IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 15, 2022, accepted May 31, 2022, date of publication June 10, 2022, date of current version June 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182104

Attention-Guided Feature Extraction and Multiscale Feature Fusion 3D ResNet for Automated Pulmonary Nodule Detection

GUANGLU ZHANG[1,2], HONGJUN ZHANG[1], YUHUA YAO[2], AND QIUHUI SHEN[1]
[1]Army Engineering University of PLA, Nanjing 210007, China
[2]Hainan Normal University, Hainan 571158, China

Corresponding author: Hongjun Zhang (jsnjzhj_lgdx@163.com)

This work was supported in part the Education Department of Hainan Province under Grant Hnky2021ZD-12; in part by the Natural Science Foundation of Hainan Province under Grant 621RC1060; in part by the Specific Research Fund of The Innovation Platform for Academicians of Hainan Province; in part by the Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, Haikou, China; and in part by the National Natural Science Foundation of China under Grant 62162025.

ABSTRACT Automatic detection of pulmonary nodules is critical for the early diagnosis and prevention of lung cancer. Computed tomography (CT) is an effective and economical lung cancer detection method. In CT images, the size and shape of pulmonary nodules appear different, and some nodules appear similar to the surrounding tissues. Therefore, the automatic localization of pulmonary nodules in CT images is a challenging task. An attention-embedded three-dimensional convolutional neural network is proposed for pulmonary nodule detection in the current study. Specifically, 1) channel-spatial attention guides 3D ResNet to down sample the input 3D CT patch. The channel pays attention to important features and the space to the region of interest. The two form a complementary feature extraction mechanism to effectively help the global flow of information in the network and refine the feature mapping to extract the nodule context features. 2) The channel-spatial attention module changes the fusion model of the feature pyramid, adaptively adjusts the pixel-level weight between features and extracts multi-scale representative node features. 3) The deep separable convolution is used to replace the standard convolution of ResNet, reducing the time cost and improving the efficiency of model training on the premise of ensuring the model's performance. 4) To adapt the distribution of nodule scale, different characteristic layers correspond to two sizes of anchors. Under the condition of ensuring the detection rate of nodules, the number of anchor frames is reduced, and the network sensitivity is improved. Finally, several ablation experiments are carried out using the LUNA16 dataset. The results revealed that the attention-guided network could extract the multi-scale representative features of nodules, and the average sensitivity was 97.7%. Additionally, the CMP score reached 0.912. The extensive experiments demonstrate that the proposed approach can effectively improve the detection sensitivity and control the number of false positive nodules, which has clinical application value and a certain reference value.

INDEX TERMS Channel-spatial attention mechanism, multi-scale features, pulmonary nodule, computed tomography scan, medical computer vision.

I. INTRODUCTION

Globally, cancer is the leading cause of death and a major obstacle to improving life expectancy [1]. Lung cancer is particularly deadly and is the leading cause of cancer deaths in men. Lung cancer ranks first among countries with high HDI (39 per 100,000 people). According to 2020 global cancer incidence and mortality estimates compiled by the International Agency for Research on Cancer, nearly 10 million people will die from cancer, of which lung cancer would account for 11.4% [2], [3]. Early lung cancer symptoms are not evident, and once a person starts to exhibit uncomfortable

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

symptoms, the cancer is at an incurable stage. Therefore, early diagnosis of lung cancer could effectively prolong the survival rate. Lung cancer might appear as pulmonary nodules in the early stage. In CT images, pulmonary nodules refer to round or oval lung tissue masses with a diameter between 3 and 30 mm, and those with a diameter of less than 3 mm are called micronodules. Clinically, pulmonary nodules can be identified according to the morphology, location, density, intensity, calcification, and changes in surrounding tissues. Therefore, detecting pulmonary nodules is vital for the early diagnosis of lung cancer.

In general, pulmonary nodules can be effectively detected with computed tomography (CT), MRI, and positron emission tomography (PET-CT), among which low-dose CT of the chest is recognized as the most efficient and economical method of diagnosis [4]. CT scan could intuitively describe the morphological characteristics of lesions, with a sensitivity of 98%–100% and specificity of 54%–93%. However, data from a single scan for each patient consists of 200 to 700 2D images (2-Dimension, 2D), and the resolution of each image is 300*300 or more. Relying entirely on doctors' subjective judgment takes time and effort and increases the rate of misdiagnosis and missed detection due to objective factors, such as fatigue and lack of professional experience and attention. Therefore, we must develop an automatic detection system for pulmonary nodules to help doctors discover potential abnormalities, reduce workload, and improve pulmonary nodules diagnosis accuracy.

Deep learning technology has rapidly developed, and deep convolutional neural network (CNN) has made remarkable achievements in various problems [5]–[8]. Deep learning techniques also play an important role in medical image analysis. CNN's powerful feature representation and end-to-end training model can obtain rich target features from image data, making it a promising method for lung nodule detection. A series of CNN-based pulmonary nodule detection methods [9]–[14] have been proposed that achieved excellent detection performance. We studied the literature on deep learning-based pulmonary nodule detection methods in the past 10 years and discovered that the existing pulmonary nodule detection system has many challenges:

1. The shapes and sizes of pulmonary nodules are diverse. The location of nodules is specific, such as near vascular nodules and lung wall nodules. Some nodules have margins similar to the surrounding lung parenchyma. The size of nodules is mainly distributed between 3 and 40 mm, belonging to small targets. Due to these peculiarities of nodules [15], their detection using the existing algorithms is inefficient, resulting in missed detections and false positives.

2. To improve the network performance of pulmonary nodules detection, some CNN-based methods primarily focus on broadening or deepening the model structure to learn more high-resolution features. However, inherent correlation studies between feature layers are insufficient, which weakens the characterization ability of CNN.

3. Some algorithms use a feature pyramid network (FPN) [16] to extract multi-scale features of nodules. However, the feature pyramid restores multiple down-sampled features by up-sampling, and the pixel values and positions of up-sampled features are inconsistent with those of the original feature images without down-sampling. The addition of concatenation operations to fuse different levels of feature maps leads to discrepancies or ambiguities in the fusion process, making it difficult to match the high-level semantic information with the underlying structure. As a result, the detailed representation of features or contextual information is corrupted, which hinders the flexibility of the network to extract features.

4. Some nodule detection systems ignore the performance and efficiency of the model while pursuing high precision.

Inspired by previous studies, we proposed a 3D Faster R-CNN pulmonary nodules detection model embedded with attention mechanism and adaptive feature pyramid to address the problems mentioned above. 3D Faster R-CNN model structure is the most advanced two-stage target detection method, with the ability to automatically extract depth features and locate targets, and the detection accuracy is better than the classical algorithm of one-stage target detection. Unlike natural images, lung nodules' size, shape, and texture are arbitrary, and the surrounding tissue might resemble the nodules. Therefore, accurate identification of their features is challenging. Using the excellent two-stage model, the regional proposal of pulmonary nodules in the first stage and then the secondary correction of the regional proposal can obtain more accurate detection results. We adopted channel-space attention to improving the residual blocks of ResNet and used the improved ResNet as the feature extraction backbone network. ResNet solves the degradation problem of the network by short-circuiting the connections and fully extracting the local spatial contextual information of the nodules. Channel attention adaptively adjusts channel weights to enhance salient features of nodes, inhibit unimportant features, and enhance the flow of shallow features. Spatial attention adaptively adjusts region of interest weights. It guides the network to learn abstract semantic features and extract representative global contextual structural information of nodules, forming a complementary mechanism with channel attention to complement and activate more location information. In the current study, a channel-space attention (CSA)-based adaptive feature fusion (AFF) network is proposed for fine-grained feature extraction. AFF performs pixel-level adaptive feature fusion of feature maps from different layers and channel and space directions, respectively. AFF models the contextual information of lung nodules by fusing high-level semantic features and shallow detailed features to extract multi-scale fine-grained nodule features. The 3D structure is more complex than the 2D structure, with several parameters and impressive time consumption. In the present paper, Depth-wise Separable Convolution (DSC) is introduced to replace the 3D convolutional block of the original residual structure and reduce the number of CNN parameters, simplify the feature encoding and improve the performance of the model. Only a few images usually contain pulmonary nodules in a CT image dataset. Therefore, the number of positive and negative samples in the data set is extremely unbalanced. The imbalance seriously affects the model's detection performance and even leads to model degradation and convergence difficulties during the training process. Therefore, in this paper, focal loss [17] is used to calculate the classification loss, and the super-parameters in the function were selected through experiments.

