DOI: 10.19884/j.1672-5220.202401002

DCA-YOLO: Detection Algorithm for YOLOv8 Pulmonary Nodules Based on Attention Mechanism Optimization

SONG Yongsheng, LIU Guohua*
School of Computer Science and Technology, Donghua University, Shanghai 201620, China

Abstract: Pulmonary nodules represent an early manifestation of lung cancer. However, pulmonary nodules only constitute a small portion of the overall image, posing challenges for physicians in image interpretation and potentially leading to false positives or missed detections. To solve these problems, the YOLOv8 network is enhanced by adding deformable convolution and atrous spatial pyramid pooling (ASPP), along with the integration of a coordinate attention (CA) mechanism. This allows the network to focus on small targets while expanding the receptive field without losing resolution. At the same time, context information on the target is gathered and feature expression is enhanced by attention modules in different directions. It effectively improves the positioning accuracy and achieves good results on the LUNA16 dataset. Compared with other detection algorithms, it improves the accuracy of pulmonary nodule detection to a certain extent.

Keywords: pulmonary nodule; YOLOv8 network; object detection; deformable convolution; atrous spatial pyramid pooling (ASPP); coordinate attention (CA) mechanism
CLC number: TP391.1 **Document code**: A
Article ID: 1672-5220(2025)01-0078-10

Open Science Identity (OSID)

0 Introduction

Lung cancer as one of the deadliest cancers worldwide, typically first presents as pulmonary nodules[1]. For certain nodules, especially those with adverse morphological characteristics and rapid growth rates, additional examinations and treatments are required to exclude lung cancer. Currently, the diagnosis of pulmonary nodules heavily relies on the computed tomography (CT) technology[2]. However, the substantial volume of CT image data increases the diagnostic burden on physicians, and extended periods of diagnostic work may lead to fatigue, potentially resulting in disease misdiagnosis or misinterpretation[3]. Furthermore, physicians primarily rely on their own experience for subjective judgments of CT images, lacking objective quantitative analyses. Therefore, harnessing computer technology to aid physicians in pulmonary nodule detection and diagnosis holds promise for markedly improving diagnostic precision and efficiency.

Currently, significant progress has been made in pulmonary nodule detection using deep learning algorithms. Sun et al.[4] proposed an attention-embedded complementary flow convolutional neural network (CNN) method to reduce false positives in pulmonary nodule detection, thereby enhancing the accuracy and reliability of the detection. Zhao et al.[5] introduced a method employing three-dimensional U-Net and CNN for pulmonary nodule detection, achieving accurate detection and classification of pulmonary nodules. Liu et al.[6] combined existing methods such as deep separable over-parameterized convolution layers, convolutional block attention modules, and focal loss functions with the base YOLOv3 architecture to improve the accuracy and efficiency of the model.

In pulmonary CT images, background structures such as vessels and bronchi resemble pulmonary nodules, posing challenges in target detection[7]. Additionally, various types of noises, including speckle noise and circular artifacts, exist in the images. These noises may interfere with feature extraction, negatively impacting the detection of small nodules and potentially leading to false positives[8]. Furthermore, pulmonary nodules are small targets with low annotation box resolution. These small nodules may only occupy a few pixels in the image, so they fail to provide sufficient detailed features. When deep learning networks undergo downsampling to a certain extent, feature loss may occur, resulting in missed detections[9].

To address the aforementioned challenges, we propose an intelligent pulmonary nodule detection algorithm based on an attention-optimized YOLOv8[10]. This algorithm enhances the YOLOv8 network by incorporating deformable convolution[11] and atrous spatial pyramid pooling (ASPP)[12] while integrating a coordinate attention (CA) mechanism[13]. This novel architecture enables the network to concentrate on small targets while broadening the receptive field concomitantly, without compromising resolution. Simultaneously, the utilization of attention modules in different directions aggregates contextual information on the target, thereby enhancing feature

Received date: 2024-01-08
* Correspondence should be addressed to LIU Guohua, email: ghliu@dhu.edu.cn
Citation: SONG Y S, LIU G H. DCA-YOLO: detection algorithm for YOLOv8 pulmonary nodules based on attention mechanism optimization[J]. *Journal of Donghua University (English Edition)*, 2025, 42(1): 78-87.

representation, improving localization accuracy, and decreasing the occurrence of false positives and false negatives.

1 Relevant Methods

1.1 YOLOv8 object detection network

The structure of YOLOv8[10] comprises three components: the main feature extraction network (backbone), the enhanced feature extraction network (FPN), and the prediction head (YOLO head).

YOLOv8 employs CSPResNeXt as the backbone for its main feature extraction network. The CSPResNeXt network integrates the advantages of residual connections, dense connections and multi-scale feature extraction. Upon inputting an image into the network, feature extraction occurs initially in the backbone network, producing feature layers that represent the image's features. Simultaneously, the backbone network extracts three feature layers of different scales, known as effective feature layers, which are utilized in subsequent network construction.

FPN serves as the enhanced feature extraction module in YOLOv8, primarily integrates effective feature layers of different scales from the main backbone network. Through both upsampling and downsampling, it facilitates the flow of information between feature layers of varying scales, effectively combining feature information from different semantic levels. In YOLOv8, FPN adopts the path aggregation network (PANet) structure, iteratively performing upsampling and downsampling to further enhance feature representation.

