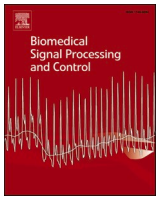
Biomedical Signal Processing and Control 94 (2024) 106318

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



YOLO-MSRF for lung nodule detection

Xiaosheng Wu [a], Hang Zhang [a], Junding Sun [a], Shuihua Wang [a,b], Yudong Zhang [a,c,*]

[a] School of Computer Science and Technology, Henan Polytechnic University, Henan 451460, China
[b] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China
[c] School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, UK

ARTICLE INFO

Keywords:
Lung nodule
Multi-scale receptive field
Object detection
Convolution

ABSTRACT

(Aim) Aiming at the problem that there are a large number of small object nodules that are difficult to detect in lung images, detection methods based on improved YOLOv7 are proposed in this paper. (Method) First, a new small object detection layer (SODL) is proposed to solve the problem of the small size and irregular shape of lung nodules being difficult to detect accurately. Secondly, aiming at the problem that the characteristics of lung nodules are blurred and difficult to detect due to the continuous downsampling of the model, a multi-scale receptive field (MSRF) module is proposed and designed to improve the model's extraction of channel features. Finally, efficient omni-dimensional convolution (EODConv) is used to improve the ability of the network to extract the space, filters, and channels of the convolution kernel. (Results) Experiments were carried out on the public Luna16 dataset, and the results showed that our mAP, precision, and recall rate reached 95.26 %, 95.41 %, and 94.02 %, respectively, surpassing many state-of-the-art models. (Conclusion) In this study, a YOLOv7-based method is proposed for detecting lung nodules. Experimental results show that the proposed modification can significantly improve detection performance and is more suitable for clinical medical diagnosis.

1. Introduction

According to the statistical data released by the National Cancer Center in 2022, lung cancer has become the most common and dangerous disease. Early symptoms of lung cancer are common, such as coughing and blood-stained sputum, which are difficult to distinguish from other diseases. Lung cancer develops rapidly, taking only a few months from the early stage to the late stage. Without timely computer tomography (CT) scans, early-stage lung cancer patients may develop late-stage lung cancer with a low survival rate. Nodules are characteristic of early-stage lung cancer, so detecting nodules is crucial for preventing lung cancer [1].

Under the present healthcare systems, hospitals adopt CT scans to screen for lung nodules [2]. Radiologists can only accurately identify the location of suspicious lung nodules after analyzing hundreds of CT images because of the small size of lung nodules and the surrounding circular and white shadows in CT images. Moreover, finding the true lung cancer nodules among the suspicious ones is challenging and time-consuming for radiologists. Therefore, developing an automatic lung nodule detection system helps radiologists make judgments, save time, and help hospitals and society save money while assisting patients in better preventing lung cancer.

Among the lung nodule detection methods, two main methods extract suspicious nodules. The first category is based on traditional methods of machine learning, such as thresholding [3], support vector machines [4], morphology [5], and clustering [6]. Traditional methods usually require the manual extraction of features from nodules. After years of research, investigators have found that traditional detection methods have many problems. When the extracted features are too single and too specific, the generalization ability of the features is weak, which leads to poor robustness in practical applications and makes it difficult to obtain remarkable detection results. Therefore, traditional methods have significant limitations.

The second category of approaches is based on deep learning, mainly using convolutional neural networks (CNNs), which can automatically learn high-level semantic information from the input images and have demonstrated powerful feature extraction capabilities in classification, object detection, and segmentation tasks in recent years. This approach has been used in many fields, including agriculture [7,8], industry, manufacturing, and medicine. In medical nodule detection, using CNNs

* Corresponding author at: School of Computer Science and Technology, Henan Polytechnic University, Henan 451460, China.
E-mail addresses: wuxs@hpu.edu.cn (X. Wu), zh@home.hpu.edu.cn (H. Zhang), sunjd@hpu.edu.cn (J. Sun), shuihuawang@ieee.org (S. Wang), yudongzhang@ieee.org (Y. Zhang).

https://doi.org/10.1016/j.bspc.2024.106318
Received 12 November 2023; Received in revised form 23 February 2024; Accepted 8 April 2024
Available online 10 April 2024

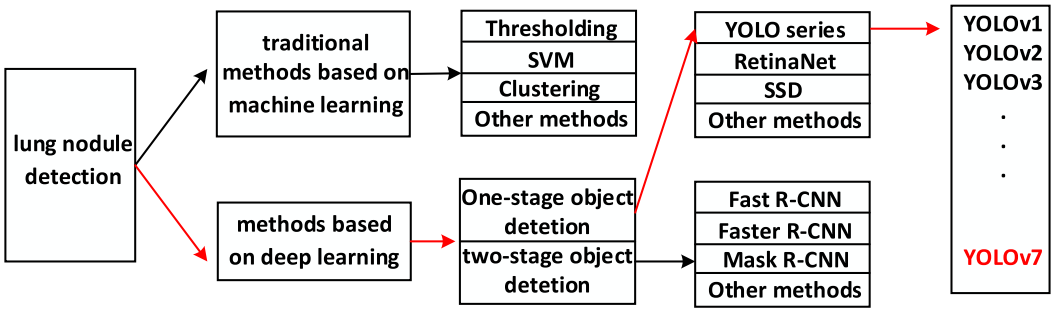


Fig. 1. Taxonomy.

allows for better robustness and generalization by automatically learning the features of lung nodules without relying on manual feature extraction. Moreover, compared to traditional methods that require manual feature design, deep learning methods do not require excessive human intervention or domain knowledge. CNNs can perform nodule detection tasks more efficiently. Therefore, deep learning-based methods can achieve better object detection results.

When using deep learning methods for object detection, the detection algorithms can be divided into two-stage and one-stage detection algorithms depending on whether it is necessary to generate candidate boxes in advance.

Two-stage algorithms, known as region-based object detection algorithms, include R-CNN [9], Fast R-CNN [10], Faster R-CNN [11], Mask R-CNN [12], and others. Ding et al. [13] used VGG16 [14] as the backbone network and extracted features through candidate regions selected by Faster R-CNN. The network also introduced a deconvolution structure and used a 3D deep convolutional neural network (DCNN) to eliminate false positives. Zhang et al. [15] proposed an improved network structure based on Faster R-CNN, which combines the encoder-decoder of U-Net with the multi-scale attention module as the backbone network. Similarly, Guo et al. [16], inspired by Faster R-CNN, added a multi-threshold detector to Cascade R-CNN [17] to increase the number of anchors, which increases the number of positive samples and reduces the impact of positive–negative sample imbalance, achieving better results. Cai et al. [18] used the Mask R-CNN algorithm to detect pulmonary nodules, using ResNet50 as the backbone network and proposing candidate frames with the Region Proposal Network (PRN), achieving an mAP score of 0.882 on the Luna16 dataset.