To summarize, our contributions are listed as follows:

1. The proposed attention-embedded pulmonary nodules detection model based on 3D Faster R-CNN. ResNet is improved by embedding attention to residual blocks and DSC instead of the standard 3D convolution in the original residual blocks. The improved ResNet is used as the backbone network for feature extraction, which adaptively focuses on the main channels and regions of interest and guides the low-level location, contour, edge, texture and shape features and high-level abstract semantic features to activate each other to form complementary feature flow and extract more representative nodular features. DSC ensures the accuracy and sensitivity of nodule detection, reduces the number of parameters and calculations of the model, and enhances its detection performance.

2. An AFF network is proposed based on the attention mechanism. The attention mechanism adaptively guides the network to construct pixel-level feature fusion weights between up-sampled and down-sampled features, effectively modeling contextual feature information and enhancing network representation capability.

3. Anchors of different scales are assigned to feature maps of different resolutions for extracting region suggestions, which efficiently fit the scale distribution of nodules and improve network detection sensitivity.

4. Focal loss function was used to calculate the classification loss, and the optimal hyperparameters were selected through many experiments to balance positive and negative samples effectively.

II. RELATION WORK

With the wide application of deep learning technology in medical image analysis, natural image detection algorithms based on deep learning demonstrate excellent performance in medical images. Two primary CNN-based lung nodule detection models are present, namely, one-stage detection represented by YOLO series [18] and SSD [7] and two-stage detection, represented by Faster R-CNN [19] and Mask R-CNN [20]. Inspired by the classical CNN network detection model, a series of automatic detection algorithms for pulmonary nodules based on CNN are proposed. Setio *et al.* [12] extracted a set of 2D patches from multiple directions at each candidate position and then input each patch into the CNN stream to learn features separately. Finally, all output features are combined through a dedicated fusion method to calculate the final score and reduce false positives. Ding *et al.* [9] used 2D Faster R-CNN [19] combined with the VGG-16 model to generate suspect candidate nodules, and 3D DCNN was used to remove false-positive nodules. The model obtained a high CMP score of 0.891. CT images are composed of several 2D images, and each 2D image is only a cross-section of the CT image sequence, which cannot completely represent the nodules with different morphologies and variable sizes. Therefore, detecting lung nodules using 2D images destroys the continuity of nodules in 3D space and cannot fully utilize the information of contextual features of nodule images.

Compared with 2D CNN structures, 3D CNN network structures are more conducive to identifying nodules by training 3D samples and extracting more representative nodule features using the spatial information of the nodule's context and achieving good results in terms of accuracy. Zhu *et al.* [14] adopted 3D Faster R-CNN with DPNs and encoder-decoder structure similar to U-NET [21] for nodule detection. Then, the gradient Boosting Machine, based on deep 3D DPN features, original nodule CT pixels, and nodule size, was designed for nodule classification to reduce false positives. Finally, the model's nodule- and patient-level diagnosis on the LDC-IDRI dataset was comparable with that of experienced doctors. Dou *et al.* [10] dopted three 3D CNNS, each encoding a specific level of background information. Finally, the final classification result is obtained by integrating the probability prediction results of these networks, and the detection result of the proposed model is better than that of most pulmonary nodules recognition algorithms. Wang *et al.* [22] proposed an improved 3D Faster R-CNN structure model. The model used VGG16 as the backbone network for feature extraction and stitched together four features from shallow to deep in the backbone network. Different scale anchors were located using the feature pyramid. The feature maps were sent to ROI for classification as candidates or non-candidates by feature sharing. Advanced feature mapping, for the classification of size nodes, has richer semantic information than single layer features. In the second stage, a simple 3D classification network is designed for candidate nodule classification to reduce false positives, and the detection method can achieve high sensitivity with few FPs. Fu *et al.* [23] proposed an improved 3D U-Net deep learning model for automatic detection of pulmonary nodules on chest CT images. The model was validated through 89 CT scans of LUNA16, a public dataset, with a CPM of 0.947 and 450 chest CT scans provided by a City University hospital in Japan, with a CPM of 0.833. This indicates that the improved 3D U-NET deep learning model has good robustness in the detection performance of pulmonary nodules. Tang *et al.* [24] based on the multi-scale feature of transfer learning, established a 3D U-NET CNN with a multi-scale feature structure, which can detect pulmonary nodules from the thoracic region containing background and noise. The accuracy of the model for detecting small nodules reached 70%.

Attention mechanism: The attention mechanism has attracted extensive attention in computer vision research.

Attention simulates sensory features in the human visual system by selectively focusing on important and evident information through local observations. The attentional mechanism adaptively enhances the weight of rich feature information by suppressing the expression of unimportant features. The attention module is lightweight and can be embedded directly into deep convolutional neural networks or replace components of CNN networks, which adaptively adjust the feature weights of contextually important regions to improve the network's performance effectively. Squeeze-excitation network (SENet) is frequently used in many target detection and image classification [25]. SENet is widely used in medical image analysis. Yuan *et al.* [26], Zhang *et al.* [27] and Zhang *et al.* [28] introduced SE modules into the feature extraction network to improve pulmonary nodules' detection and classification performance. However, the SE module compresses the pixel values of each channel into a real number by using the global average pooling (GAP) layer, which unilaterally considers channel correlation without considering location information. However, spatial attention plays an important role in determining the ''where'' to focus on. In addition, the SE module only exploits the average pooling feature, ignoring the importance of the maximum pooling feature, which encodes significant features and compensates for the GAP feature. Pulmonary internal environment is complex, the size and shape of nodules are diverse, and the nodules near blood vessels and lung walls are blurred with surrounding tissues. The SE module inevitably leads to the loss of nodule structural information, resulting in missed detection or false-positive. Convolutional block attention module (CBAM) [29] consists of two parts: channel attention and spatial attention, where channel attention uses global maximum pooling and average pooling operations to adaptively adjust channel weights to focus on important features, whereas spatial attention adaptively refines spatial features to focus on key regions, forming a complementary mechanism that effectively compensates for the shortcomings of SE. The effectiveness of CBAM on pulmonary nodule detection has not been extensively explored yet.

Feature Pyramid Network (FPN): FPN [16] has become an important module of the target detection algorithm. In the detection network of pulmonary nodules, FPN was established on the backbone network, and multi-scale feature images were fused to obtain multilevel features with different resolutions. Pulmonary nodules with different scales were assigned to feature images with different resolutions. It can effectively alleviate the scale change and heterogeneity problems of pulmonary nodules detection, effectively extract fine-grained characteristic context information, and assist small target detection. Zhang *et al.* [28] proposed 3D FPN lung nodule detection network to extract multi-scale features of small targets. Feature pyramids can be used to obtain high-resolution features by up-sampling higher-level low-resolution features and then fusing them with lower-level high-resolution features through addition or concatenation. However, differences between two feature layers with different resolutions in the backbone are present to a certain extent, and direct addition will destroy the representation of features in both layers. Moreover, direct concatenation is not conducive to certain region details or the detection of small targets and accurate target localization. To solve this problem, in the current paper, we proposed an AFF pyramid structure, which by embedding the CBAM module, uses the high- and low-level features to predict the pixel-level fusion weights and effectively extract the 3D contextual information of nodules.

III. MATERIALS

Lung nodule detection is a target detection task, and two main types of representative methods for current target detection are One-step detection methods and a two-stage detection model. One-step detection methods use CNN to extract features and then directly classify and regress. The whole process only needs one step, so it is relatively fast with low accuracy. The 3D Faster R-CNN model structure is the state-of-the-art and two-stage approach to detecting targets. First, all target suggestions are predicted as much as possible from CT images, and then features in the target suggestions are extracted using ROI Align [20] for classification and fine coordinate regression. Two-stage detection usually combines the feature pyramid module to propose multi-scale features and then predicts target suggestions based on each scale feature, making full use of contextual information to improve the sensitivity of small nodule detection. The lung environment is complex, and nodules are not discernible from the surrounding tissues. Often nodules are confused with lung wall, lymph nodes, blood vessels, bronchi, and other pathological tissues on CT images, which requires accurate detection algorithms to improve the nodule detection performance. 3D Faster R-CNN's two-stage detection model is beneficial for lung nodule detection. We proposed an attention-embedded lung nodule detection model based on a 3D fast R-CNN model. The overall process of the automatic detection system for pulmonary nodules is shown in Fig.1, which aims to extract contextual feature information of the 3D nodules fully and accurately detect pulmonary nodules. It mainly consists of data acquisition, pre-processing feature extraction, fusion, and nodule detection.