YOLO head functions as the classification and regression branches, further extracting target information based on the enhanced feature layers fused by FPN. The input feature layer of YOLO head consists of a spatial dimension (width and height) and a channel dimension and can be considered as consisting of countless feature points, each of which contains sample semantic information. YOLO head achieves detection by assessing whether each feature point corresponds to an object by judging the inclusion of prior bounding boxes. The decoupled design of the head, namely classification and regression, is realized through an independent 1×1 convolution. The final output displays prior bounding boxes containing detected objects on the original image. YOLOv8, as the latest iteration in the YOLO series of object detection models, has achieved state-of-the-art performance across multiple object detection datasets and is considered a preferred algorithm for the majority of object detection tasks[14].

1.2 Deformable convolution

Deformable convolution is an improved convolution operation that has emerged in the fields of object detection and semantic segmentation. Unlike standard convolution kernels where the sampling positions are fixed, deformable convolution introduces an offset mechanism based on input data. This allows the convolution operation to automatically adjust its shape while extracting features, ensuring that it adequately focuses on the target object for feature extraction. Consequently, this approach alleviates the issue of false positives caused by background and noise interference in images.

In standard convolutional operations, the feature map is divided into segments that correspond to the size of the convolutional kernel. Subsequently, convolutional processes are performed on these segments, each fixed in its position on the feature map, the standard convolution process can be represented as

$$\boldsymbol{y}(p_0) = \sum_{p_n \in R} \boldsymbol{w}(p_n) \cdot \boldsymbol{x}(p_0 + p_n), \tag{1}$$

where p_0 represents the feature point before convolution on the image; p_n represents the offset of each point in the convolution kernel; R enumerates the offsets relative to the central point, $R = \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$; $\boldsymbol{w}(p_n)$ denotes the weight of the current convolution position in the convolution kernel; $\boldsymbol{x}(p_0+p_n)$ represents the pixel value at the offset point on the feature map; $\boldsymbol{y}(p_0)$ is the resulting feature value after convolution for that specific feature point.

For deformable convolution, an offset Δp_n needs to be added after computing the position, the offset is typically learned by the network. The convolution operation of deformable convolution can be expressed as

$$\boldsymbol{y}^*(p_0) = \sum_{p_n \in R} \boldsymbol{w}(p_n) \cdot \boldsymbol{x}(p_0 + p_n + \Delta p_n), \tag{2}$$

where $\boldsymbol{y}^*(p_0)$ is the resulting feature value after deformable convolution.

In convolutional neural networks, the offset Δp_n is typically a decimal, while the feature map can only correspond to integer positions. To address this issue, a bilinear interpolation method is employed, estimating the feature value at the decimal position by calculating the feature values of the nearest neighboring points. The calculation process can be represented as

$$\boldsymbol{x}(p) = \sum_q G(q, p) \cdot \boldsymbol{x}(q), \tag{3}$$

$$G(q, p) = g(q_x, p_x) \cdot g(q_y, p_y), \tag{4}$$

$$g(a, b) = \max(0, 1 - |a - b|), \tag{5}$$

where p represents any decimal position, in the x-axis is p_x, in the y-axis is p_y; q denotes all neighboring integer positions of p on feature map x; g represents the kernel function for bilinear interpolation; G denotes the influence weights of integer points on p; a and b represent the pixel positions in the image, a denotes the coordinate of the current fractional position being computed, and b denotes the coordinate of the neighboring integer position.

1.3 ASPP

ASPP is a crucial module in deep learning. In object detection tasks, multiscale information is paramount for accurately distinguishing between different targets. The ASPP module effectively captures semantic information at different scales by combining multiple dilated convolutions and pooling operations of varying scales. In this way, the ASPP module can extract rich features and fuse them, bringing a certain degree of accuracy

improvement to object detection tasks. In object detection, the scale, shape and contextual information of objects are essential for precise localization and identification. The ASPP module enhances the ability of the target detection algorithm to identify multi-scale targets by integrating multi-scale information, thus improving the detection accuracy. Different dilation rates of convolutions with a kernel size of 3×3 are illustrated in Fig. 1. As shown in Fig. 1, conv represents convolution, conv 3×3 represents a matrix with a convolution kernel size of 3×3, and rate represents the dilation rate.

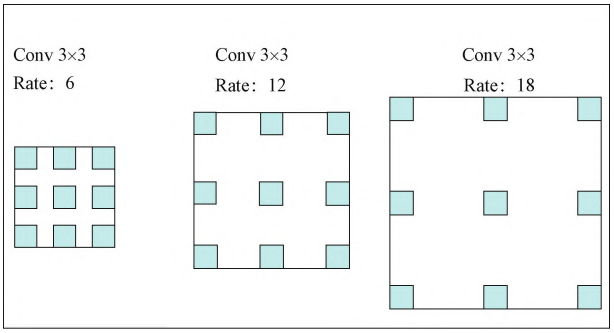


Fig. 1 Different dilation rates of convolutions with a kernel size of 3×3

Specifically, the ASPP module can effectively capture multi-scale features of the input image by utilizing dilated convolutions with different dilation rates and multi-scale pooling operations. Consequently, even objects of varying scales can be accurately identified and localized. The ASPP module also provides a broader range of contextual information, contributing to the improvement of both the classification and localization accuracy of the target. The structure of the ASPP module is depicted in Fig. 2, and the ASPP convolution operation can be expressed as

$$\boldsymbol{M}(u) = \sum_k w_k \cdot \boldsymbol{F}(u + r_k \cdot p_k), \tag{6}$$

where $\boldsymbol{M}$ represents the output feature map; $\boldsymbol{F}$ represents the input feature map; u represents the coordinate of a pixel position in the output feature map $\boldsymbol{M}$; w_k represents the kth dilated convolution kernel; r_k represents the dilation rate of the kth dilated convolution kernel; p_k represents the central position corresponding to the kth dilated convolution kernel. Conv 1×1 in Fig. 2 represents point convolution and is mainly used to adjust the number of channels, and concat is short for concatenation.

