scientific reports

Check for updates

OPEN

LN-DETR: cross-scale feature fusion and re-weighting for lung nodule detection

Dibin Zhou, Honggang Xu, Wenhao Liu & Fuchang Liu✉

With advancements in technology, lung nodule detection has significantly improved in both speed and accuracy. However, challenges remain in deploying these methods in complex real-world scenarios. This paper introduces an enhanced lung nodule detection algorithm base on RT-DETR, called LN-DETR. First, we designed a Deep and Shallow Detail Fusion layer that effectively fuses cross-scale features from both shallow and deep layers. Second, we optimized the computational load of the backbone network, effectively reducing the overall scale of the model. Finally, an efficient downsampling is designed to enhance the detection of lung nodules by re-weighting contextual information. Experiments conducted on the public LUNA16 dataset demonstrate that the proposed method, with a reduced number of parameters and computational overhead, achieves 83.7% mAP@0.5 and 36.3% mAP@0.5:0.95, outperforming RT-DETR in both model size and accuracy. These results highlight the superior detection accuracy of the proposed network while maintaining computational efficiency.

Keywords Deep learning, Object detection, Lung nodule, Medical image processing

Lung cancer is one of the malignant tumors with the fastest growing incidence and mortality rates, and it is also the leading cause of cancer-related deaths[1,2]. Due to the unique characteristics of lung cancer, most patients are diagnosed at an advanced stage, resulting in a five-year survival rate of only 10–20% in many countries[3]. However, if the relevant lesions are detected early and treated promptly, the survival rate can exceed 50%[4]. Lung nodules are one of the early signs of lung cancer, making early screening and diagnosis of lung nodules crucial. Computed tomography (CT) has been widelyused in medical diagnosis to obtain interior information of patients[5]. CT has a very high density resolution and is considered one of the most mature and effective imaging techniques for screening tasks. In recent years, with the widespread application of CT screening[6], the number of CT images has increased exponentially, which has placed significant pressure on radiologists and increased the risk of misdiagnosis due to subjectivity and fatigue.

In traditional diagnostic methods, radiologists need to combine personal experience with CT images to make a diagnosis, which is a cumbersome process. In recent years, with the rise of artificial intelligence, intelligent lung nodule detection methods have greatly facilitated the diagnostic work of radiologists. However, the detection of lung nodules poses various difficulties and challenges. Numerous studies have explored different methods and techniques to overcome these challenges, focusing primarily on the following aspects: Lung nodules vary in size, and small nodules are particularly difficult to distinguish from the background because they have similar texture features and occupy a very small portion of the image, making them easy to overlook[7,8]. Moreover, the limitations of traditional detection algorithms result in poor generalization ability, making them unsuitable for complex datasets[9]. The shape and edges of nodules are complex, including solid nodules, part-solid nodules, and ground-glass nodules, among others, which makes it challenging for detection models to capture all types of nodule features[10]. Since the density of lung nodules is similar to that of surrounding normal tissues, especially for less obvious nodules like ground-glass nodules, the contrast between the nodule and surrounding tissue is low, leading to a higher likelihood of missed detections by the model[11].

The complex image background in lung nodule detection tasks can severely affect the model's training efficiency. For example, lung nodules are often embedded in the complex anatomical structures of CT scans, such as blood vessels, alveoli, and lung textures. These background regions may have similar density and morphological features, leading to the extraction of distracting features that interfere with the model's learning. Additionally, the significant structural differences in the lungs of different patients, as well as the varying degrees of background texture disruption caused by different types of lung diseases, can further hinder the model's feature learning.

School of Information Science and Technology, Hangzhou Normal University, Hang Zhou 311121, China. ✉email: liufc@hznu.edu.cn

The advent of deep convolutional neural networks (CNN) has, to some extent, ameliorated these issues, owing to their exceptional performance in capturing local features. This enables CNN to learn from induced biases and be robust against translation variances[12]. However, CNN have a bias due to the nature of local feature extraction[13]. Besides, their capacity to retain feature dependencies is compromised due to the progressive expansion of their visual field across layers[14]. The emergence of Detection Transformer(DETR), which adept at map long-range dependencies between features, has effectively addressed this issue[15]. Its advantage lies in its ability to transform the detection problem into an unordered sequence output problem, turning traditional dense detection into sparse detection[16]. Besides, DETR typically possess more parameters than CNN, enabling them to outperform in complex images. However, in lung nodule detection, DETR still exhibits certain shortcomings, such as insufficient sensitivity for detecting lung nodules, and high computational costs, which limit its real-time application and may result in false positives or false negatives in complex backgrounds. Moreover, such networks consume relatively large amounts of GPU resources, increasing hardware costs. Therefore, lung nodule detection in complex scenarios requires more efficient and accurate solutions.