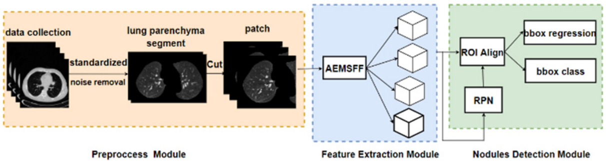

FIGURE 1. Lung nodule detection process with embedded attention mechanism.

A. DATA COLLECTION AND PRE-PROCESSING

Data collection: The present paper uses the LUNA16 dataset (Tenchi competition dataset) to evaluate the proposed algorithm. The data can be downloaded at https://luna16.grand-challenge.org. LUNA16 is a subset of data from The Lung

Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). The LDC-IDRI [30], [31] data were collected by the National Cancer Institute, the world's largest publicly cited database of lung nodules, with 1018 studies, to analyze early cancer detection in high-risk populations. In LUNA16 [32] data set, the CT of specimens with slice thickness greater than 3 mm was removed, and specimens with inconsistent slice space and missing sections were also removed. Finally, 888 CT pieces were generated. A total of 36,378 nodules were identified (labeled in LDC-IDRI) in 888 CT scans, and 5765 nodules with diameters >3 mm were screened and used. Nodules whose distance between the centers of two nodes was less than the sum of the radii of the two nodes were merged, leaving 2290 nodes after the merger. These nodules are also the data set used for the experiments in the current paper. The lung nodule annotations in the dataset were collected by four professional radiologists during a two-stage image annotation process. The number of nodules annotated by at least one, two, three, and four radiologists was 2290, 1602, 1186, and 777, respectively. LUNA16 was used for 1186 nodules labeled by at least three radiologists as positive samples in the reference standard. We used the diameter of each nodule provided in the LUNA16 match to generate a bounding box as a label for detecting nodules.

Pre-processing: We refer to the data pre-processing process in an article [28], [33]. Data pre-processing mainly includes resampling, normalization, denoising, and lung parenchyma segmentation. Original CT image scanning consists of two-dimensional images with different pixel intervals because CT images were collected on different devices. Different sampling frequencies lead to different CT data sizes and nodular diameters. All CT scans are resampled at 1 mm × 1 mm × 1 mm pixel spacing using linear interpolation and adjusted to the same orientation to eliminate inconsistent resolution of different CT scans. A Gaussian kernel with a size of 3 × 3 was used to perform Gaussian filtering on image slices to remove the noise during acquisition. According to the HU (Hounsfield Unit) value of lung CT, the lung areas with an HU value between [−1000,400] are reserved, and other irrelevant areas are omitted. The mask of the lung region was obtained using a lung segmentation image provided by LUNA16. Lung segments, including all nodules, were identified using the convex wrap and dilation method. The final image pixel values were cropped to [−1200,600] and normalized to [0, 255]. The non-lung region was filled with pixels at 170, and the cavity was expanded, convexly packed, and dilated by bifurcated morphological operations to preserve more boundary information and segment the lung parenchyma.

B. ATTENTION-GUIDED 3D RESNET AND ADAPTIVE MULTISCALE FUSION NETWORK

Inspired by previous research results, we proposed a 3D ResNet based on attention-guided contextual feature extraction and an FPN with an adaptive fusion of multi-scale features which can effectively extract representative nodal features. ResNet achieves feature reuse by adding jump connection channels, making full use of 3D contextual information of lung nodules, and solving the network's gradient disappearance and gradient explosion problems. ResNet is widely used in the field of computer vision. CSA is embedded in the ResNet network, which models the correlation between channels, improves the feature representation of the region of interest, and extracts more representative nodal features. DSC is used to replace the ordinary convolution of the original residual network, which improves the performance and efficiency of model training while ensuring model detection accuracy. Attention is introduced to develop a multi-scale feature fusion and extraction network, which adaptively context models high-level abstract semantic features and shallow location structure features and achieves pixel-level fine-grained feature fusion. The network framework consists of two main components: the 3D residual network with embedded attention and the adaptive multi-scale feature fusion network. The structure is shown in Fig. 2.

1) GENERAL STRUCTURE OF FEATURE EXTRACTION AND FUSION

Encoding networks with embedded attention: The network input is a 3D patch, which first passes through two convolutional layers (kernels = 3 × 3 × 3, channels = 24) and then passes through five 3D Residue (3D Res-AM) modules of the embedding Attentional Mechanisms (AM) structure to extract features. Each 3D Res-AM is followed by a max pooling (kernel = 2 × 2 × 2, stride = 2) layer to reduce the size of the feature map, extracting the features of the image patches from shallow to deep layers, layer by layer. The size of the 3D feature maps from shallow to deep are obtained as (128^3@24, 64^3@32, 32^3@64, 16^3@64,8^3@64,4^3@64). In the present paper, ResNet has been improved and readjusted as the backbone network. Each of the five 3D Res-AM modules consists of residual cells with the same structure and the number of iterations of the residual cells, namely 3, 4, 6, 4, and 3. 3D Res-AM enriches the attention graph by effectively combining spatial attention and channel attention, utilizing global contextual information to selectively highlight or weaken features, and guide the network to extract more representative features and regions of interest.

Decoding networks with AFF Network: In the feature fusion path, a deconvolution layer (kernel = 2 × 2 × 2, step = 2) is first used to up-sample the feature map with the lowest resolution obtained from down-sampling, and a feature map with a size of 8^3@64 is obtained. The up-sampled feature map and the down-sampled feature map with the same size are fused using the AFF module to obtain the feature map with the scale of 83@64. Then, four feature images fused with the corresponding size of down-sampling are obtained using three sets of deconvolutions, AFF, and 3D RES-AM operations. As shown in Fig. 2. The predicted feature maps of the four scales obtained in the feature extraction stage were (64^3@64, 32^3@64, 16^3@64, 8^3@64). Next, we fed

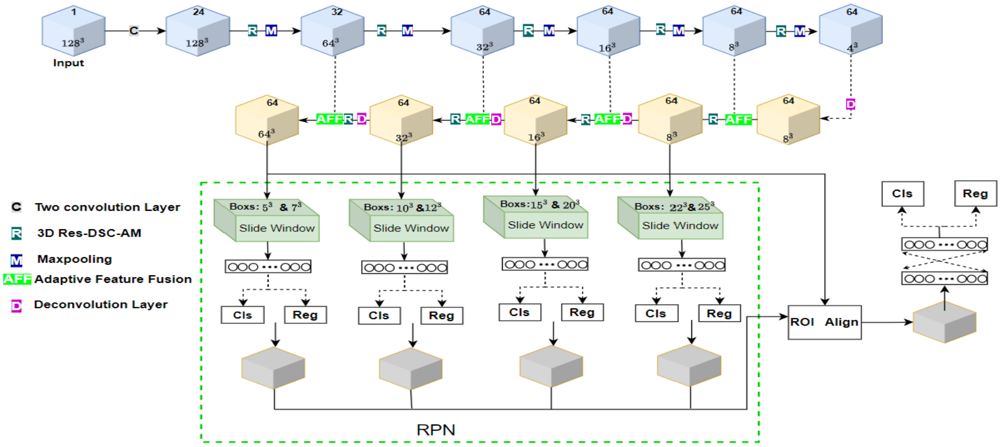

FIGURE 2. **Overall framework of general structure of feature extraction and fusion.**

each feature layer into the RPN network to extract the region suggestion box. The feature fusion module with embedded attention achieves pixel-level AFF by contextual modeling of high-level features and shallow-level features.

2) CONVOLUTIONAL BLOCK ATTENTION MODULE

The attention mechanism adaptively enhances the weight of rich feature information while suppressing the expression of unimportant features, which effectively improves the performance of the network in recognizing features. SENet selectively scales the channels to capture the channel dependence between channels. A lung nodule detection algorithm has been proposed based on SENet and has achieved excellent detection performance. However, the SE module only considers channel correlation and does not account for location information; however, spatial attention plays a major role in determining the "where" aspect of attention. In complex pulmonary nodule states, the SE module loses nodule details, which reduces the detection performance. CBAM [29] guides the network to focus adaptively on important features and regions of interest along both channel and spatial directions, respectively, effectively compensating for the lack of SE. In the present paper, we used the CBAM module to improve the residual network and construct the AFF pyramid to improve the expression of features. An illustration of the CBAM is shown in Fig. 3.