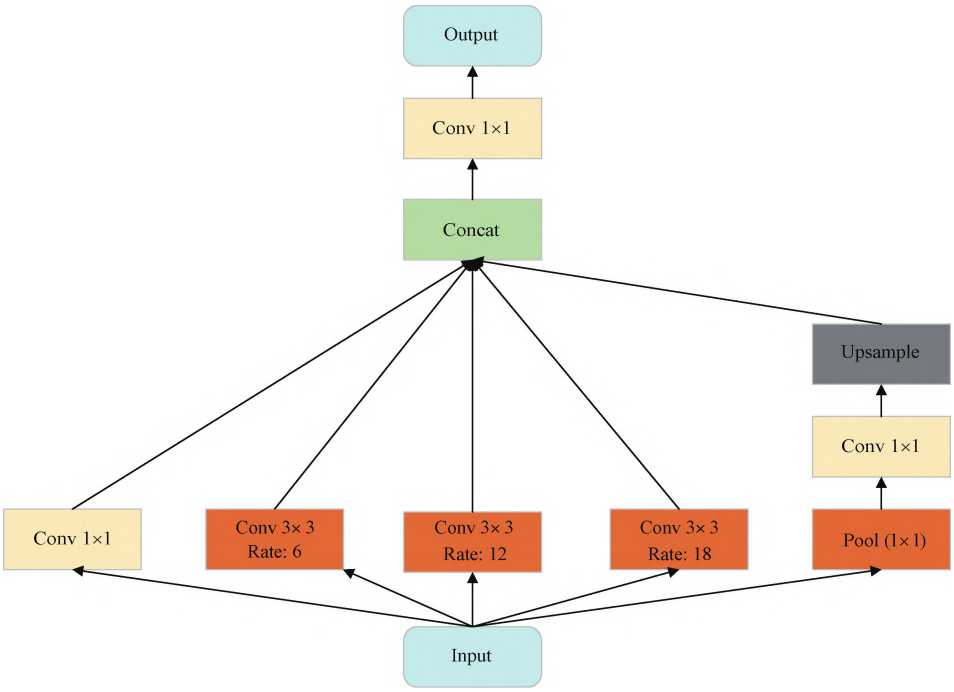


Fig. 2 Structure of ASPP module

1.4 Attention mechanism

The attention mechanism is designed to simulate human attention, allowing the network to concentrate on vital parts of the input data while disregarding irrelevant components. This improvement enables the network to perform better when tackling complex tasks.

Spatial attention mechanism[15-16] and channel attention mechanism[17] are commonly encountered attention mechanisms. The algorithmic flowchart of the channel attention mechanism is illustrated in Fig. 3, and its process can be outlined in four steps.

1) Transformation (F_{tr}). This step involves passing the input feature map $\boldsymbol{X}$, which has dimensions $H' \times W' \times C'$ (H', W' and C' represent the height, width and number of channels of X, respectively), through a regular convolutional layer to further extract features, resulting in the feature map $\boldsymbol{U}$.

2) Squeeze (F_{sq}). In this step, the feature map

undergoes global average pooling, which involves averaging all pixels of each feature map. It generates a vector $\boldsymbol{Z}_C$ of size $1 \times 1 \times C$, where C is the number of channels. Through this operation, each channel can be represented by a single value indicating its feature information. This can be expressed by

$$\boldsymbol{Z}_C = F_{sq}(\boldsymbol{U}_C) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{U}_C(i, j), \quad (7)$$

where $\boldsymbol{Z}_C$ represents the C-channel of the output feature map; $\boldsymbol{U}_C$ represents the C-channel of the feature map $\boldsymbol{U}$; H and W represent the height and the width, respectively; i and j represent the enumeration of the feature points of the current channel in the height and the width, respectively.

3) Excitation (F_{ex}). In this step, two fully connected layers are utilized to generate the desired weight information using the learned weights $\boldsymbol{E}$. Through learning, the weights $\boldsymbol{E}$ are acquired and utilized to capture intricate non-linear relationships among feature channels, ultimately resulting in the generation of suitable channel attention weights. This can be expressed by

$$\boldsymbol{S} = F_{ex}(\boldsymbol{Z}, \boldsymbol{E}) = \sigma(\boldsymbol{E}_2 \delta(\boldsymbol{E}_1 \boldsymbol{Z})), \quad (8)$$

where $\boldsymbol{S}$ denotes the weighted feature map; $\boldsymbol{Z}$ denotes the vector generated by the squeze operation; $\boldsymbol{E}_1$ represents the weights of the first fully connected layer; δ is the relu activation function; $\boldsymbol{E}_2$ represents the weights of the second fully connected layer; σ is the sigmoid activation function.

4) Scale (F_{scale}). This process involves element-wise multiplication of the feature map and the original feature map. Each of the $H \times W$ value in feature map $\boldsymbol{U}$ is multiplied by the corresponding weight in $\boldsymbol{S}$. The result is the feature map after incorporating the channel attention mechanism.