Although the performance of the above two-stage based network for lung nodule detection is improved, the two-stage algorithm needs to generate candidate frames in advance, and this process is seriously time-consuming, leading to the network's slow overall detection operation. Meanwhile, in the two-stage target detection network, the generation of candidate frames is usually based on the extracted feature maps and some a priori knowledge (e.g., anchor frames), which may lead to inaccurate localisation of candidate frames and the inability to accurately cover the target object for some cases where the shape of the target varies a lot or the size is small.

Recently, many networks about 3D deep learning for detecting lung nodules have been developed, but 3D networks have a long training time. Compared to 2D networks, 3D networks need to deal with three-dimensional data, with a larger number of parameters and higher computational cost. Meanwhile, compared with 2D networks, the design of 3D networks is more complex, and more factors need to be considered, such as network depth, convolution kernel size, step size, and so on. In addition to this 3D networks require higher computational resources for training and inference, which limits their application on some devices. Therefore, this paper uses a 2D One-stage object detection network to detect lung nodules.

One-stage or regression-based object detection algorithms include the YOLO series, SSD [19], RetinaNet [20], and others. The YOLO series can perform end-to-end detection and is much faster than two-stage detection algorithms and many traditional detection algorithms, with the characteristics of rapid detection speed and better universality. The network structure of YOLOv1 [21] is improved based on GoogLeNet by replacing the inception layer with 1×1 and 3×3 convolutions and the core idea is to regard object detection as a regression task. In 2018, Redmon further improved YOLOv2 [22] and proposed YOLOv3 [23]. The network draws on the residual structure of ResNet [24]. YOLOv3 offers DarkNet-53 to make the backbone network deeper (from DarkNet-19 of v2 to DarkNet-53 of v3), which can achieve accuracy comparable to ResNet-101 and ResNet-152. It was one of the algorithms that achieved the best balance between accuracy and speed at that time.

Recently, one-stage algorithms have been widely used in medical imaging and achieved better results. Huang et al. [25] proposed a 3D nodule detection system based on YOLOv3, which combines YOLOv3 with a one-shot aggregation module (OSA) and a receptive field block module (RFB), achieving a competitive performance metric (CPM) of 0.905 on the LIDC dataset. Mei et al. [26] performed a lightweight model with an accuracy of 93.6 % by using YOLOv4 as the backbone network, adding an attention mechanism and focal loss, and pruning the network at the end. Liu et al. [27] designed a nodule detection network based on YOLOv5, which adopted the Bi-FPN feature extraction and stochastic pooling methods and replaced the model loss function, improving the nodule recall rate.

YOLOv7 [28] was the latest one-stage object detection algorithm in 2022. It adopts the Efficient Layer Aggregation Network (ELAN) and the re-parameterization convolutional network. The ELAN controls the gradient path and designs a deeper network for learning the model effectively. Meanwhile, a new label assignment strategy is proposed to optimize gradient descent. Although YOLOv7 performs well in natural image detection, it performs poorly on small object datasets. Therefore, if YOLOv7 is applied to medical images for nodule detection, it is challenging to extract feature information effectively.

Although two-stage or one-stage methods for lung nodule detection have been proposed and have achieved acceptable results, there are still some problems, such as low detection accuracy and the inability to fully utilize the surrounding feature information of nodules. Therefore, a lung nodule detection network is based on YOLOv7. YOLO multi-scale receptive field (YOLO-MSRF) is proposed in this paper:

1) We proposed a small object detection layer (SODL), which focuses more on feature extraction for small objects compared to the other three detection layers, to address the issue of the original network's insufficient capture of feature information from small nodules.
2) We proposed a multi-scale receptive field (MSRF) module that extracts features around the nodules to obtain richer nodule information.
3) We proposed an efficient omni-dimensional convolution (EODConv) model, which enables attention weighting of the input data. This convolution improves accuracy and enhances the extraction of information.

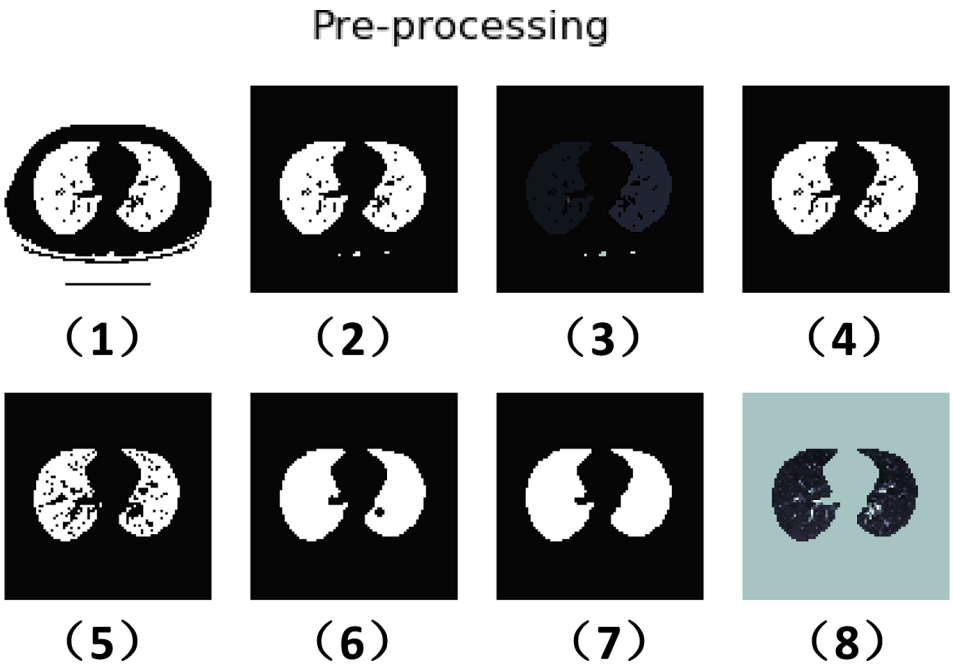


Fig. 2. Image preprocessing.