Channel Attention architecture (CA): For any input intermediate layer feature F∈ $R^{D\times H\times W@C}$ The average pooling and the global maximum pooling operations aggregate spatial feature information to form two different contextual descriptors; a shared network consisting of a two-layer perceptron and a hidden layer is applied to approximate each descriptor and merge the output feature vectors using element-wise summation. Generate channel attention map $CA_F \in R^{1\times 1\times 1@C}$. The channel structure is shown in Fig. 3(a). In short, the detailed operation is described as follows:

$$CA_F = \sigma\left(MLP\left(AvgPool\left(F\right)\right) + \mathrm{MLP}\left(MaxPool\left(F\right)\right)\right) \quad (1)$$

where σ denotes the sigmoid function, The parameter r of MLP is eight in the current study. AvgPool and MaxPool denote the GAP and max pooling, respectively. Average pooling aggregates spatial features, and max pooling focuses on essential cues of the target features, which infer fine-grained channel attention. Both operations are used simultaneously to improve the representation capability of the network.

Spatial attention architecture (SA): Spatial attention focuses on the region of interest and plays an effective complementary role in channeling attention. For any input intermediate layer feature F ∈ $R^{D\times H\times W@C}$, the average pooling and the global max pooling operations are used to encode each pixel at all spatial locations along the channel direction, generating two feature descriptors. The two concatenated descriptors are convolved to generate a spatial feature map $SA_F \in R^{D\times H\times W@1}$. The spatial structure is shown in Fig. 3(b), and the detailed operation is described as follows:

$$SA_F = \sigma(\mathrm{conv}\left(\left[\mathrm{AvgPool}\left(\mathrm{F}\right) : \mathrm{MaxPool}\left(\mathrm{F}\right)\right]\right) \quad (2)$$

where σ denotes the sigmoid function, Conv represents a convolution operation. The filter size of the convolution operation in this study is 3 × 3 × 3. In the SA structure, pooling operations are applied along the channel axes to highlight regions of information effectively. Convolution layers are applied to cascade feature descriptors to emphasize or suppress spatial locations.

Channel-Spatial Order attention (CSA): Channel and space focus on representative features and regions of interest and compute complimentary attention. The two modules can be inserted into an existing efficient network in parallel or series. In paper [29], after extensive compatible experimental studies, the CSA tandem module is superior to other approaches; namely, the CBAM approach is optimal. For any input intermediate layer feature F∈ $R^{D\times H\times W@C}$. First feature map F is calculated using channel attention, and the obtained channel features are multiplied by F to obtain the middle layer feature map $F_M \in R^{D\times H\times W@C}$. Then, F_M is used as the input of SA to calculating the spatial feature

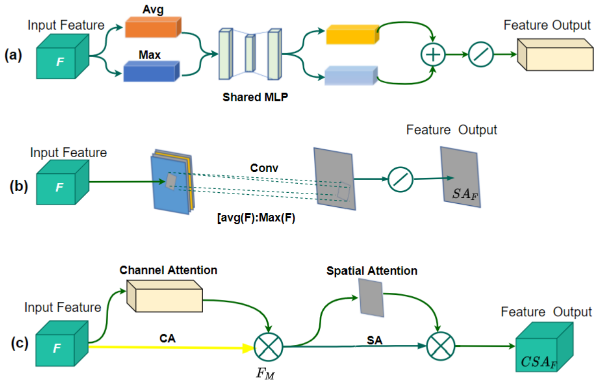

FIGURE 3. **F Channel-spatial attention structure diagram: (a) Channel attention (CA), (b) Spatial attention (SA), (c) Channel-Spatial Attention (C-SA).**

map $SA_F \in R^{D\times H\times W@1}$. Finally, SA is multiplied with the feature map F_M to obtain the CBAM feature map $CSA_F \in R^{D\times H\times W@C}$. The CSA structure is shown in Fig. 3(c), and the detailed operation is described as follows:

$$F_M = CA(F) \otimes F \tag{3}$$

$$CSA_F = SA(F_M) \otimes F_M \tag{4}$$

3) RESIDUAL NETWORK STRUCTURE BASED ON DSC AND CSA (RES-DSC-AM)

ResNet exhibited excellent performance in the field of computer vision, and its core is the introduction of residual structure, which enables feature reuse by adding shortcut connection channels, solves the gradient disappearance/explosion problem of deep networks, and enhances the information flow of feature propagation. However, ResNet ignores the correlation between channels and the importance of regions in spatial, therefore, it cannot extract global and local features simultaneously. To fully utilize the 3D spatial information of lung nodules, we chose a 3D patch as the input of the network. Compared with the 2D network model, the 3D network has several computationally expensive and complex parameters. We improved the residual structure by introducing 3D DSC and embedding attention. The structure is shown in Fig. 4.

Depth-wise separable ResNet (Res-DSC): The convolution operation is decomposed into two steps, Depth-Wise Convolution (DWC) and Point-Wise Convolution, which can effectively reduce the computation while ensuring the accuracy of the model. In the present paper, the 3D DSC is used to replace the original standard convolution, and the improved residual structure is shown if 4(b).

Embedded-Attention ResNet (ResNet-DSC-AM): Given an input map, channel and space, two attention modules compute complementary attention, focusing on the "what" and "where", respectively. The two modules can be placed in a parallel or sequential manner. The attentional convolution operation extracts information features by mixing cross-channel and spatial information and emphasizes meaningful features along the two main dimensions of the channel and spatial axis. The channel and SA modules are embedded sequentially in ResNet so that each of the branches can learn "representative features" and "important regions" on the channel and spatial axes, respectively. As a result, the attention module effectively helps information flow globally in the network by learning which features to emphasize or suppress and further refines the feature mapping to enhance the nodule contextual feature extraction. Fig. 4(c).

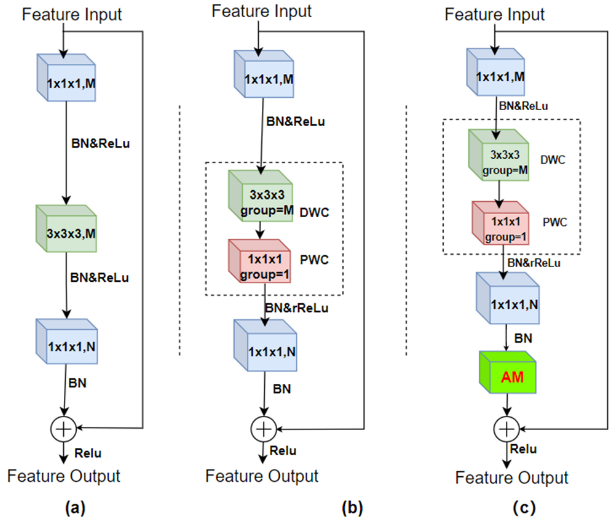

FIGURE 4. **Improved residual block structure:(a) Original 3D residual block. (b) 3D Res-DSC. (C) 3D Res-DSC-AM.**

4) ADAPTIVE FEATURE FUSION (AFF) Network

FPNs fuse the up-sampled feature maps with down-sampled feature maps with the same structure in the backbone network and predict targets of different sizes on feature maps, which is an essential structure in target detection algorithms and effectively solves the small target detection problem. Therefore, the feature pyramid structure is widely used in target detection models. However, FPN has some drawbacks. The FPN recovers the feature maps of the backbone network after being down-sampled several times by up-sampling, and the recovered feature maps are prone to misalignment. Therefore, the feature maps obtained after fusion are discrepant or blurred. FPNs fuse feature maps from different layers by traditional summation of corresponding elements or simple splicing operations. The up-sampled feature maps differ from the backbone network response structure feature maps, and direct summation will destroy the feature representation of the feature maps. If the fusion is directly collocated, the regional details are easily lost, unfavorable for detecting small target nodules near lung walls and blood vessels. An attention-based adaptive FPN structure is proposed. The structure is shown in Figure 5. The adaptive fusion process can be summarized as follow: For any input, high-level low-resolution feature map $F_{HL} \in R^{D\times H\times W@C}$, up-sampled two times to obtain the feature map $F_{HL-UP} \in R^{D\times H\times W@C}$, then, F_{HL-UP} and the low-level high-resolution feature map $F_{LH} \in R^{D\times H\times W@C}$

from the backbone network are concatenated, and then the intermediate layer F_M is obtained, which achieves coarse-grained spatial feature fusion. $F_M \in R^{D\times H\times W@2C}$ is fed into the CSA module after descending convolution and smoothing convolution, respectively, to obtain the feature map $F_{CSA} \in R^{D\times H\times W@C}$. The CSA module adaptively adjusts the pixel weights from the pixel-level along the channel and spatial directions, respectively, to capture multi-scale information of lung nodules. Finally, F_{HL-UP}, F_{LH} and F_{CSA} are summed to achieve fine-grained feature fusion $F_{AFF} \in R^{D\times H\times W@C}$. it is described as:

$$F_M = cat\,(F_{HL-UP}, F_{LH}) \tag{5}$$

$$F_{CSA} = CSA(conv1\,(conv2\,(F_M))) \tag{6}$$

$$F_{AFF} = F_{CSA} \oplus' F_{HL-UP} \oplus' F_{LH} \tag{7}$$

where *cat* represents concatenation operation. *conv*1, *conv*2 represent convolution operations which kernel size is 1 × 1 × 1 and 3 × 3 × 3, respectively. ⊕ represents element-wise add operation.