The traditional attention mechanism usually employs global max pooling or average pooling for channel attention, leading to the loss of spatial information about objects. To address this issue, we aim to introduce both channel attention and spatial attention simultaneously. The CA mechanism embeds positional information into channel attention. For small objects, the CA mechanism learns that they are more concentrated in certain regions of the image. When generating attention weights, these regions receive higher weight assignments. Consequently, in subsequent feature extraction, regions with a higher probability of target occurrence are enhanced by the attention mechanism, allowing for a more significant expression of features for small targets. The CA captures the spatial information and contextual dependencies of target features through position encoding, thereby enhancing the representational capacity for small targets.

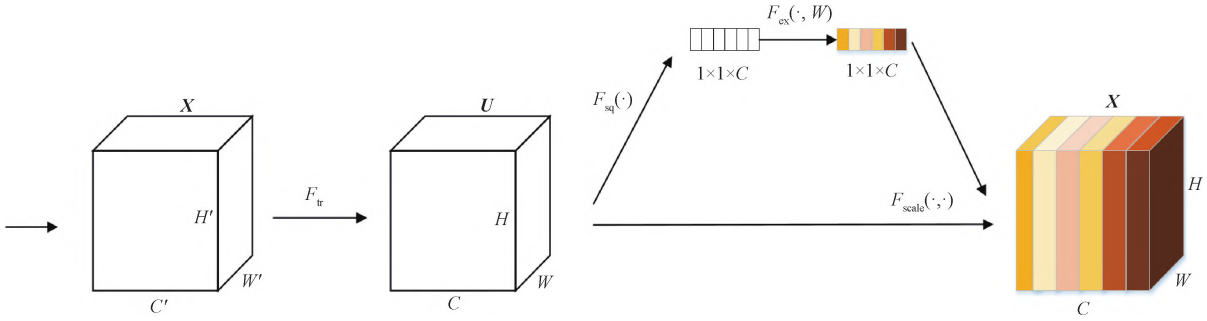


Fig. 3 Algorithmic flowchart of channel attention mechanism

2 Algorithm for YOLOv8 Pulmonary Nodules Based on Attention Mechanism Optimization

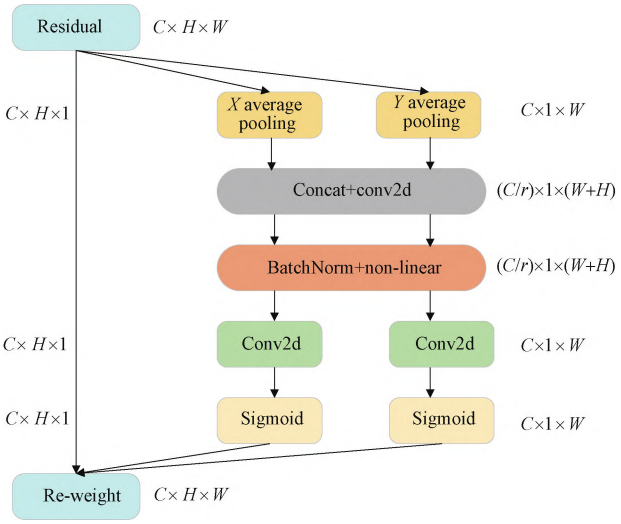
2.1 Improvement of backbone network

The backbone feature extraction network of YOLOv8 consists of multiple standard convolutions and cross stage partial (CSP) modules[18]. However, when dealing with small target objects like pulmonary nodules, traditional convolutions fail to focus adequately, resulting in insufficient extraction of target features. To address this issue, we introduce deformable convolutions in the backbone network to replace traditional convolutions. This allows the convolution to adapt its shape, better focusing on the target and extracting features, thereby mitigating the issue of false detections. Due to the small size and limited feature information of pulmonary nodules, as well as their low pixels to background structures such as bronchi and blood vessels, we employ ASPP for feature extraction in the shallow layers of the backbone feature extraction network. The fusion of these shallow features with the last layer feature upsampled from the enhanced feature extraction network helps supplement high-level features, improving the model's localization accuracy.

2.2 Model enhancement with CA mechanism

The CA mechanism integrates coordinate information with the original feature map through convolutional operations, generating precise attention weights. During the weight generation phase, the system assigns higher weights to regions with a higher probability of containing the target. This enables the attention mechanism to strengthen regions where the target is more likely to

appear, significantly expressing the features of small targets during subsequent feature extraction. This approach effectively enhances the model's detection performance for small targets, thereby improving the accuracy and robustness of pulmonary nodule detection. The structure of the CA mechanism is illustrated in Fig. 4.



r — the reduction factor data; conv2d — a convolution operation on multi-dimensional.

Fig. 4 Structure of CA mechanism

The process of generating CA mechanism involves two main phases.

1) Coordinate embedding. In this phase, a set of coordinate encodings are generated for each spatial position in the input feature map to represent the absolute spatial information of that location. A group of the learned embedding parameters are utilized, combined with the coordinate indices of each position, to map and generate high-dimensional coordinate embedding vectors. These embedding vectors comprehensively encode the spatial information of each position.

2) Attention generation. In this phase, a convolutional neural network learns the attention weight distribution for each position. The network synthesizes the coordinate embedding vectors for each position along with the surrounding feature representations to assess the probability of the presence of a target at that location, generating an attention distribution. Finally, the attention weights are applied to the feature map through a weighted summation.