In summary, to address the issues of insensitivity to small lung nodules and overall low detection accuracy, this paper proposes LN-DETR (Lung Nodule DETR), a model based on RT-DETR, designed for high-precision detection of lung nodules in CT images with complex backgrounds. The main contributions of this study are as follows:

- A feature extraction module, FasterBlock, is proposed to improve the backbone network, increasing the network's feature extraction capability and detection accuracy while reducing floating-point computations and the number of parameters;
- A Deep and Shallow Detail Fusion (DSDF) module is introduced to effectively enhance the network's ability to aggregate cross-scale feature from both shallow and deep features, improving detection accuracy;
- A Joint Context Feature (JCF) module is proposed to improve the network's downsampling function. This module uses global contextual semantic information to re-weight the joint features intelligently across channels, emphasizing useful components and suppressing irrelevant ones, thereby improving detection accuracy during the downsampling process.

Related work

The early detection and accurate classification of lung nodules play a crucial role in improving patient prognosis and survival rates[17]. However, accurate detection and classification of these nodules are often challenging due to their varying sizes, shapes, and densities, as well as the complex background of lung structures in computed tomography (CT) scans. These factors significantly increase the difficulty of detection, prompting many studies to explore various methods and techniques[18].

In this section, we summarize the research on lung nodule detection. We primarily focus on common methods in lung nodule detection tasks:

Methods based on CNN in lung nodule detection

CNN learn feature representations directly from data, eliminating the need for manual feature extraction. This development substantially enhanced the accuracy and efficiency of detection[19]. As a result, CNN play a crucial role in medical image processing due to their excellent performance[20]. For predicting lung cancer, Yu Gu et al.[11], developed a novel technique to assist the radiologist to detect Lung nodule using multi-scale prediction strategy combined with 3D deep convolution neural network thereby provides a second opinion to the radiologist on efficient detection if-nodule-which is an important step for effective diagnosis of Lung Cancer. And when compared with 2 Dimensional-convolution-neuralnetwork 3 Dimensional-Convolution-neural-network utilizes rich spatial contextual information and results in generating more discriminative features. The proposed technique is tested on 889 thin slice CT scans with 1187-nodules. The sensitivity of the techniques reached 87.94%. However, due to the structural limitations of the 3 Dimensional-Convolution-neural-network, the training time is long, which significantly affects detection efficiency. Liu et al.[21] applied an improved drip segmentation algorithm to preprocess CT images, filtering out irrelevant information and obtaining enhanced lung parenchyma images with better visualization. Subsequently, they adjusted the structure and optimized the parameters of the YOLOv4[22] model to improve its performance in lung nodule detection. The results indicate that while the model's accuracy is relatively low at only 80.1%, it demonstrates significantly faster processing speed compared to other networks, achieving 47.6 frames per second (FPS). Wang et al.[10], leveraging the basic symmetric U-shaped architecture of U-Net, redesigned two new U-shaped deep learning (U-DL) models that were expanded to six levels of convolutional layers. After that, an ensemble layer was used to combine the two U-DL models into the H-DL model. The results indicated that the H-DL model could achieve segmentation accuracy comparable to radiologists'segmentation for nodules with wide ranges of image characteristics. However, its training dataset only includes nodules ranging from 7 to 45 mm in size, resulting in poor detection performance for nodules smaller than 7 mm. Despite the significant success of CNN in lung nodule detection, but they each have their own limitations: the training time is long, the ability to detect small targets is weak , low overall detection accuracy.

Methods based on transformer architecture in lung nodule detection

The Transformer, originally developed for Natural Language Processing (NLP), has been successfully adapted for object detection tasks through the DETR framework[23]. This architecture employs a sophisticated attention mechanism for image processing, which is systematically organized into three distinct stages: the patch embedding layer, encoder, and decoder components[24,25]. DETR represents a groundbreaking application of Transformer architecture in the domain of lung nodule detection. Although it still maintains the utilization of CNN in its

backbone network to extract image features, its core innovation lies in leveraging the Transformer architecture to process feature maps, thereby enabling end-to-end object detection. The emergence of DETR is considered an alternative method of performing object detection tasks other than CNN[26]. However, due to the structural characteristics of DETR, this kind of method still face significant drawbacks, including high computational resource consumption and relatively poor detection capabilities for small targets. RT-DETR[27] optimizes these issues by introducing improvements such as a design without NMS and two-layer attention mechanisms, reducing processing latency and enhancing the detection ability for objects of varying sizes, particularly small targets. However, the accuracy of lung nodule detection in complex CT images remains unsatisfactory.