The adaptive fusion pyramid can dynamically adjust feature fusion by adjusting the weighting of higher-level low-resolution features and lower-level high-resolution features at the pixel-level fine-grained, which preserves more structural information for regions requiring more detail (e.g., small target nodules, near lung walls, and vascular nodules). It can efficiently extract multi-scale spatial information at a finer granularity level, and form long-range channel dependencies and learning spaces are richer multi-scale feature representations.

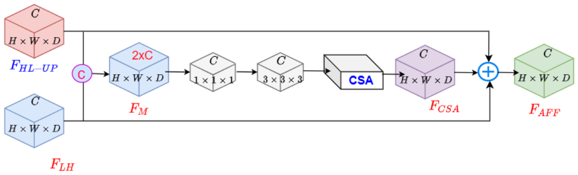

FIGURE 5. Illustration of adaptive feature fusion network (AFF) structure.

C. RPN NETWORK

RPN accepts four feature layers of different scales after adaptive feature pyramid fusion, and then the sliding window is run through the feature map obtained in the previous step. The sliding window size is n × n × n (here, it is 3 × 3 × 3). For each sliding window, a specific set of anchors is generated. In this study, each voxel of the feature map corresponds to two scales of anchor boxes. The feature maps are fed into the anchor boxes classifier and regressor, respectively, after passing through the FC layer. The classifier is used to learn the probability value (p) of foreground or background, and the anchor boxes with nodes are classified as foreground and others as background. The regressor is used to learn the offset (x, y, z, d) of the foreground boxes, the first three indicate the coordinates of the proposed boxes in the region, and d denotes the diameter. The loss functions for classification and regression are defined in the loss function section. In the ROI alignment step, several anchor boxes are generated, and anchor boxes of different sizes correspond to different scales of detection target regions because the nodule size is highly variable and a single layer of feature layers lacks semantic or structural information, so anchors of different scales are assigned on multiple layers. To make the size of the Anchor fit the nodular scale distribution as much as possible to make anchor size fit nodule scale distribution as much as possible, we refer to Zhang *et al.* [28] to calculate the distribution results of nodule size and, combined with prior clinical knowledge, assign anchors of different scales to the four scale feature maps after the fusion of adaptive feature pyramid. Following the principle that high-level low-resolution features contain highly abstract linguistic features, which are beneficial for detecting large scale targets, and low-level high-resolution features contain more information on location, boundary, and morphology, which are suitable for detection of small-scale nodules, each feature map corresponds to two-scale anchors. The results are shown in Table 1.

TABLE 1. Feature map corresponding to anchors size parameters.

Future map(D×H×W@C)	Anchors size(L×W×H)
64×64×64@64	5×5×5,10×10×10
32×32×32@64	12×12×12,15×15×15
16×16×16@64	20×20×20,22×22×22,
8×8×8@64	25×25×25,30×30×30,

The positive and negative samples are classified by calculating Intersection over Union (IoU) for each candidate region. If the IoU ratio of the candidate region to the labeled is greater than 0.5, it is classified as a positive sample. If IoU < 0.02, it is determined that the candidate region does not contain nodules and is classified as a negative sample, and regions with IoU between 0.02 and 0.5 are not involved in the training process.

D. ROI ALIGN CLASSIFICATION NETWORK

The ROI module classifies the proposed region suggestions while fine-tuning the regression parameters of the nodal bounding boxes. The proposed region suggestion box is mapped to feature mapping of the same size by the ROI Align method. The feature layer is mapped to the feature vector by two layers of FC. Finally, the proposed region suggestion box is precisely regressed with offset (x, y, z) and diameter d using the bounding box regressor, respectively, and whether the proposed region contains nodules is predicted by the classifier. The prediction probability p is output according to the prediction result. The object in the proposed region is further precisely adjusted, whether it is a pulmonary nodule or not and the coordinate position of the nodule.

E. LOSS FUNCTION

Candidate nodule detection and nodule classification prediction share a 3D residual network, and the lung nodule detection model is a multi-task learning model.Our loss function is composed of classification loss probability score p for the anchor box and regression loss for nodule coordinate (x, y, z), and nodule size d.

The total loss of the model is described as:

$$L_{loss} = L_{cls} + L_{reg} \tag{8}$$

Nodule classification is a binary problem, but the positive and negative samples in lung nodule data are extremely unbalanced, and the heterogeneity of nodules leads to different difficulty of nodule detection. The Focal Loss function [17] is able to improve the imbalance of the samples better by introducing the balance factor and modulation coefficient. Therefore, the algorithm in this paper chooses the Focal Loss function to calculate the classification loss and avoid the weakness of positive and negative samples, which is defined as follows:

$$L_{cls} = -\alpha(1-p)^{\gamma} log_2\,(p) \tag{9}$$

where, p is defined as prediction probability for binary classification, p is the probability of the class with label 1 estimated by the model, α is a balanced weighting factor, which is used to balance the loss of positive and negative samples in the retraining process. γ is a tunable focusing parameter and $(1-p)^{\gamma}$ is modulation coefficient, which is used to control the weights of the difficult samples.

For L_{reg}, we used smooth L1-norm regression loss function [34], which is defined as:

$$L_{reg} = smooth_{L1}(t,\hat{t}) = \begin{cases} \left|t-\hat{t}\right| -0.5 & if\ \left|t-\hat{t}\right| > 1 \\ 0.5\left(t-\hat{t}\right)^2 & else \end{cases} \tag{10}$$

where t is the offset of the ground truth box relative to the anchor i, and $\hat{t}$ is the predicted value of the same position, are given by:

$$\mathrm{t} = \left(\frac{x-x_a}{d_a}, \frac{y-y_a}{d_a}, \frac{z-z_a}{d_a}, \log\left(\frac{d}{d_a}\right)\right) \tag{11}$$

where (x, y, z, d) are the predicted nodule coordinates and diameter in the original space, $\left(x_{a,}y_a, z_a, d_a\right)$ are the coordinates and scale for the anchor i.

$$\hat{t} = \left(\frac{\hat{x}-x_a}{d_a}, \frac{\hat{y}-y_a}{d_a}, \frac{\hat{z}-z_a}{d_a}, \log\left(\frac{\hat{d}}{d_a}\right)\right) \tag{12}$$

where $(\hat{x}, \hat{y}, \hat{z}, \hat{d})$ are nodule ground truth coordinates and diameter.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT DESIGN

The experiments were conducted using Ubuntu 20.04. The network model was developed using PyTorch 3.8, and the experiments were carried out on an NVIDIA GeForce RTX 3080 graphics card with 24 GB of video memory through the PyTorch parallel computing framework. The Batch Size parameter was set to 8, and the number of model iterations was 150. the fit rate of each parameter was adaptively adjusted using the Adam optimizer. The initial learning rate is 0.01, which decays to 0.005 at 38 iterations, 0.001 at 80 iterations, 0.0005 at 133 iterations, and 0.0001 at 142 iterations. Due to GPU capacity limitation, 128 × 128 × 128.1 (D × H × W@C) patches were cropped from CT as the input to the network. To prevent overfitting of the model, a scaling factor between [0.75, 1.25] and random inversion and rotation were used to enhance the data. In the testing phase, the lung CT was cropped to 192 × 192 × 192.1 patch to include all nodes as much as possible, and the segmentation blocks were cropped so that 24 pixels were overlapped between them to eliminate boundary effects during the convolution calculation. In the whole training and testing, the method of cross-validation was adopted. Among the ten subsets of the LUN16 data set, nine groups were randomly selected as the training set and the remaining one group as the test set. After ten cycles, the results of all test sets were summarized for analysis.

