Images denoising for COVID-19 chest X-Ray based on multiresolution parallel residual CNN, *Mach. Vision Appl.*, **32** (2021), 1–15. https://doi.org/10.1007/s00138-021-01224-3

32. A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P. R. Pinheiro, Covidgan: data augmentation using auxiliary classifier gan for improved COVID-19 detection, *IEEE Access*, **8** (2020), 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762

33. A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, J. J. P. C. Rodriguese, Identifying pneumonia in chest X-Rays: a deep learning approach, *Med. Hypotheses*, **145** (2019), 511–518. https://doi.org/10.1016/j.measurement.2019.05.076

34. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 770–788.

35. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv: 1409.1556.

36. S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Single-shot refinement neural network for object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), 4203–4212.

37. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2020).

38. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Proceedings of the International Conference on Machine Learning*, (2015).

39. X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, (2011), 315–323.

40. G. Huang, Z. Liu, L. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 4700–4708.

41. T. Y. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), 2117–2125.

42. F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, (2018), 2403–2412.

43. J. Park, S. Woo, J. Y. Lee, I. S. Kweonet, Bam: Bottleneck attention module, preprint, arXiv: 1807.06514.

44. S. Woo, J. Park, J. Y. Lee, I. S. Kweonet, Cbam: Convolutional block attention module, in *Proceedings of the European conference on computer vision (ECCV)* , (2018), 3–19.

45. H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, et al., VinDr-CXR: An open dataset of chest X-Rays with radiologist's annotations, *Sci. Data*, **9** (2022), 429. https://doi.org/10.1038/s41597-022-01498-w

46. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., PyTorch: An imperative style, high-performance deep learning library, *Adv. Neural Inf. Process. Syst.*, **32** (2019).

47. R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, et al., ScratchDet: Training single-shot object detectors from scratch, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 2268–2277.

48. M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, *Int. J. Comput. Vision*, **88** (2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4