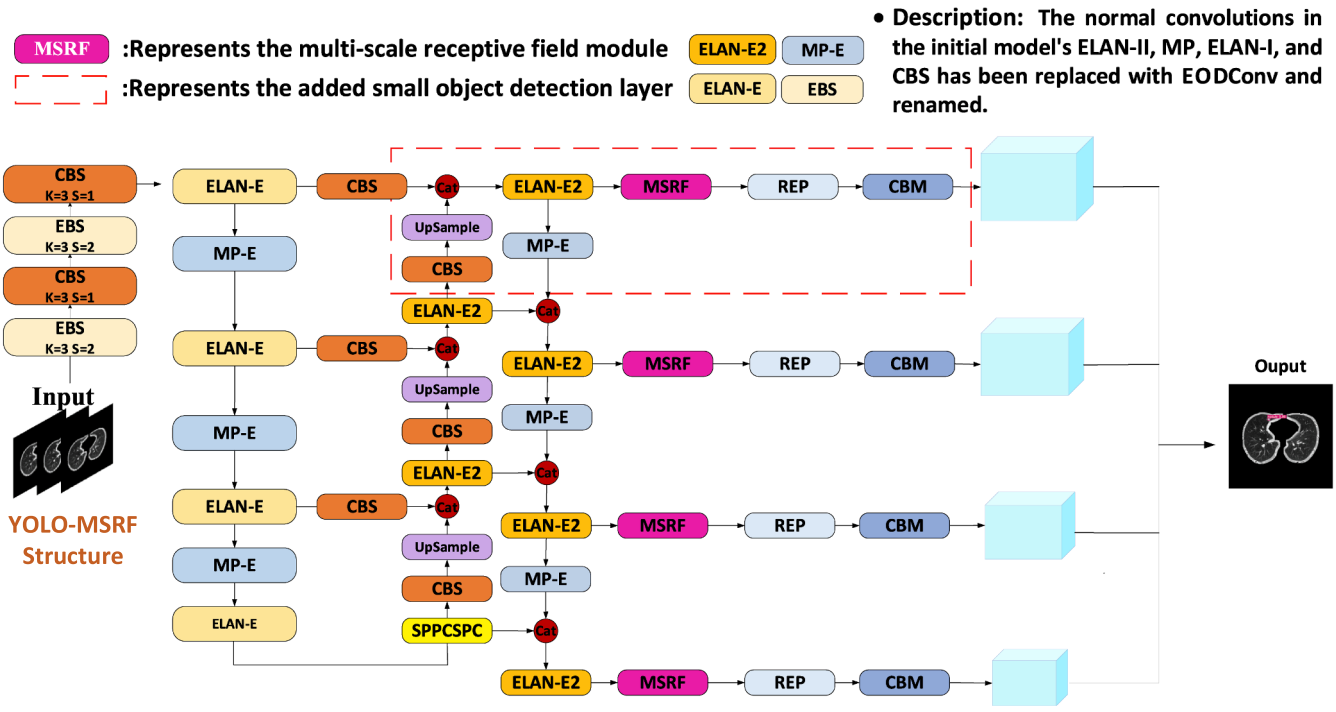


Fig. 3. YOLO-MSRF Structure.

The experimental results are validated on the Luna16 dataset and show that the methods proposed in this paper are effective. The taxonomy is shown in Fig. 1.

The red color represents the direction of this research.

2. Methodology of YOLO-MSRF

2.1. Image preprocessing

To eliminate the interference of extraneous information in the original image and be convenient to feature extraction, it is necessary to preprocess the image before feeding it into the network. Preprocessing provides better-quality images for experiments.

1. first, convert the image to a binarised image, keeping only black and white colours and removing redundant colour interference. 2. remove points connected to the image boundary, removing pixel point information from boundary points. Integrate the information of the region around the lung nodule. 3. identify the left and right lungs, extract the information about the lungs at different locations, and facilitate the subsequent localisation of the lung nodule. 4. Join the two lung regions,

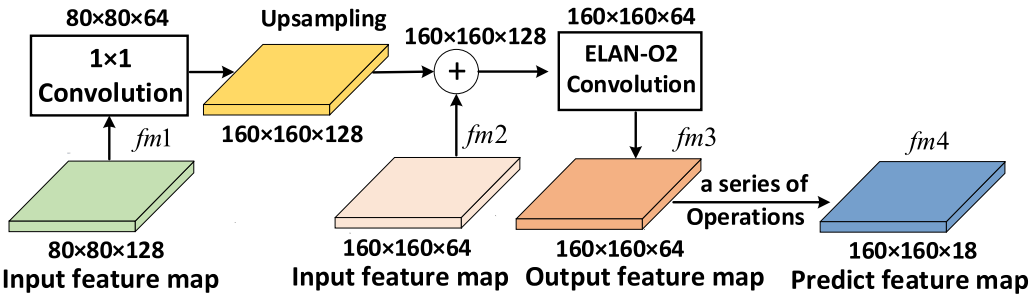


Fig. 4. SODL Structure.

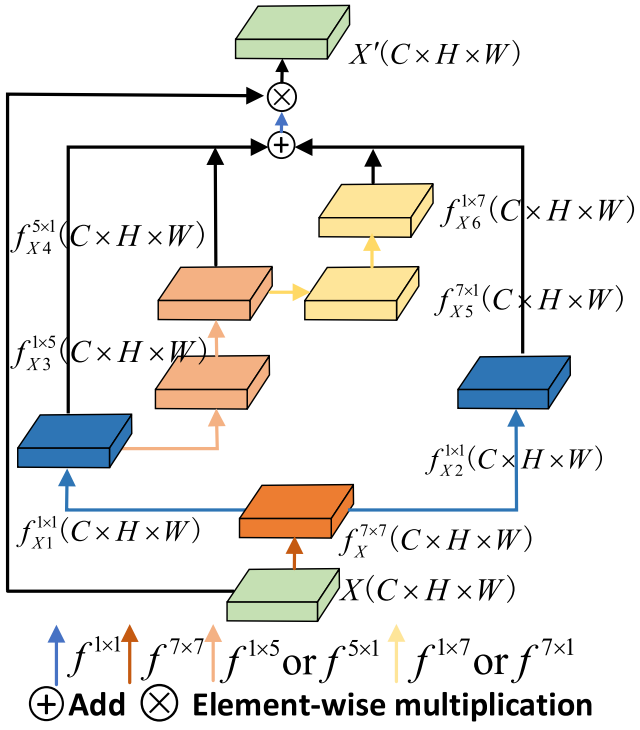


Fig. 5. MSRF Structure.

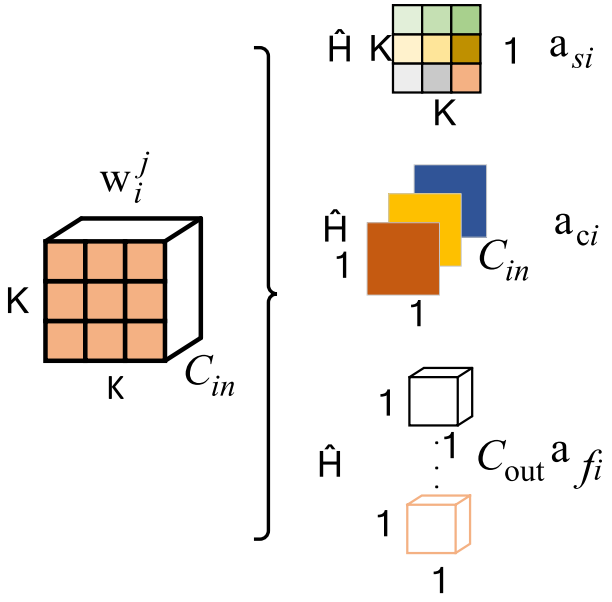


Fig. 6. EODConv structure.

black and white words, the area of the lung region, and keep the maximum shadow area between the two lungs. 5. Erode and separate the tiny nodules attached to the blood vessels, meticulously extract the information of the nodules from the internal regions of the lungs. 6. Expansion operation to fill up the left and right lungs and remove the small lung cavities. 7. Fill the small holes in the lung masks. 8. Superimpose a binary mask onto the input image. The preprocessing process is shown in Fig. 2.