B. EVALUATION METRICS

In medical image recognition, three metrics are commonly used to evaluate the detection performance of a system. They are sensitivity, free Receiver Operating characteristics and the competition performance metric. In this study, three metrics are also selected to evaluate the effectiveness and performance of the proposed model.

Sensitivity: the percentage of true positive lung nodules detected, defined as:

$$\mathrm{Sen} = \frac{TP}{TP+FN} \tag{13}$$

where TP represents the number of correctly predicted true nodules, FN is the number of predicted non-nodules that are true nodules.

Free Receiver Operating Characteristic (FROC): FROC curve reflects the decreasing trend of false-positive nodules. The horizontal coordinate is the false positive rate point in each CT slice, noted as FPs/scan, FPs/scan ∈ (0.125, 0.25, 0.5, 1, 2, 4, 8)

The competition performance metric (CPM): CPM is the average sensitivity at seven predefined false positive rates. In this paper, the seven predefined false positive rates are defined as (0.125, 0.25, 0.5, 1, 2, 4, 8). The higher the CPM score, the better the system performance.

C. ABLATION STUDY

To verify the performance and effectiveness of the proposed lung nodule detection model, we design a series of ablation experiments in terms of embedded attentional feature extraction, AFF, improved residual structure and loss function, respectively.

1) COMPARISON OF DIFFERENT ATTENTION MODULES

Channel-SA is a lightweight module embedded in any CNN network to guide the network to focus on essential features and regions of interest. The attention module can be split, and channel and SA can be used individually or in series in some order of priority. To verify the effect of different

TABLE 2. CPM scores of different attention mechanisms.

Attention module	Avg.FP rate							
	0.125	0.25	0.5	1	2	4	8	CMP
SA	0.783	0.812	0.823	0.857	0.918	0.929	0.943	0.866
CA	0.757	0.791	0.835	0.872	0.927	0.943	0.964	0.870
SCA	0.781	0.823	0.847	0.916	0.948	0.954	0.978	0.892
CSA	0.813	0.846	0.897	0.923	0.951	0.967	0.984	0.912

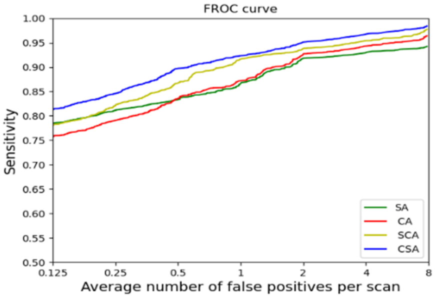

FIGURE 6. FROC of different attention mechanisms.

attention combinations embedded in 3DResNet on detection performance, different combinations of attention modules are replaced with substitutions while other hyperparameters remain unchanged. We designed four comparative experiments in which four attention structures were inserted into the same position of the residual block, respectively. They are separate SA, separate CA, spatial-channel sequential attention, and CSA. At the same time, we embed the same attention module onto the multi-scale feature fusion path. We compared the performance of four possible arrangements using the CMP and FROC metrics. The results are shown in Table 2, Figure 6. The CSA arrangement is the best. This also coincides with the validation results of CBAM proposed in the paper [29]. Channels first adaptively adjust the weights of each channel, aggregating the space and extracting key features. SA focuses on regions of interest and extracts features at a finer granularity at the pixel-level for the channel results. Thus, CSA effectively helps information flow globally in the network and further refines the feature mapping and guides new work in learning more contextual information about the nodal feature.

2) ADAPTIVE MULTI-SCALE FEATURE FUSION PERFORMANCE

TABLE 3. CPM scores of AFF and no-AFF.

Method	Avg.FP rate							
	0.125	0.25	0.5	1	2	4	8	CMP
FPN	0.746	0.813	0.865	0.918	0.937	0.952	0.978	0.887
FPN+CSA	0.813	0.846	0.897	0.923	0.951	0.967	0.984	0.912

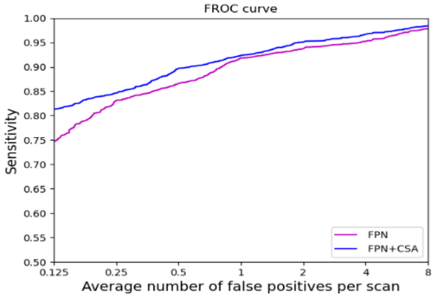

FIGURE 7. FROC of detection performance of attention-embedded FPNs.

TABLE 4. Comparison of different FL parameters.

gamma (γ)	Alpha(α)	Average CPM	Sen
0	0.2	0.836	0.920
1	0.2	0.843	0.942
2	0.2	0.828	0.918
0	0.65	0.832	0.921
1	0.65	0.839	0.932
2	**0.65**	**0.912**	**0.977**
0	0.9	0.847	0.965
1	0.9	0.838	0.929
2	0.9	0.823	0.918
0	1	0.844	0.958

To verify the effectiveness of adaptive multi-scale feature fusion, we compare the attention network embedded in channel-spatial order (FPN+CSA) with the original feature fusion network without embedded attention (FPN). However, other hyperparameters remain unchanged. The feature extraction network is embedded with channel-space attributes of attention. The results are shown in Fig. 7, Table 2. Feature fusion with embedded attention can extract more fine-grained contextual information about the nodes. The number of candidate nodules is reduced, which improves the detection performance of pulmonary nodules. This indicates that the attention mechanism dynamically adjusts the feature fusion, enhances the useful nodule feature representation and suppresses the useless information so higher detection sensitivity.

3) EFFECT OF FOCAL LOSS

Focal Loss is applied to solve the positive and negative sample imbalance and model skewing problems. There are two hyperparameters of FL in formula (7). Therefore, we let γ and α take a series of values for cross verification and compare the experiment with the cross-entropy loss function (i.e., r = 0, a = 1). The CPM score was utilized to evaluate the different combinations of the parameters. Table 5 lists the CPM scores. Experimental results reveal that after using focal loss to replace the cross-entropy loss function, CMP performance is better, indicating that FL effectively reduces the unbalance between nodules and non-nodules. Specifically, in the multi-parameter cross experiment, when $\gamma = 2$ and $\alpha = 0.65$, the model detection effect is best, so this parameter is selected as the loss function. To further verify the validity of focal loss, we select a 3D Faster R-CNN model using the cross-entropy loss function to calculate the loss for comparison experiments. Fig.8 shows that this paper's module's convergence speed and effect are significantly better than the 3D faster. Therefore, the algorithm proposed effectively improves the detection accuracy with reduced computational cost.

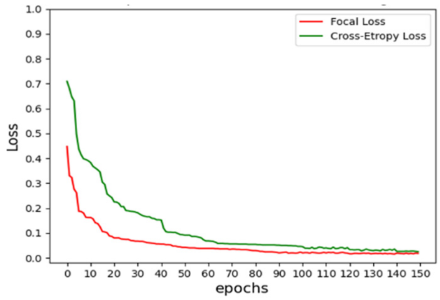

FIGURE 8. Convergence of different classification loss functions.

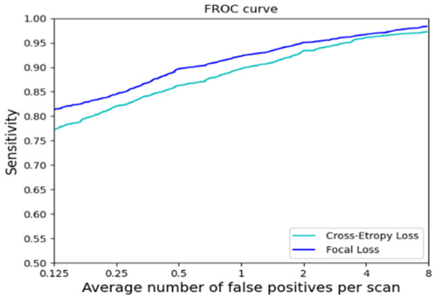

FIGURE 9. Influence of different classification loss functions on detection performance.

TABLE 5. Effect of different extraction structures on the model.

Backbone Net	Sen (%)	CPM (%)	Inferred time(ms/scan)
VGG16	90.13	83.23	191
ResNet18	94.32	89.69	124
ResNet50	95.02	88.41	158
ours	97.7	91.2	131

4) COMPARISON OF DIFFERENT FEATURE EXTRACTION STRUCTURES

To verify the effectiveness of deeply-separable convolution in the proposed model, VGG16, ResNet18 and ResNet50 were selected as the backbone feature extraction networks under the condition that other hyperparameters remained unchanged. Table 5 shows the average CPM score, sensitivity (Sen) and inference time. As can be seen from Table 6, among the three relay networks, ResNet18 has the shortest inference time, and ResNet50 has the best performance. The sensitivity and CMP of AFF, CBAM and DSC are increased by 3% and 1%∼2%, respectively, compared with the original ResNet50 backbone network. Inference time is second only to ResNet18. Therefore, the change of residual structure by depth-separable convolution ensures the sensitivity and accuracy of detection and reduces the parameters and computational effort of the model.