In conclusion, these common lung nodule detection methods share some common disadvantages: First, the sensitivity to lung nodules is insufficient, resulting in low overall detection accuracy. Second, the ability to detect small targets is weak, which leads to unsatisfactory performance in the lung nodule detection task. Therefore, this study aims to address these two issues in order to improve the model's performance in lung nodule detection. Additionally, although Transformer models have shown excellent performance, there is limited research applying them to lung nodule detection tasks. Consequently, in this research, RT-DETR was selected as the baseline for further investigation.

Methods

Overview

The structure of the LN-DETR algorithm is shown in Fig. 1. The algorithm consists of three main components: backbon, neck, and head. First, the lightweight FasterBlock is used to extract features from lung nodule images, obtaining three feature layers at different levels-shallow, mid, and deep-that contain rich semantic information. These three output feature layers are then fed into the neck network. Second, the neck network processes the feature layers, transforming them into utilizable information. The neck network consists of the AIFI(Attention-based Intra-scale Feature Interaction) and a feature fusion network. The AIFI encodes the high-level features, capturing more semantic information while significantly reducing computational load and improving processing speed without compromising model performance[27]. The feature fusion network uses the DSDF and JCF to perform semantic information filtering and feature fusion along two paths-bottom-up and top-down-converting multi-scale features into feature maps at different resolution ratio-high, medium, and low. image features. Finally, the head network performs decoding operations on the feature maps of varying resolution ratio, where the IoU-aware query and auxiliary prediction heads are used to iteratively optimize and generate predicted bounding boxes.

Light-weight FasterBlock module

The backbone network of RT-DETR has a large computational load and high hardware requirements. Additionally, due to the specificity of lung nodule images, there is a certain sensitivity requirement for small targets in the feature extraction network. To further improve the model's feature extraction capability, this paper proposes the FasterBlock module to replace the original feature extraction module in the network. Figure 2 shows the structure of the FasterBlock module. This module draws extensively from the design principles of FasterNet[28], particularly the use of Partial Convolution (PConv), which effectively reduces memory access and computational redundancy(details of the experiments can be found in Table 3). To enhance the feature extraction network's detection performance for lung nodules, this paper adds a 3x3 convolution layer before the PConv, forming a dual-path residual block structure with subsequent convolution layers. This structure aims to create a more comprehensive feature representation, improving the sensitivity and recognition performance for lung nodules while further reducing information loss during the feature extraction process.

The core of the FasterBlock structure lies in the basic operator PConv. Figure 2 shows a comparison between this operator and convolution. PConv leverages the redundancy in the feature map by performing convolution operations on only part of the channels, while keeping some input channels unchanged. During memory access, only the first or last consecutive channels are considered as representatives of the entire feature map for computation. Additionally, PConv preserves part of the original features of the input image within the feature map, providing richer semantic information for high-level feature fusion in the subsequent feature fusion network. Without compromising generality, the input and output feature maps have the same number of channels. Therefore, the floating-point operations and memory access volume of PConv can be expressed by Eqs. (1) and (2):

$$h * w * k^2 * c_p^2, \tag{1}$$

$$h * w * 2c_p + k^2 * c_p^2 \approx h * w * 2c_p. \tag{2}$$

Taking a typical local ratio as an example, when $r = c_p/c = 1/4$, the FLOPs of PConv are only 1/16 of those of a standard convolution, and PConv also has a smaller memory access volume, being only 1/4 of the memory access volume of a standard convolution with $r = 1/4$.

To fully and effectively utilize the semantic information extracted by PConv, two pointwise convolutions (PwConv) are added after PConv, forming an inverted residual block. This design leverages redundancy between filters to further reduce floating-point operations (FLOPs), thereby improving the computational speed of the feature extraction network. In addition to the above operations, normalization and activation layers are also essential for high-performance neural networks. However, excessive use of such layers in the network can limit the diversity of features, thereby reducing the performance of feature extraction and slowing down the overall computation speed of the network. Therefore, these layers are only placed after the Conv and the first PwConv, maintaining feature diversity while achieving faster computation. In this paper, batch normalization (BN) and