Through this joint approach of position encoding and attention learning, the CA mechanism can fully exploit the spatial correlations of features, learning the prior distribution of target occurrences. This effectively enhances the modeling capability for small objects and long-distance dependencies. Such a design reinforces the sensitivity and learning capacity of the detection model to the spatial distribution of targets. For an input feature map $\boldsymbol{X}$, the initial encoding involves using pooling kernels of size $(H, 1)$ and $(1, W)$ to encode the channels in the vertical and horizontal directions. The output features for the C-channel are expressed as

$$\boldsymbol{Z}_C^H(H) = \frac{1}{W}\sum_{0\leqslant i\leqslant W}\boldsymbol{X}_C(H, i), \tag{9}$$

$$\boldsymbol{Z}_C^W(W) = \frac{1}{H}\sum_{0\leqslant j\leqslant H}\boldsymbol{X}_C(j, W), \tag{10}$$

where $\boldsymbol{X}_C$ represents the C-channel of the input feature; $\boldsymbol{Z}_C^H$ represents the C-channel of the output feature in the vertical direction at height H; $\boldsymbol{Z}_C^W$ represents the C-channel of the output feature in the horizontal direction at width W.

Next is attention generation, utilizing a 1×1 convolution F_1 for channel adjustment, yielding the intermediate feature map $\boldsymbol{f}$. The formula is

$$\boldsymbol{f} = F_1(\boldsymbol{Z}^H, \boldsymbol{Z}^W), \tag{11}$$

where $\boldsymbol{Z}^H$ represents the output feature in the vertical direction at height H; $\boldsymbol{Z}^W$ represents the output feature in the horizontal direction at width W. The feature map $\boldsymbol{f}$ is then split along the height and width dimensions. Split refers to the readjustment of $(C/r) \times 1 \times (H + W)$ to $C \times H \times 1$ and $C \times 1 \times W$ by convolution, resulting in two independent tensors $\boldsymbol{f}^{\mathrm{h}} \in R^{(C/r)\times H}$ and $\boldsymbol{f}^{\mathrm{w}} \in R^{(C/r)\times W}$, which represent the feature maps for the horizontal and vertical directions, respectively. The channels of both feature maps are adjusted to match the same number of channels as the input $\boldsymbol{X}$, denoted by $\boldsymbol{g}^{\mathrm{h}}$ and $\boldsymbol{g}^{\mathrm{w}}$, which represent the spatial attention feature maps generated in the horizontal and vertical directions, respectively. This adjustment can be expressed as

$$\boldsymbol{g}^{\mathrm{h}} = \sigma(F_{\mathrm{h}}(\delta(\boldsymbol{f}^{\mathrm{h}}))), \tag{12}$$

$$\boldsymbol{g}^{\mathrm{w}} = \sigma(F_{\mathrm{w}}(\delta(\boldsymbol{f}^{\mathrm{w}}))), \tag{13}$$

where F_{h} denotes the 1×1 convolution in the vertical direction; F_{w} represents the 1×1 convolution in the horizontal direction, respectively.

Expanding $\boldsymbol{g}^{\mathrm{h}}$ and $\boldsymbol{g}^{\mathrm{w}}$, the CA mechanism can be expressed as

$$\boldsymbol{y}_C(i, j) = \boldsymbol{X}_C(i, j) \times \boldsymbol{g}_C^{\mathrm{h}}(i) \times \boldsymbol{g}_C^{\mathrm{w}}(j), \tag{14}$$

where $\boldsymbol{y}_C$ represents the C-channel of output after the CA mechanism; $\boldsymbol{g}_C^{\mathrm{h}}$ and $\boldsymbol{g}_C^{\mathrm{w}}$ represent the C-channel of spatial attention feature maps generated in the horizontal and vertical directions, respectively.

The three multi-scale feature layers extracted from the backbone network, referred to as effective feature layers, are input into the subsequent network. These three effective feature layers represent visual features at different abstraction levels, providing a foundation of features for subsequent object detection. Following the output of the three effective feature layers from the backbone feature extraction network, the CA module is applied for fusion. Based on the feature maps, attention weights are generated for each spatial position,

indicating the probability of containing the target at that position. The calculation of attention weights incorporates prior information about the target's spatial position, allowing the network to focus on the coordinate region where the target is likely to appear in the feature map. The design of the CA mechanism enables the model to pay more attention to the spatial distribution of pulmonary nodules, thereby enhancing the detection capability for small targets, reducing the probability of false negatives, and improving the overall detection performance of the model. The overall architecture of YOLOv8 is illustrated in Fig. 5. In Fig. 5, CBS stands for convolution-batch normalization-Silu activation (Conv-BN-Silu); s represents the step size; k signifies the convolution kernel size; C2f indicates cross stage feature fusion, with $N = 3$ specifying the number of bottleneck blocks; SPPF denoting spatial pyramid pooling-fast, with $k = 6$ specifies the convolution kernel size. DCN are newly added deformable convolutions, which output three effective feature layers after convolution. Then the attention mechanism is added to introduce the subsequent enhanced feature extraction network. At the same time, the ASPP module is used to extract shallow features in the initial part of the image and then fused with the features of the enhanced feature extraction network.