By preprocessing the lung nodule images, the proposed YOLO-MSRF recall and accuracy can be improved, and misdetection and missed detection can be reduced, thus improving the early diagnosis of lung cancer.

2.2. Framework

The structure of the YOLO-MSRF framework is shown in Fig. 3. The overall network adopts a feature pyramid network (FPN) with a path aggregation network (PAN) structure.

First, the nodule images are continuously downsampled, and then the ELAN extracts deeper semantic information. Next, we use the UpSample modules for upsampling. The network learns richer feature information by combining the relative position information acquired from upsampling with the deep semantic information obtained from downsampling. After the final upsampling, SODL is added to enhance the ability to detect small objects. Secondly, the MSRF modules are added to get more prominent receptive field information before obtaining the final output results for each layer. Finally, in the initial modules of the network, such as CBS, ELAN-I, ELAN-II, and MP, we have used the EODConv to replace the regular convolutions so that the network can focus more on marking node region information during the convolution operation.

To facilitate a better understanding of the network structure, we renamed these modules EBS, ELAN-E, ELAN-E2, and MP-E. EBS stands for the convolution operation using EODConv + BN + Silu, and ELAN-E and ELAN-E2 stand for replacing regular convolution operation with EODConv convolution in two different ELAN operations. MP-E represents the addition of the EODConv convolution after Maxpooling. This nomenclature helps to get a clearer picture of the structural composition of the network.

2.3. Proposed SODL

The original application scenario of the YOLOv7 detection algorithm was natural images. Its pre-trained model used images from the COCO dataset, which contains 80 categories of objects, most of which have relatively large shapes. The original YOLOv7 finally acquires only downsampled 8x, 16x, and 32x feature maps, which are mapped back to

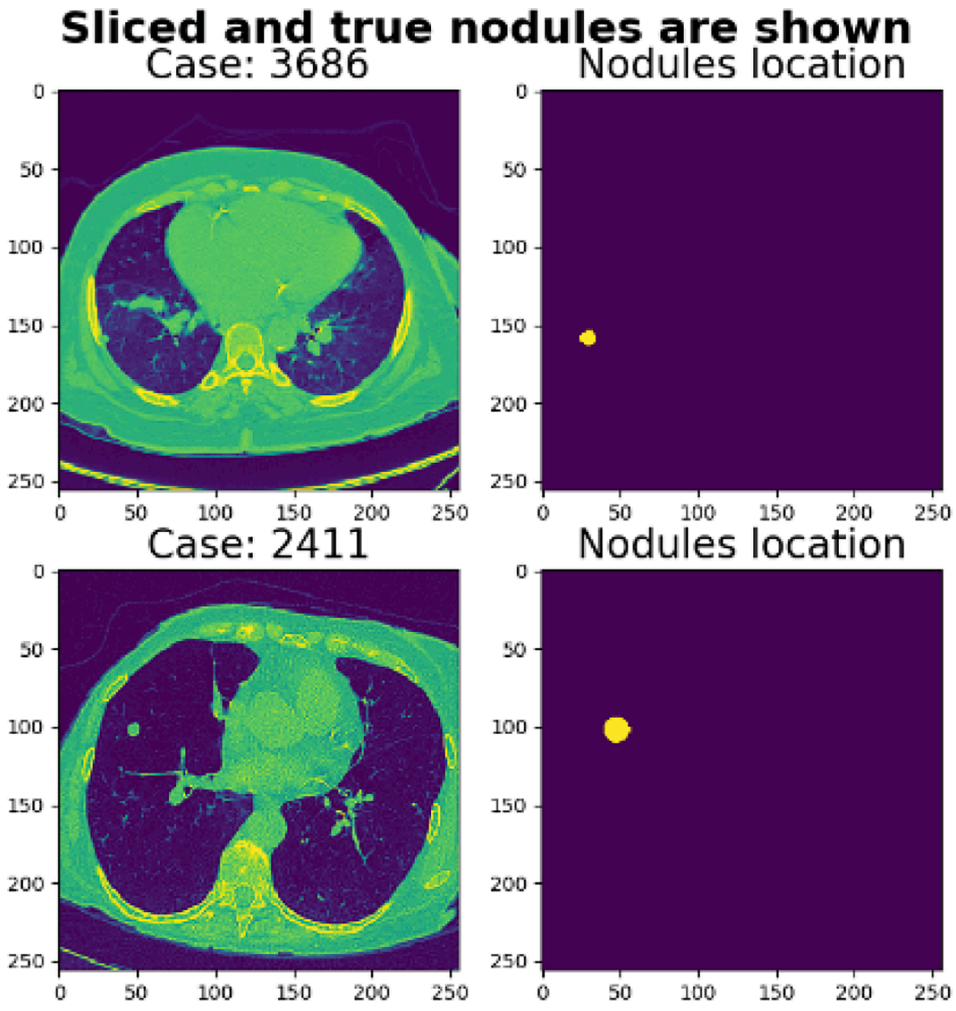


Fig. 7. Original images and GT nodules.

Table 1
Experimental parameter.

Parameter	Value
Weight decay	0.0005
batch size	4
learning rate	0.01
epoch	250

Table 2
YOLO-MSRF ablation experiments.

Algorithm	SODL	EODConv	MSRF	Precision	Recall	mAP
YOLOv7				0.6801	0.6479	0.7053
YOLOv7 + SODL	√			0.7818	0.7226	0.7227
YOLOv7 + EODConv		√		0.7945	0.7065	0.7495
YOLOv7 + MSRF			√	0.7848	0.7058	0.7219
YOLOv7 + EODConv + MSRF		√	√	0.8240	0.7479	0.7883
YOLOv7 + SODL + EODConv	√	√		0.8461	0.7394	0.7721
YOLOv7 + SODL + MSRF	√		√	0.8181	0.7560	0.7572
YOLO-MSRF (Ours)	√	√	√	**0.8567**	**0.8379**	**0.8094**

the original image and represent different regions, with the downsampled 8x feature map being the largest and used to detect small objects—for example, baseballs, toothbrushes, and spoons. When the lung images are input to the network, the shape and size of the lung nodules are much smaller than the small objects in the natural images, and only the downsampled 8x feature map is not sufficient to extract the feature information of the nodules, so it is not desirable to directly detect the nodule dataset with YOLOv7.

For this purpose, SODL is designed to capture the connection between feature maps, reduce the interference of image background information, enhance the accuracy of small objects, and ensure that the network can efficiently identify lung nodule features. The structure of SODL is shown in Fig. 4.

Assuming the network input size is 640 × 640 × 3, after a series of downsampling operations, the feature map *fm*1 with a size of 80 × 80 × 128 is obtained. Several additional steps must be taken for the feature map to contain more context information. First, the channel number of *fm*1 is changed by applying a 1 × 1 convolution to achieve cross-channel information interaction. Then, it is upsampled and restored to the size of 160 × 160 × 64, after which it is summed with the input feature map *fm*2 (*fm*2 is obtained by 4x downsampling of the initial network). The output feature map *fm*3 is obtained via the ELAN-O2 module, and then after convolution and other operations, the final predicted feature map *fm*4 with a size of 160 × 160 × 18 is generated. The predicted feature

Table 3
Comparison of different attention models with MSRF experimental ablation.