5) NODULE DETECTION PERFORMANCE AT DIFFERENT SCALES AND TYPES

In this part, we choose a 3D Faster R-CNN (proposed by zhu *et al.*) [14] based lung nodule detection algorithm as a baseline and compare nodule detection performance at different scales and the detection effect of different types of nodules for analysis. In the LUN16 dataset, there are 1186 nodes, including 272 nodes for 3 ∼ 5mm, 633 nodes for 5 ∼ 10mm, 231 nodes for 10 ∼ 20mm, and 50 nodes for greater than 20mm, and 888 CT scans. The existing algorithms mainly detect large nodes with high efficiency, while small nodes are easily ignored. Table 6 compares the detection results of nodules of different sizes by the model proposed in this paper and the basic algorithm 3D Faster R-CNN [14]. The experimental results demonstrate that the model can effectively improve the efficiency of small nodule detection. Compared with the traditional base algorithm, this paper's proposed method of embedding attention structure can extract richer global and local features.

Four typical pulmonary nodules are selected to compare their detection performance. The result is shown in Fig.10. The first row is the gold standard map of pulmonary nodules. The second and third-row are the prediction results of different types of nodules by the baseline algorithm and the algorithm in this paper, respectively. The confidence of predicted nodules was compared. Solid nodules are relatively easy to detect, and the detection results of the two algorithms are similar. Ground glass nodules are irregular in shape, similar to surrounding tissues, and nodules near the vascular and lung wall are located in special locations, so it is challenging to detect such nodules. The detection accuracy of the algorithm proposed in this paper is higher than that of the benchmark algorithm. By comprehensive comparison, it can be seen that the algorithm in this paper has better detection results for nodules of different sizes and types and has a high confidence level.

6) COMPARISON WITH EXISTING MODELS

We compared the proposed network with some advanced models based on deep learning. The results are listed in Table 7. Setio *et al.* [12] and Dou *et al.* [10] detected pulmonary nodules on multiple 2D CTs, which did not take full advantage of the volumetric information of 3D pulmonary nodules to extract spatial contextual information of 3D pulmonary nodule features. Therefore, the sensitivity of detection was low. Zhu *et al.* [14] chose to detect pulmonary nodules on 3D CT images, which were able to extract rich contextual information. However, it did not focus on the extraction of spatial contextual features at multiple scales and different receptive fields, which would lead to the lack of representativeness of the extracted features and thus affect the sensitivity of pulmonary nodule detection, especially for small nodule detection. Zhang *et al.* [28] and Yu *et al.* [26] adaptively extract nodule-rich contextual features by embedding SE into the 3D CNN, and the SE module selectively emphasizes salient features while suppressing nonsignificant features. As a result, representative features of lung nodules can be better captured and utilized. However, The SE module unilaterally focuses on the correlation of channels without attaching importance to positional information. However, SA is vital in determining "where" is the focus. The internal environment of the lung is complex, with nodules of

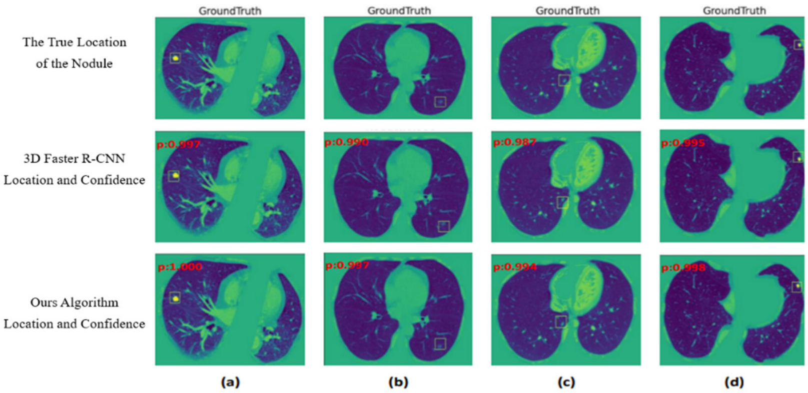

FIGURE 10. Effectiveness of detection of different types of nodules: (a) Isolated Nodule, (b) Ground glass nodule, (c) Near vascular nodule, (d) Near lung wall nodule.

TABLE 6. Effectiveness of nodule detection at different scale sizes.

Method	diameter(3~5mm)		diameter(5~10mm)		diameter(10~20mm)		diameter(>20mm)		Total number of detected	Sen (%)
	Number	Detection rate (%)	Number	Detection rate (%)	Number	Detection rate (%)	Number	Detection rate (%)		
Zhu et al.(Res18)[14]	251	92.2	604	95.4	220	95.2	47	94.0	1 122	94.6
ours	266	97.8	618	97.6	226	97.8	49	98.0	1 157	97.7

various sizes and shapes. The SE module will inevitably lose information on nodal structure, resulting in missed or false positives. We propose a 3D-based two-stage CNN model for lung nodule detection. The 3Dpatch as input makes full use of the contextual spatial information of the nodules. Embedding CSA to 3D ResNet guides the network to adaptively adjust the weights along the channel and spatial directions, respectively, to extract the salient features of nodules. The attention is embedded in a multi-scale feature fusion network, which guides the network to adaptively, pixel-level, a fine-grained fusion of features for feature maps of different resolutions to extract nodule detail features in complex spatial locations effectively. AFF achieves multi-scale nodal target detection performance, facilitating small nodules and nodules detection into the blood vessels and near the lung wall. A depth-separable convolution replaces the standard convolution in the original residual block to improve model training efficiency and generalization ability while ensuring model accuracy. Overall, the proposed 3D detection model can detect multi-scale lung nodules of different shapes, and its performance is significantly better than some advanced models based on deep learning.

V. DISCUSSION

In this study, a new 3D CNN network with attention-guided feature extraction and multi-scale feature fusion is proposed for lung nodule detection. Lung nodules are variable in size and morphology, and some nodules are similar to the surrounding tissues where characteristics are not significant and are more difficult to detect. Therefore, we chose the two-stage target detection model 3D Faster R-CNN to detect lung nodules. Faster R-CNN is the most classical model for two-stage target detection, and the detection accuracy is better than that of the one-stage model. Suitable for complex and variable lung nodule detection. We select 3D ResNet as the backbone network for feature extraction to achieve feature reuse, prevent gradient explosion/disappearance in the deep network, and improve the model's generalization ability. ResNet enables local feature reuse. However, it fails to establish the mapping of global features of nodules. As the network deepens, the flow of shallow features is hindered, which results in the loss of a large amount of small target information. In ResNet, CSA attention is introduced to guide the network to dynamically adjust feature weights along with both channel and space directions, respectively, to enhance useful feature information of nodes, and inhibit the irrelevant features, which effectively extracts global contextual information of nodes. In FPN, CSA is embedded to adaptively fuse features from different layers, changing the original simple fusion method to achieve fine-grained fusion of features at the pixel-level, which retains more structural and detailed information about nodules, and extracts a feature of representative nodules and improves the sensitivity of nodule detection.

In this study, to evaluate the performance of the proposed model, many ablation experiments were performed. The channel-SA module can be embedded in parallel or serially into the CNN. We selected four arrangements for comparison experiments, and the results showed that the channel-space sequential structure has the highest average detection sensitivity. The results in Table 5 show that our model structure has higher sensitivity and CMP scores than ResNet50 without the embedded CSA structure, which indicates the effectiveness of the attention module. FIG. 7 and Table 3 compare

TABLE 7. Comparison of detection performance of pulmonary nodules.

Method	Avg. FP rate							
	0.125	0.25	0.5	1	2	4	8	CMP
Zhu et al.(Res18)[14]	0.622	0.746	0.815	0.864	0.902	0.918	0.932	0.834
Zhu et al.(DPN26)[14]	0.692	0.769	0.824	0.865	0.893	0.917	0.933	0.842
setio et al.[12]	0.692	0.771	0.809	0.863	0.859	0.914	0.923	0.838
CHI[35]	0.677	0.791	0.865	0.930	0.968	0.971	0.971	0.882
Zhang et al.[28]	0.754	0.848	0.878	0.918	0.946	0.948	0.971	0.893
Dou et al. [10]	0.659	0.747	0.819	0.865	0.906	0.933	0.946	0.839
Yu et al.[26]	0.697	0.789	0.843	0.891	0.920	0.940	0.947	0.861
Ours	0.813	0.846	0.897	0.923	0.951	0.967	0.984	0.912

the influence of attention-guided feature adaptive fusion on model performance. The results show that the feature fusion approach with embedded attention improves the model's performance significantly, with an increase in CMP score close to 0.025. With attention-guided feature fusion in different layers, the network can learn more representative nodal features. The depth-separated convolution significantly improves the training efficiency of the model due to its fewer parameters and lower complexity. We use the focal loss function to calculate the classification loss, balance the positive and negative samples, and select the ideal hyperparameters through experiments. According to the experimental results, focal loss makes the trained model converge faster and better, and the average sensitivity of detection is also higher than that of the model with the applied cross-entropy loss function. We compared the detection performance of the baseline lung nodule detection model for different scales of lung nodules (shown in Table 7). Based on the evaluation of the experimental results and theoretical analysis, the proposed model is a valid model for nodule detection, which is helpful for others' studies and obtained a CMP score of 0.912. Satisfactory results were obtained.