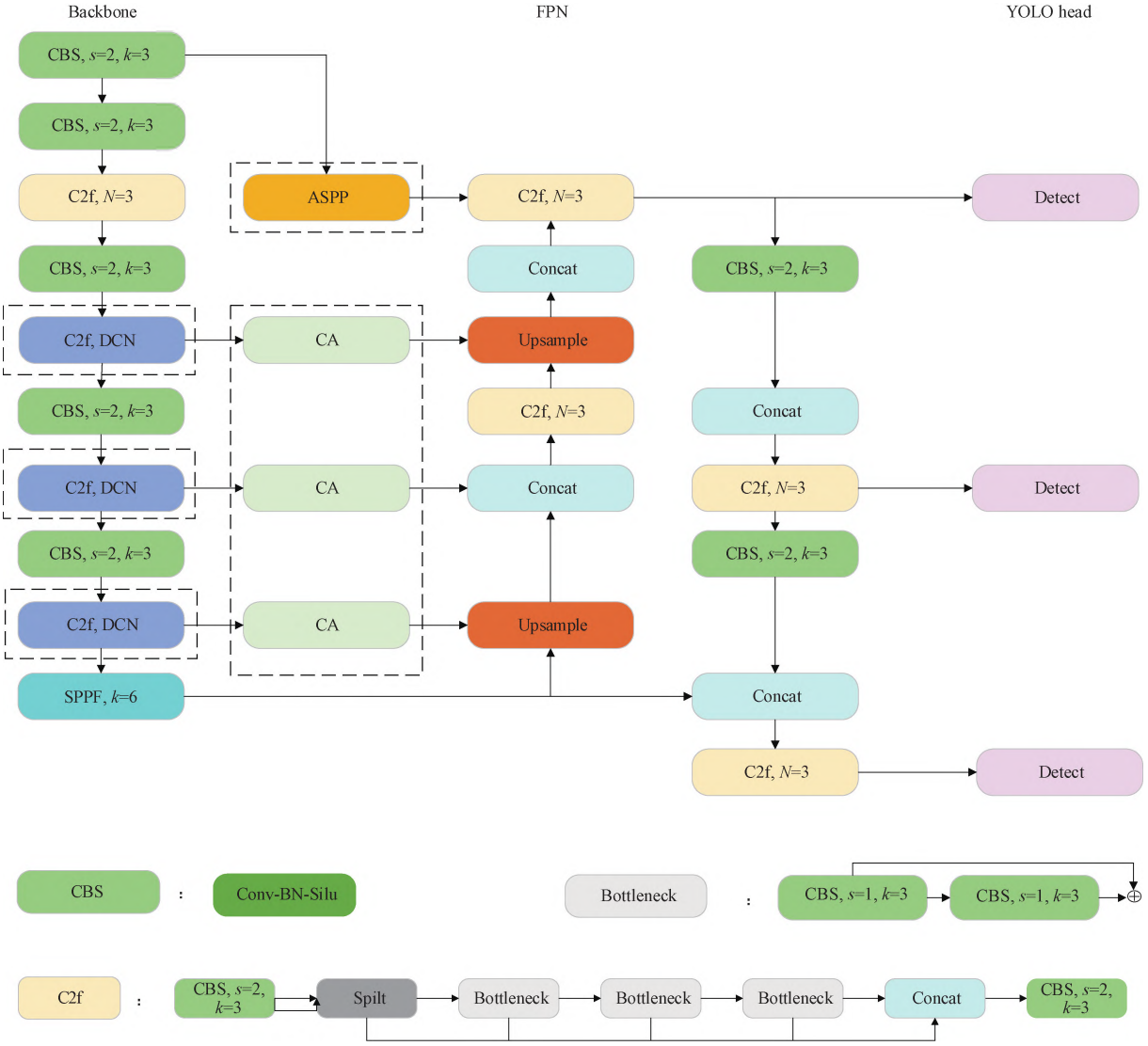


Fig. 5 Overall architecture of YOLOv8

3 Experimental Results and Analyses

3.1 Dataset and preprocessing

The dataset used in the experiment is LUNA16[19], which includes 888 CT images of the lungs, containing 1 186 nodule samples. These nodules consist of 100 malignant nodules and 1 086 benign nodules, with diameters ranging from 3 mm to 30 mm, covering various shapes and positions.

For lung CT images, the first step is to convert them into binary images. Threshold segmentation methods can be employed to divide pixel values in the grayscale image into two categories: lung tissue and background. By selecting a suitable threshold, the lung tissue appears prominently in white in the binary image. Conducting connected component analysis on the binary image retains the two largest connected components, typically corresponding to the left and right lungs. Morphological operations are then applied to remove noise and excess parts such as blood vessels within the lung. There may be some larger holes within the lung parenchyma, and an algorithm based on connected regions is used to automatically fill these holes[20]. Overlaying the filled binary mask with the original image (500 pixel×500 pixel) completes the extraction of the lung parenchyma, as illustrated in Fig. 6.