Model	mAP	Precision	Recall
CBAM [31]	0.7615	0.7476	0.7310
NAMAttention [32]	0.7356	0.7496	0.6794
GAMAttention [33]	0.7559	0.7886	0.7212
ShuffleAttention [34]	0.7807	0.7417	0.7983
SKAttention [35]	0.7766	0.8299	0.7395
SimAM [36]	0.7549	0.7577	0.7647
SE [36]	0.7735	0.7885	0.7227
CrissCrossAttention [37]	0.7747	0.7271	0.8319
YOLO-MSRF (Ours)	**0.8094**	**0.8567**	**0.8379**

map *fm*4 contains feature information downsampled 4× from the initial network, which means the original map is partitioned into 160 × 160 squares. Each small grid can predict the lung nodules. The more accurate partitioning of the network allows the model to learn more detailed and fine-grained features of the lung nodules.

2.4. Proposed MSRF

With the widespread application of self-attention models, many models in various fields have added these modules and achieved unexpected results. Research involving attention modules has also emerged in medical detection in recent years. These modules can allow the network to focus on areas of interest, accurately locate lesion areas, and provide a more accurate judgment basis for doctors.

To improve nodule detection accuracy further, we used different self-attention mechanisms in YOLO-MSRF. Although we obtained some remarkable results, as the network continues to downsample to extract lung nodule features, the feature maps become smaller and smaller. Therefore, nodule features will also shrink and occupy a relatively small proportion of the feature map. Because attention mechanisms cannot concentrate on very small areas, this leads to an apparent decrease in nodule detection accuracy and the problem of difficult detection and recognition of nodules. Therefore, we need to find other methods to focus on nodule features. We found that the receptive field plays a crucial role in improving nodule detection accuracy. Increasing the receptive field can effectively increase the network's attention to nodule features and improve nodule detection accuracy.

In principle, if the range of the receptive field is too small, the context information that the network can extract is insufficient, which will lead to a lack of learning ability on the part of the network for the object area and prevent it from receiving the change in the object location. Suppose the receptive field area is too large. In that case, too much information is extracted in the training process, and it will contain more redundant data, which is not conducive to the network making a correct judgment and reduces the network's performance, especially when the training sample is limited.

To better detect the region of interest, this paper proposes a model called MSRF, which fuses different scales of receptive field information to allow the model to focus on the region of interest and better capture the surrounding information, thus accurately and efficiently guiding the model to locate the lung nodules.

The structure of the MSRF model proposed in this paper is shown in Fig. 5. It is constructed as follows: First, a 7 × 7 convolution is performed on the feature map $X(C \times H \times W)$ of the detection layer to get the feature map $f_X^{7\times7}$, followed by a 1 × 1 convolution to enhance the non-linear learning capability of the network to get the feature maps $f_{X1}^{1\times1}$ and $f_{X2}^{1\times1}$. Then two spatially separated convolutions are performed on $f_{X1}^{1\times1}$ to split the 5 × 5 convolution kernel into 1 × 5 and 5 × 1 to reduce the calculated amount and thus speed up the operation of the network. This step obtains the feature maps $f_{X3}^{1\times5}$ and $f_{X4}^{5\times1}$. To further enhance the receptive field, we perform two separate convolutions on the feature map $f_{X4}^{5\times1}$ to get the feature maps $f_{X5}^{7\times1}$ and $f_{X6}^{1\times7}$. Next, the feature maps $f_{X1}^{1\times1}$, $f_{X2}^{1\times1}$, $f_{X4}^{5\times1}$, and $f_{X6}^{1\times7}$ obtained from different receptive field scales are summed, and then the information is integrated using 1 × 1 convolution. Finally, the receptive field information obtained is multiplied by the original feature map X to get the final feature map $X^{'}(C \times H \times W)$ can be expressed as formula:

$$X^{'} = \left(f^{1\times1}\left(Concat\left(f_{X1}^{1\times1}, f_{X2}^{1\times1}, f_{X4}^{5\times1}, f_{X6}^{1\times7}\right)\right)\right) \otimes X,$$

where $\otimes$ is the element-wise matrix multiplication operation.

2.5. Proposed EODConv

Deep learning models have robust feature extraction capabilities primarily attributed to the convolution operation, which allows networks to become more powerful and effective, resulting in better detection.

However, in traditional convolutional operations, the convolutional kernels are independent of the input. They are not correlated with each other, leading to a lack of detailed feature extraction and efficient convolutional learning in the network.

The ODConv [29] module is capable of linearly weighting the kernel space in four dimensions (number of convolutional kernels, space size, input channels, and output channels) to enhance the learning ability of convolution by correlating information in different dimensions, enabling the network to extract features in more detail and perform convolutional computations more efficiently. However, in detecting lung nodules, we found that learning about the number of convolutional kernels does not improve model performance; instead, it may lead to overfitting and waste of computational resources. Therefore, we propose an EODConv module optimized based on ODConv. We remove dimensional information about the number of convolutional kernels to improve the performance and learning efficiency of the network and reduce the number of parameters.

The description formula for a static convolution operation is as follows:

$$y = w_i x + b.$$

In the formula, $x \in \mathbb{R}^{h\times w\times c_{in}}, y \in \mathbb{R}^{h\times w\times c_{out}}$ represent the input feature map and output feature map, respectively;w_i represents different convolution kernels, and b is bias.

The EODConv structure is shown in the Fig. 6.

The EODConv operation is described by the following formula:

$$y = \left(\begin{array}{c} a_{f1} \odot a_{c1} \odot a_{s1} \odot w_1 + \cdot\cdot \\ \cdot + \odot a_{fn} \odot a_{cn} \odot a_{sn} \odot w_n \end{array}\right) * x + b,$$

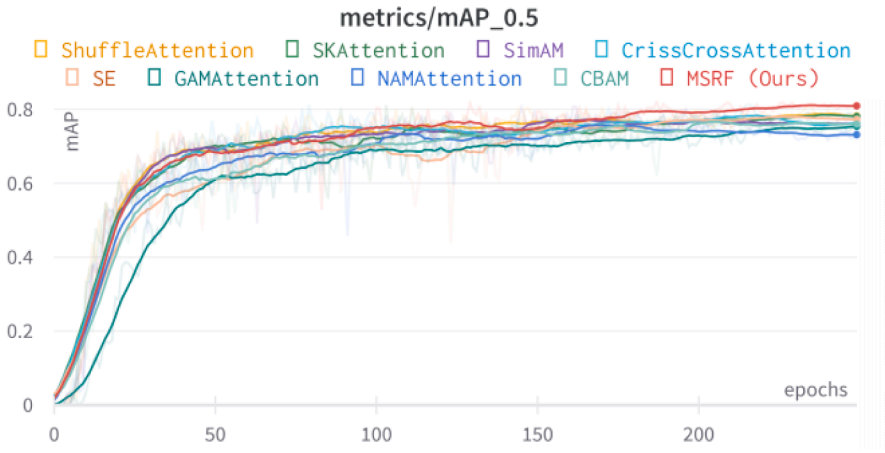
where a_{fi}, a_{ci}, a_{si} represents the attention scalar introduced into the computational convolution kernel. Where $a_{si} \in \mathbb{R}^{k\times k}, a_{ci} \in \mathbb{R}^{c_{in}}, a_{fi} \in \mathbb{R}^{c_{out}}$ is the attention scalar along the space dimension, the input channel dimension, and the output channel dimension. * denotes the convolution operation; $\odot$ denotes the multiplication operations along different dimensions of the kernel space.w_i represents different convolution kernels, and b is bias.