VI. CONCLUSION

After studying and analyzing the available literature references, we proposed a 3D Faster R-CNN pulmonary nodules detection model embedded with attention mechanism and adaptive feature pyramid. The channel and spatial attention modules are embedded in ResNet to guide the network to adaptively adjust the correlation between channels and highlight important feature regions, both of which form complementary mechanisms for feature extraction and enhance the flow of shallow features. Deeply separable convolution replaces the 3D convolutional blocks in ResNet, effectively reducing the computational effort while ensuring the accuracy of the model. The results of a large number of ablation experiments show that the improved 3D ResNet serves as the backbone network for feature extraction to effectively extract representative global contextual information of nodules. We develop an attention-guided adaptive feature fusion pyramid network that refines the pixel-value weights of feature extraction to improve the fine-grained fusion of pixel-level feature information, reduce the redundancy of fused information, and enhance the feature representation of FPNs. A series of comparative experiments were performed on the LUNA dataset, and the results showed that the method achieved high sensitivity in each of the first four low FPs/scans and could outstandingly achieve accurate detection of lung nodules.

However, the algorithm still has some limitations. The detection accuracy has improved but still has not reached the desired height. The sensitivity of a low false-positive rate is not high. Focal loss can improve data imbalance but cannot eliminate the effect of the imbalanced data set. Although appropriate measures are taken to enhance the data, more data volume is required to generalize the model capability with the deepening of the network. The next step can be considered to study how to take the generative adversarial network to expand the positive sample set or combine the electronic medical record information to increase the attributes of the samples, thus improving the nodule detection performance.

REFERENCES

[1] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li, and W. Chen, "Cancer statistics in China and United States, 2022: Profiles, trends, and determinants," *Chin. Med. J.*, vol. 135, no. 5, pp. 584–590, Mar. 2022.

[2] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray, "Cancer statistics for the year 2020: An overview," *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, Apr. 2021.

[3] R. Sharma, "Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050," *Int. J. Clin. Oncol.*, vol. 27, no. 4, pp. 665–675, Apr. 2022.

[4] J. S. Rhodes, T. R. P. Ford, J. A. Lynch, P. J. Liepins, and R. V. Curtis, "Micro-computed tomography: A new tool for experimental endodontology," *Int. Endodontic J.*, vol. 32, no. 3, pp. 165–170, May 1999.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, Amsterdam, The Netherlands, 2016, pp. 21–37.

[8] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, Montreal, QC, Canada, Dec. 2016, pp. 1–9.

[9] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *Proc. MICCAI*, Sep. 2017, pp. 559–567.

[10] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, ''Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection,'' *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[11] L. Gong, S. Jiang, Z. Yang, G. Zhang, and L. Wang, ''Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks,'' *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 11, pp. 1969–1979, Nov. 2019.

[12] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, and C. Jacobs, ''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.

[13] X. Hongtao, D. Yang, N. Sun, Z. Chen, and Y. Zhang, ''Automated pulmonary nodule detection in CT images using deep convolutional neural networks,'' *Pattern Recognit.*, vol. 85, pp. 109–119, Jan. 2019.

[14] W. Zhu, C. Liu, W. Fan, and X. Xie, ''DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 673–681.

[15] F. Ciompi, K. Chung, S. J. van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. W. Wille, A. Marchiano, U. Pastorino, M. Prokop, and B. van Ginneken, ''Towards automatic pulmonary nodule management in lung cancer screening with deep learning,'' *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Apr. 2017.

[16] 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.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2016, pp. 779–788.

[19] 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.

[20] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, ''Mask R-CNN,'' in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[21] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Med. Image Comput. Comput.-Assisted Intervent.*, Munich, Germany, 2015, pp. 234–241.

[22] S. Zheng, J. Guo, X. Cui, and R. N. J. Veldhuis, ''Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection,'' *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 797–805, Mar. 2020.

[23] R. Fu, C. Zhang, T. Zhang, X.-P. Chu, W.-F. Tang, X.-N. Yang, M.-P. Huang, J. Zhuang, Y.-L. Wu, and W.-Z. Zhong, ''A three-dimensional printing navigational template combined with mixed reality technique for localizing pulmonary nodules,'' *Interact. CardioVascular Thoracic Surgery*, vol. 32, no. 4, pp. 552–559, Apr. 2021.

[24] S. Tang, M. Yang, and J. Bai, ''Detection of pulmonary nodules based on a multiscale feature 3D U-Net convolutional neural network of transfer learning,'' *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0235672.

[25] 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, Jun. 2020.

[26] H. Yuan, Y. Wu, J. Cheng, Z. Fan, and Z. Zeng, ''Pulmonary nodule detection using 3-D residual U-Net oriented context-guided attention and multi-branch classification network,'' *IEEE Access*, vol. 10, pp. 82–98, 2022.

[27] H. Zhang, Y. Peng, and Y. Guo, ''Pulmonary nodules detection based on multi-scale attention networks,'' *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, Jan. 2022.

[28] M. Zhang, Z. Kong, W. Zhu, F. Yan, and C. Xie, ''Pulmonary nodule detection based on 3D feature pyramid network with incorporated squeeze-and-excitation-attention mechanism,'' *Concurrency Comput., Pract. Exper.*, Mar. 2021, Art. no. e6237, doi: 10.1002/cpe.6237.

[29] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.

[30] W. Wang, J. Luo, X. Yang, and H. Lin, ''Data analysis of the lung imaging database consortium and image database resource initiative,'' *Acad. Radiol.*, vol. 22, no. 4, pp. 488–495, Apr. 2015.

[31] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, and E. A. Kazerooni, ''The lung image database consortium, (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans,'' *Med. Phys.*, vol. 38, no. 2, pp. 915–931, Feb. 2011.

[32] A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, and C. van den Bogaard, ''Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge,'' *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.

[33] H. W. Zhang and H. Zhang, ''LungSeek: 3D Selective kernel residual network for pulmonary nodule diagnosis,'' in *The Visual Computer*. 2022, pp. 1432–2315, doi: 10.1007/s00371-021-02366-1.

[34] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Dec. 2015, pp. 2961–2969.

[35] C. C. Nguyen, G. S. Tran, V. T. Nguyen, J.-C. Burie, and T. P. Nghiem, ''Pulmonary nodule detection based on faster R-CNN with adaptive anchor box,'' *IEEE Access*, vol. 9, pp. 154740–154751, 2021.

GUANGLU ZHANG received the B.S. degree in computer science and applications from the Henan University of Telecommunications, in 2002, and the M.S. degree in communication and information systems from Hainan University, in 2009. She is currently pursuing the Ph.D. degree in software engineering with the Army Engineering University of PLA. Her research interests include the area of data mining, computer vision, medical image analysis, and processing technology.

HONGJUN ZHANG was born in 1963. He is currently a Professor and a Ph.D. Supervisor. His research interests include data engineering and military modeling and simulation.

YUHUA YAO received the Ph.D. degree in computational mathematics from the Dalian University of Technology, China, in 2006. From 2006 to 2017, he held a Professorship with the College of Life Sciences, Zhejiang Sci-Tech University. He has been a Professor with the School of Mathematics and Statistics, Hainan Normal University, since 2017. His current research interests include computational biology, biological mathematics, and bioinformatics.

QIUHUI SHEN was born in Zhoukou, Henan, China, in 1987. She received the B.S. and M.S. degrees in computer science and technology from Central South University. She is currently pursuing the Ph.D. degree in software engineering with the Army Engineering University of PLA. Since 2018, she has been a Lecturer with the Computer Science and Technology College, Zhoukou Normal University. Her research interests include the area of knowledge graph and cloud computing.