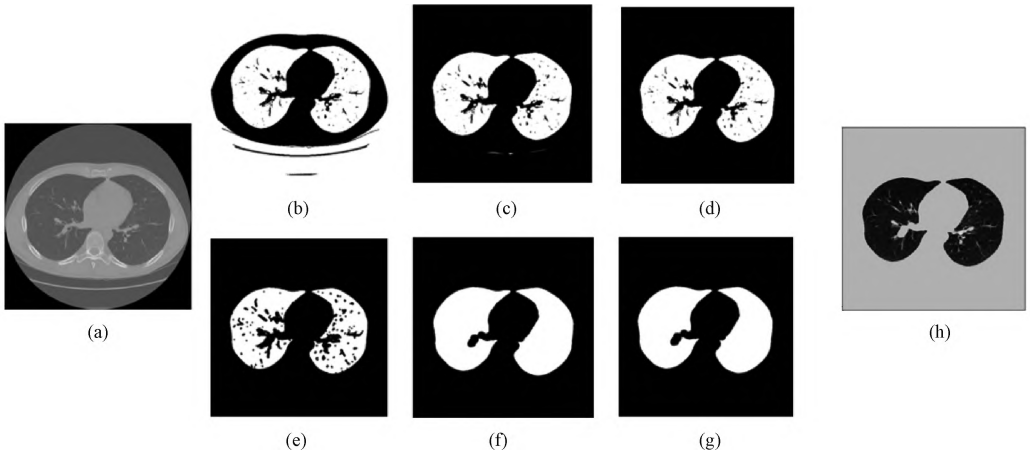


Fig. 6 Extraction process of lung parenchyma: (a) original image; (b) binarization; (c) clearing boundaries; (d) two largest connected regions; (e) removing noise; (f) relocating nodules attached to the lung wall; (g) filling small holes; (h) resulting image

3.2 Experimental environment and evaluation metrics

The experimental training platform used for this study consists of an Intel(R) Xeon(R) W-2265 CPU @ 3.50 GHz with 32 GB of RAM and 24 GB of GPU memory. The operating system is Ubuntu 18.04.6 LTS on a Linux environment, and the GPU version is GeForce RTX 3090. The Python version is Python 3.8, and the deep learning framework is PyTorch. The initial learning rate is set to 0.001, and the batch size is eight. The stochastic gradient descent (SGD) optimizer is used for training on the PyTorch 2.0.14 platform. The total number of training iterations is 500, with 100 iterations specifically for training the frozen backbone feature extraction network. The dataset comprises 1 186 images, with a training-to-testing split ratio of 9∶1.

For evaluation, the experiment employs average precision (AP) and sensitivity as performance metrics. AP is utilized to measure the accuracy of the object detection algorithm across different object categories. It is assessed by calculating the area under the precision-recall curve. Sensitivity is employed to gauge the recall rate of the object detection algorithm, indicating the proportion of real targets that the algorithm correctly detects. These metrics offer a holistic evaluation of the algorithm's precision and its capacity to effectively discern genuine targets.

3.3 Experimental results

To compare the effectiveness of our algorithm with different series of YOLO in pulmonary nodule detection, we conducted various experiments to compare AP and sensitivity. The experimental results are shown in Table 1.

Table 1 Comparison of results from different improvement strategies

Algorithm	AP/%	Sensitivity/%
YOLOv8	78.17	79.76
YOLOv8+CA	81.16	82.86
YOLOv8+DCN	80.26	83.49
YOLOv8+ASPP	79.61	80.97
YOLOv8+CA+DCN	84.07	85.17
YOLOv8+CA+ASPP	83.89	84.84
YOLOv8+DCN+ASPP	81.10	86.14
DCA-YOLO	86.28	89.30

This work aims to improve YOLOv8 by introducing deformable convolutions, multi-scale modules and the CA mechanism. The improved algorithm (DCA-YOLO),

compared with the original YOLOv8, achieved an increase of 8.11% (from 78.17% to 86.28%) in AP and 9.54% in sensitivity. Note that all percentage increases mentioned in this section are absolute improvements. We also compared the improvement of the individual module with the original YOLOv8. The addition of the CA module resulted in an increase of 2.99% in AP and 3.10% in sensitivity over the original YOLOv8, demonstrating that the CA module can make the algorithm more attentive to the spatial distribution of pulmonary nodules, thereby improving the accuracy. The addition of the DCN module resulted in an increase of 2.09% in AP and 3.73% in sensitivity over the original YOLOv8, proving that DCN can focus on the target object for feature extraction, enhancing the detection accuracy. The addition of the ASPP module resulted in an increase of 1.44% in AP and 1.21% in sensitivity over the original YOLOv8, demonstrating that ASPP captures a larger receptive field, improving the accuracy. The experiments show that each module contributes to the improvement of the original algorithm's detection accuracy, validating the effectiveness of the proposed improvement strategy. The detection comparison results of the algorithms are shown in Fig. 7.

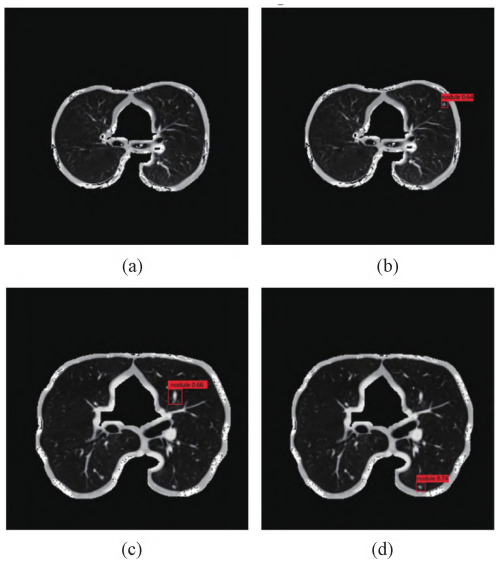


Fig. 7 Comparison of the results before and after improvement: (a) missed detection before improvement; (b) correct detection after improvement compared with Fig. 7(a); (c) false detection before improvement; (d) correct detection after improvement compared with Fig. 7(c)

This study also compared other algorithms in the YOLO series and other mainstream object detection algorithms. The specific comparison results are shown in Table 2. The proposed algorithm in this paper performs excellently in small object detection tasks, and its performance is significantly better than that of other mainstream object detection algorithms. This finding emphasizes the applicability of the proposed algorithm in tasks related to small object detection.