2.6. Evaluation metrics

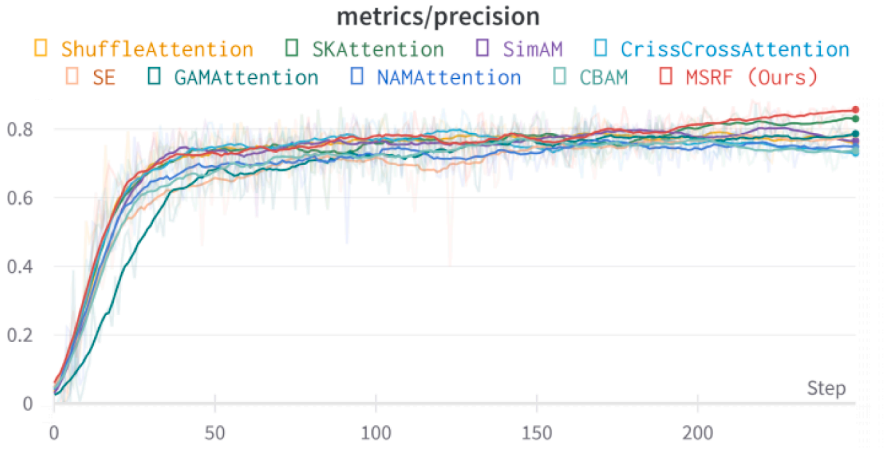
The methods in our study only involved the 2D lung nodule detection stage, so we selected accuracy, recall rate, and mAP as evaluation indicators. The formula is as follows:

$$\begin{cases} \text{precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \end{cases}$$

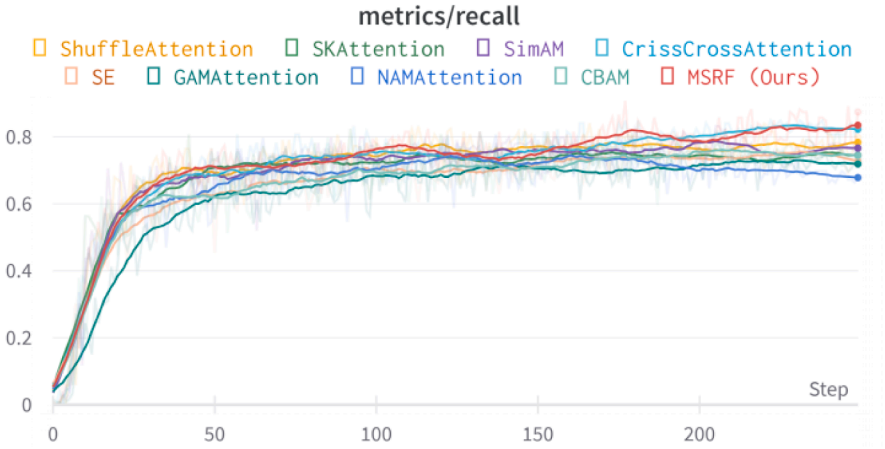
TP, FN, and TN refer to the number of True Positives, False Negatives, and True Negatives, respectively, and TP + FN is the total number



(a) mAP Performance Comparison



(b) Precision Performance Comparison



(c) Recall Performance Comparison

Fig. 8. Comparison of MSRF effect with different attentions.

of positive nodules. TP + FP is the total number of nodules predicted as positive. Average Precision (AP) represents the model's overall performance at different thresholds, while mAP represents the average AP across all categories.

Table 4
Different spatially separable convolutional performance.

Spatial separable convolution size	mAP	Precision	Recall
(1 × 3, 3 × 1) and (1 × 5, 5 × 1)	0.7938	0.8033	0.7863
(1 × 3, 3 × 1) and (1 × 7, 7 × 1)	0.7611	0.7368	0.8235
(1 × 3, 3 × 1) and (1 × 9, 9 × 1)	0.7663	0.7758	0.7562
(1 × 3, 3 × 1) and (1 × 11, 11 × 1)	0.7367	0.7924	0.7058
(5 × 1, 1 × 5) and (1 × 9, 9 × 1)	0.8047	0.7881	0.7814
(5 × 1, 1 × 5) and (1 × 11, 1 × 11)	0.7650	0.7723	0.7981
(7 × 1, 1 × 7) and (1 × 9, 1 × 9)	0.7562	0.7142	0.7981
(7 × 1, 1 × 7) and (1 × 11, 1 × 11)	0.7737	0.7580	0.7899
(9 × 1, 1 × 9) and (1 × 11, 1 × 11)	0.7840	0.7997	0.8067
(5 × 1, 1 × 5) and (1 × 7, 7 × 1) (Ours)	**0.8094**	**0.8567**	**0.8379**

Table 5
Effect of proposed EODConv.

Different dimensions	mAP	Precision	Recall
ODConv (Four dimensions)	0.7969	0.8443	0.8102
Remove spatial	0.7305	0.7652	0.7483
Remove filter	0.7201	0.7538	0.7309
Remove channel	0.7435	0.8129	0.7311
EODConv (Ours, Remove kernel)	**0.8094**	**0.8567**	**0.8379**

Table 6
Performance on EODConv.

Model	parameters	MAdd	Flops
ODConv (Four dimensions)	4509	8608	4525
EODConv (Ours, Remove kernel)	4442	8479	4457

Table 7
Comparison between state-of-the-art networks and YOLO-MSRF.

Models	mAP
Mask R-CNN	0.882
Faster R-CNN	0.919
SSD	0.835
EfficientDet-d0 [38]	0.874
CenterNet [39]	0.853
Cascade R-CNN	0.879
YOLO-MSRF (Ours)	**0.946**

3. Experiments and results

3.1. Dataset

Luna16 [30] is generated from the LIDC-IDRI dataset by excluding slices thicker than 2.5 mm and inconsistent slice spacing or missing slices, eventually producing 888 scanned cases. The dataset contained 1186 positive nodules, and nodules flagged by at least three of the four radiologists are considered ground truth (GT). To facilitate a better understanding of this dataset, we applied data augmentation to the original images and presented realistic nodal information, as shown in Fig. 7.

3.2. Experimental parameters

All experiments for this project were completed on the GPU. The adopted model is built using Python 3.8. The network adopts a stochastic gradient descent optimization algorithm. The model parameter settings are shown in Table 1.