Table 2 Comparison of results from different mainstream algorithms

Algorithm	AP/%	Sensitivity/%
YOLOv4	78.74	74.68
YOLOv5	76.89	76.82
SSD	74.38	82.46
Faster RCNN	81.86	84.59
DCA-YOLO	86.28	89.30

4 Conclusions

This study proposes an enhancement to the YOLOv8 network by incorporating deformable convolutions and ASPP, along with the integration of the CA mechanism. This modification enables the network to concentrate on small targets while concurrently enlarging the receptive field without compromising resolution. Additionally, it leverages attention modules in different directions to gather contextual information on the targets, enhancing feature representation. The proposed algorithm effectively improves localization accuracy, reduces false detection rates, and minimizes missed detections. However, the current results indicate that the existing algorithm has not fully addressed the challenges of small object detection. In future work, optimizations can be explored in the following two aspects. For model training and optimization, a suitable model structure can be selected based on the specific algorithm and the model can be trained and optimized to extract effective features of pulmonary nodules. For the loss function of the algorithm, consideration should be given to replacing the loss function during the training process with a function that is more suitable for small object detection.

The research and application of pulmonary nodule detection will continue to advance. Through continuous improvement and innovation, there is potential to enhance the accuracy and automation of pulmonary nodule detection. This advancement aims to enhance early detection and treatment support for lung cancer.

References

[1] LYU J, LING S H. Using multi-level convolutional neural network for classification of lung nodules on CT images [C]//2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). New York: IEEE, 2018: 686-689.

[2] WAN C Y, MA L, LIU X B, et al. Computer-aided classification of lung nodules on CT images with expert knowledge [C]//Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. Bellingham: SPIE,

2021: 89.
[3] KIM J, KIM K H. Role of chest radiographs in early lung cancer detection [J]. *Translational Lung Cancer Research*, 2020, 9(3): 522-531.
[4] SUN L M, WANG Z R, PU H, et al. Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection [J]. *Computers in Biology and Medicine*, 2021, 133: 104357.
[5] ZHAO C, HAN J G, JIA Y, et al. Lung nodule detection via 3D U-net and contextual convolutional neural network [C]//2018 International Conference on Networking and Network Applications (NaNA). New York: IEEE, 2018: 356-361.
[6] LIU C Y, HU S C, WANG C H, et al. Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data [J]. *Quantitative Imaging in Medicine and Surgery*, 2020, 10(10): 1917-1929.
[7] YAN K, WANG X S, LU L, et al. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning [J]. *Journal of Medical Imaging*, 2018, 5(3): 036501.
[8] QIU B, HUANG Z Y, LIU X, et al. Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function [J]. *Biomedical Optics Express*, 2020, 11(2): 817-830.
[9] HUANG X J, SHAN J J, VAIDYA V. Lung nodule detection in CT using 3D convolutional neural networks [C]//2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). New York: IEEE, 2017: 379-383.
[10] KIM K, KIM K, JEONG S. Application of YOLOv5 and v8 for recognition of safety risk factors at construction sites [J]. *Sustainability*, 2023, 15(20): 15179.
[11] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 764-773.
[12] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//15th European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 833-851.
[13] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE, 2021: 13708-13717.
[14] CHETOUI M, AKHLOUFI M. Object detection model-based quality inspection using a deep CNN [C]//Sixteenth International Conference on Quality Control by Artificial Vision. [S. l.]: SPIE, 2023: 9.
[15] HE Z, HE D. Bilinear squeeze-and-excitation network for fine-grained classification of tree species [J]. *IEEE Geoscience and Remote Sensing Letters*, 2021, 18(7): 1139-1143.
[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.
[17] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]// European Conference on Computer Vision. Cham: Springer, 2018: 3-19.
[18] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Los Alamitos: IEEE, 2020: 1571-1580.
[19] SETIO A A A, TRAVERSO A, DE BEL T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge [J]. *Medical Image Analysis*, 2017, 42: 1-13.
[20] WU X J, WANG M Y, HAN B. An automatic hole-filling algorithm for polygon meshes [J]. *Computer-Aided Design and Applications*, 2008, 5(6): 889-899.

DCA-YOLO：一种基于注意力机制优化的 YOLOv8 肺结节检测算法

宋勇胜，刘国华*
东华大学 计算机科学与技术学院，上海 201620

摘　要：肺结节是肺癌的早期表现，然而肺结节在图像中占比较小，不仅导致医生阅片难度大，而且还可能出现误检和漏检的情况。针对这些问题，该文提出在 YOLOv8 网络的基础上加入可变形卷积和空洞空间金字塔池化（atrous spatial pyramid pooling，ASPP）并融合坐标注意力（coordinate attention，CA）机制，使得网络在聚焦小目标的同时又扩大感受野而不丢失分辨率。同时利用不同方向的注意力模块来聚集目标上的上下文信息，增强特征表达，有效提高定位精确度。所得算法在 LUNA16 的数据集上取得了良好的效果，相比于其他检测算法，该算法对肺结节检测的精度有一定改善。

关键词：肺结节；YOLOv8 网络；目标检测；可变形卷积；空洞空间金字塔池化（ASPP）；坐标注意力（CA）机制