3.3. Ablation experiments on YOLO-MSRF

To facilitate comparison with the initial network, we did not use data augmentation techniques to expand the dataset. Instead, we used a dataset with an initial sample size of 1186 for the ablation experiments. The dataset is randomly and equally split into ten subsets for five cross-validations. For each fold, one subset is selected for testing, eight subsets for training, and one subset for validation.

In order to validate our proposed YOLO-MSRF structure, we performed corresponding ablation experiments on the original YOLOv7 model. The experimental results are shown in Table 2.

Table 2 shows that YOLO-MSRF significantly improves the accuracy, recall, and mAP of detecting small target nodules compared to the initial YOLOv7, where SODL enhances contextual semantic information and improves the model's ability to extract features from nodules. EODConv learns information about the three dimensions in the convolution module, resulting in a remarkable performance gain for the model. The MSRF module fuses limited feature information by increasing the receptive field, allowing the network to focus more on the information around the small nodule features to train the network and enhance the network's ability to recognize small nodules. The different combinations produced a positive optimization of the overall performance of YOLOv7, with the best results when all three improvements were used simultaneously.

3.4. Performance comparison

We further performed ablation experiments to validate the proposed MSRF module and compared it with many classical attention models. The results of the experiments are shown in Table 3, and the resulting demonstration figures are shown in Fig. 8. We only replaced the MSRF in the final detection layer with the attention module. We kept the rest of the network structure the same.

Based on the data in Table 3, MSRF is more effective in detecting small nodules than different attention models, demonstrating the effectiveness of an enhanced receptive field. Therefore, our study's MSRF method enables the model to detect lung nodules more accurately and effectively than attention methods.

As seen from Fig. 8, our MSRF module has achieved the best effect in detecting the direction of lung nodules; it indirectly proves that expanding the receptive field is conducive to better detection of lung nodules.

3.5. MSRF model ablation experiments

We conducted detailed ablation experiments to verify the feasibility of the spatially separable convolution employed in the MSRF module. The results are summarized in Table 4.

As shown in Table 4 above, we found that when the size of the spatially separable convolution is too large, it results in a larger receptive field. Thus, the network extracts less detailed feature information. Conversely, if the size is too small, the receptive field becomes smaller, and the network may focus more on redundant feature information. Therefore, we must choose the appropriate convolution size to extract nodal features fully. We found experimentally that the best metrics can be achieved by using convolution sizes of (5 × 1, 1 × 5) and (1 × 7, 7 × 1).

3.6. EODConv model ablation experiments

To verify the superiority of our proposed EODConv for nodule detection, we performed the corresponding ablation experiments, and the results are shown in Table 5.

The table above shows that when the dimensions of input channels, output channels, and space size are removed, the performance of the network decreases significantly compared to the initial use of ODConv.

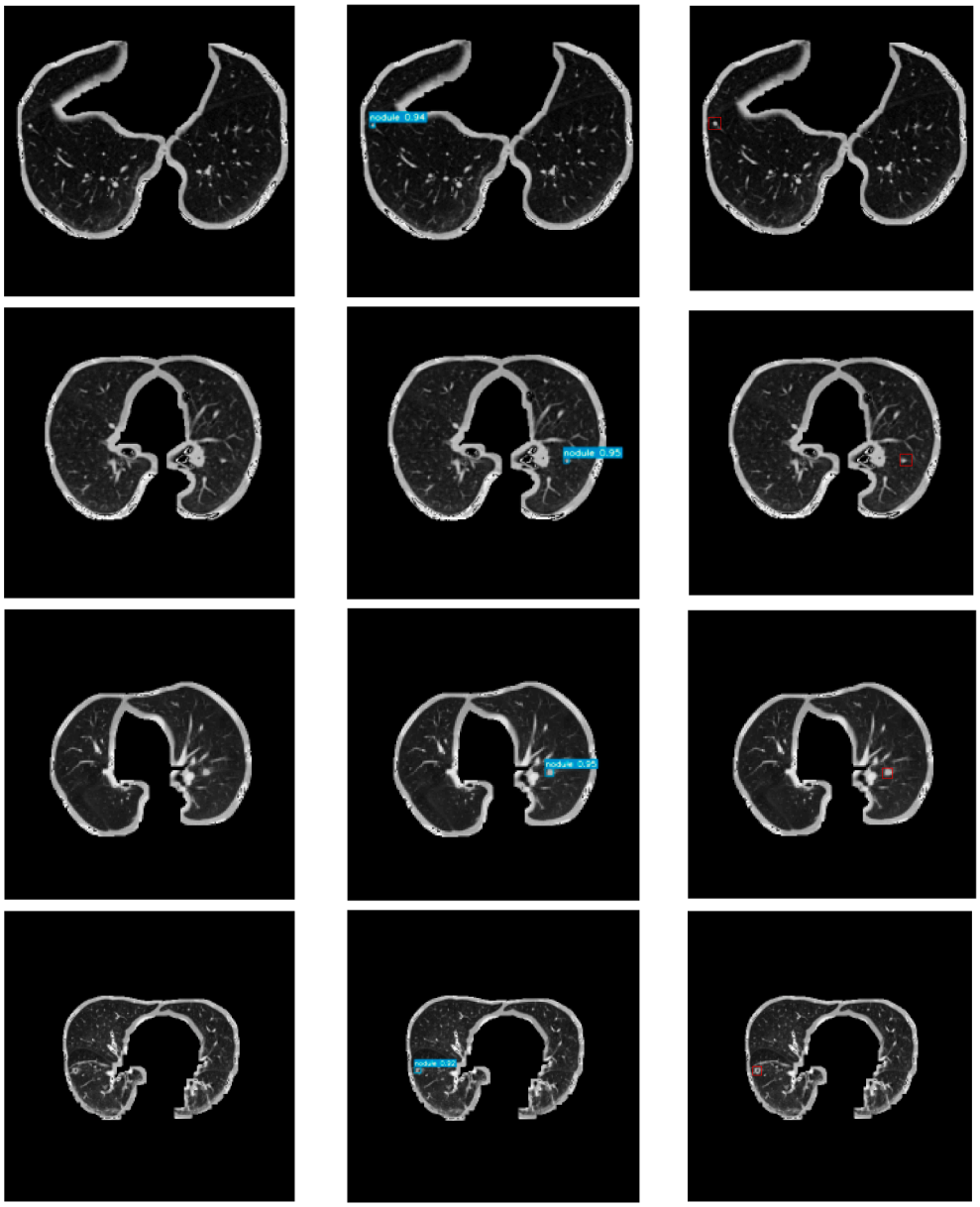


Fig. 9. Detection Results.

However, when we remove the dimensional information of the convolutional kernels, the performance metric improves, which proves the feasibility of our proposed EODConv module.

As shown in Table 6 above, the same variable settings are used in this experiment, i.e., input and output channels of 64 and 128, respectively, and a convolution kernel size of 3×3. In the table, the MAdd metric represents the number of addition operations per layer of the neural network, usually measured in millions of operations, and is used to measure the computational complexity of the neural network model. In addition, Flops (floating point operations) is an important metric to evaluate a neural network model's computational complexity and performance. A higher Flops value indicates that the model has a higher computational complexity and may require more computational resources and time for training and inference.

According to the results, the model's parameters, computation, and floating point operations were reduced by 1.5 %. This means that using EODConv convolution is beneficial for improving the computational efficiency and performance of the model.

3.7. omparison with state-of-the-art detection models

To better verify the performance of the model, we compare the proposed model with many state-of-the-art object detection network models. In addition, to achieve the same comparison effect, we also use data augmentation technology to expand the dataset. The results are shown in Table 7.

As shown above, our results achieve the best results under the relatively popular one- or two-stage object detection frameworks, proving our method's feasibility.

3.8. Results shown

Fig. 9(a) shows the pre-processed images, and Fig. 9(b) and (c) show the predicted results and the real nodules, respectively. The presence of blood vessels and airways around the nodule can be seen in the figure. This information could potentially contain pathological information that could lead the radiologists to misjudge the results of the nodule detection. However, we proposed in this paper the YOLO-MSRF method, which performs best in detecting lung nodules and helps radiologists make accurate judgments.

4. Conclusion

In this paper, we propose YOLO-MSRF, a 2D method for detecting lung nodules based on YOLOv7, to achieve accurate nodule detection. To address the problem that conventional detection algorithms are ineffective at detecting small objects, we propose SODL, which has the largest feature map and contains more detailed information about small targets. In addition, we propose the EODconv module to learn the kernel space information in three dimensions in a weighted manner, effectively reducing the computation and number of parameters and improving performance.

At the same time, we found that the object detection details of nodules are quickly lost with the deepening of the network, and the object information of nodules is not easy to locate. To better amplify nodule features, we adopted the receptive field enhancement module MSRF to learn the fine-grained peripheral nodules' features better.

The results show that a 95.26 % mAP index was obtained in this paper. The accuracy and recall rates reached 95.41 % and 94.02 %, respectively, surpassing many state-of-the-art 2D nodular detection methods. The superiority of this method indicates that it is more suitable for hospitals to deploy it to detect lung nodules in the early stages. Next, we will try to construct a 3D false-positive elimination network to meet the actual needs of clinical work.

CRediT authorship contribution statement

Xiaosheng Wu: Conceptualization, Investigation, Methodology, Software, Writing – original draft. **Hang Zhang:** Formal analysis, Investigation, Resources, Validation, Visualization. **Junding Sun:** Investigation, Methodology, Software, Visualization, Writing – original draft. **Shuihua Wang:** Funding acquisition, Investigation, Supervision, Writing – original draft, Writing – review & editing. **Yudong Zhang:** Funding acquisition, Methodology, Supervision, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] C. Wild, E. Weiderpass, B.W. Stewart, World cancer report: cancer research for cancer prevention. 2020: International Agency for Research on Cancer.

[2] S. Gao, et al., Lung cancer in People's Republic of China, J. Thoracic Oncol. 15 (10) (2020) 1567–1576.

[3] P. Mohamed Shakeel, M.I. Desa, M. Burhanuddin, Improved watershed histogram thresholding with probabilistic neural networks for lung cancer diagnosis for CBMIR systems, Multimedia Tools Appl. 79 (2020) 17115–17133.

[4] H.F. Kareem, et al., Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset, Indonesian J. Electr. Eng. Comput. Sci. 21 (3) (2021) 1731.

[5] F. Grossi, et al., 919P artificial intelligence supporting lung cancer screening: computer aided diagnosis of lung lesions driven by morphological feature extraction, Ann. Oncol. 33 (2022) S967.

[6] P.M. Shakeel, M.A. Burhanuddin, M.I. Desa, Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks, Measurement 145 (2019) 702–712.

[7] Z. Chen, et al., Plant disease recognition model based on improved YOLOv5, Agronomy 12 (2) (2022) 365.

[8] G. Gao, et al., Recognition and detection of greenhouse tomatoes in complex environment, Traitement Du Signal 39 (1) (2022).

[9] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014.

[10] R. Girshick, Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[11] R. Faster, Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 9199 (10.5555) (2015) 2969239–2969250.

[12] K. He, et al., Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[13] J. Ding, et al., Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. Medical Image Computing and Computer Assisted Intervention– MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part III 20, Springer, 2017.

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] H. Zhang, Y. Peng, Y. Guo, Pulmonary nodules detection based on multi-scale attention networks, Sci. Rep. 12 (1) (2022) 1466.

[16] N. Guo, Z. Bai, Multi-Scale Pulmonary Nodule Detection by Fusion of Cascade R-CNN and FPN. 2021 International Conference on Computer Communication and Artificial Intelligence (CCAI), IEEE, 2021.

[17] Z. Cai, N. Vasconcelos, Cascade r-cnn: delving into high quality object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] L. Cai, et al., Mask R-CNN-based detection and segmentation for pulmonary nodule 3D visualization diagnosis, Ieee Access 8 (2020) 44400–44409.

[19] W. Liu, et al., Ssd: single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016.

[20] T.-Y. Lin, et al., Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[21] J. Redmon, et al., You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] J. Redmon, A. Farhadi, Yolov3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.

[24] K. He, et al., Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[25] Y.-S. Huang, et al., One-stage pulmonary nodule detection using 3-D DCNN with feature fusion and attention mechanism in CT image, Comput. Methods Programs Biomed. 220 (2022) 106786.

[26] S. Mei, H. Jiang, L. Ma, YOLO-lung: a practical detector based on imporved YOLOv4 for pulmonary nodule detection 2021. 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2021.

[27] K. Liu, STBi-YOLO: a real-time object detection method for lung nodule recognition, IEEE Access 10 (2022) 75385–75394.

[28] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv: 2207.02696, 2022.

[29] C. Li, A. Zhou, A. Yao, Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947, 2022.

[30] A.A.A. Setio, et al., Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge, Med. Image Anal. 42 (2017) 1–13.

[31] S. Woo, et al., Cbam: convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[32] Liu, Y., et al., *NAM: Normalization-based attention module.* arXiv preprint arXiv: 2111.12419, 2021.

[33] Liu, Y., Z. Shao, and N. Hoffmann, *Global attention mechanism: Retain information to enhance channel-spatial interactions.* arXiv preprint arXiv:2112.05561, 2021.

[34] Q.-L. Zhang, Y.-B. Yang, Sa-Net: Shuffle Attention for Deep Convolutional Neural Networks, IEEE, 2021.

[35] X. Li, et al., Selective kernel networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[36] Yang, L., et al. *Simam: A simple, parameter-free attention module for convolutional neural networks*. in *International conference on machine learning*. 2021. PMLR.

[37] Z. Huang, et al., Ccnet: criss-cross attention for semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[38] M. Tan, R. Pang, Q.V. Le, Efficientdet: scalable and efficient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[39] X. Zhou, D. Wang, P. Krähenbühl, Objects as points. arXiv preprint arXiv: 1904.07850, 2019.