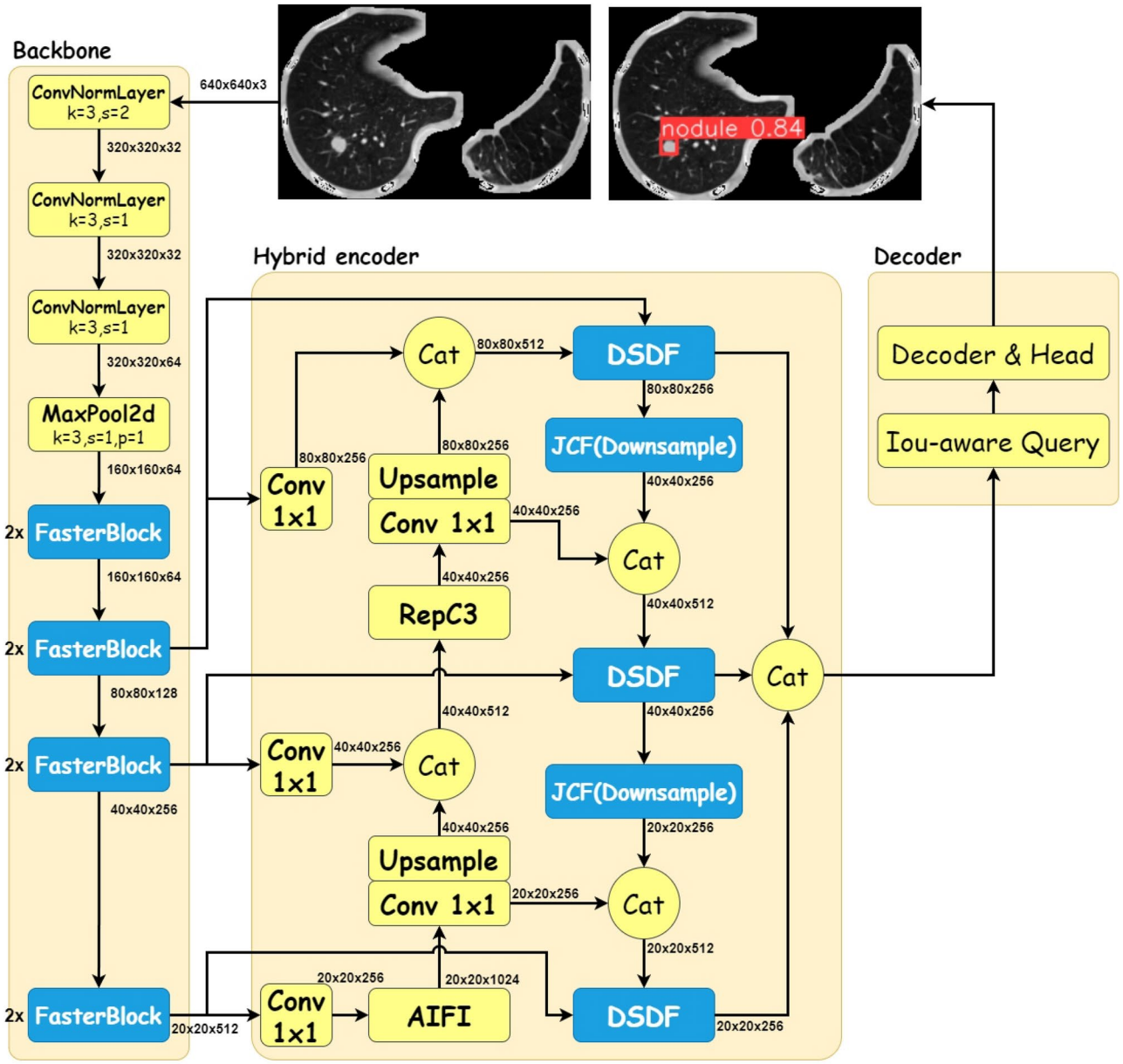


Figure 1. The structure of the LN-DETR. The yellow parts represent the modules from the baseline, RT-DETR, while the blue parts represent the proposed modules.

ReLU are used as the normalization and activation layers. The advantage of these layers is that they can be merged with adjacent convolution layers, enabling faster inference.

Deep and shallow detail fusion module

As an essential component of the hybrid encoder, the design of the feature fusion module is of paramount importance. In this regard, this paper investigates and conducts experiments on other feature fusion modules(details of the experiments can be found in Table 4). It was found that most of these modules focus on shallow image features and do not efficiently utilize deep image features. Moreover, they do not specifically optimize for lung nodules.

In the task of lung nodule detection in images, small object detection accounts for a significant proportion. Small objects refer to target objects that occupy a relatively small area in the image, and their size is usually insufficient to provide enough feature information, making accurate detection difficult. After feature extraction by the backbone network's convolution modules, shallow image features have high resolution, containing more detailed information but less semantic content and more noise. In contrast, deep image features contain stronger semantic information but lack fine details. The design of the feature fusion module aims to better utilize the features extracted by the backbone network by reprocessing and reasonably using features from different stages. To enhance the model's perception of lung nodules, the feature fusion network in the RT-DETR hybrid encoder

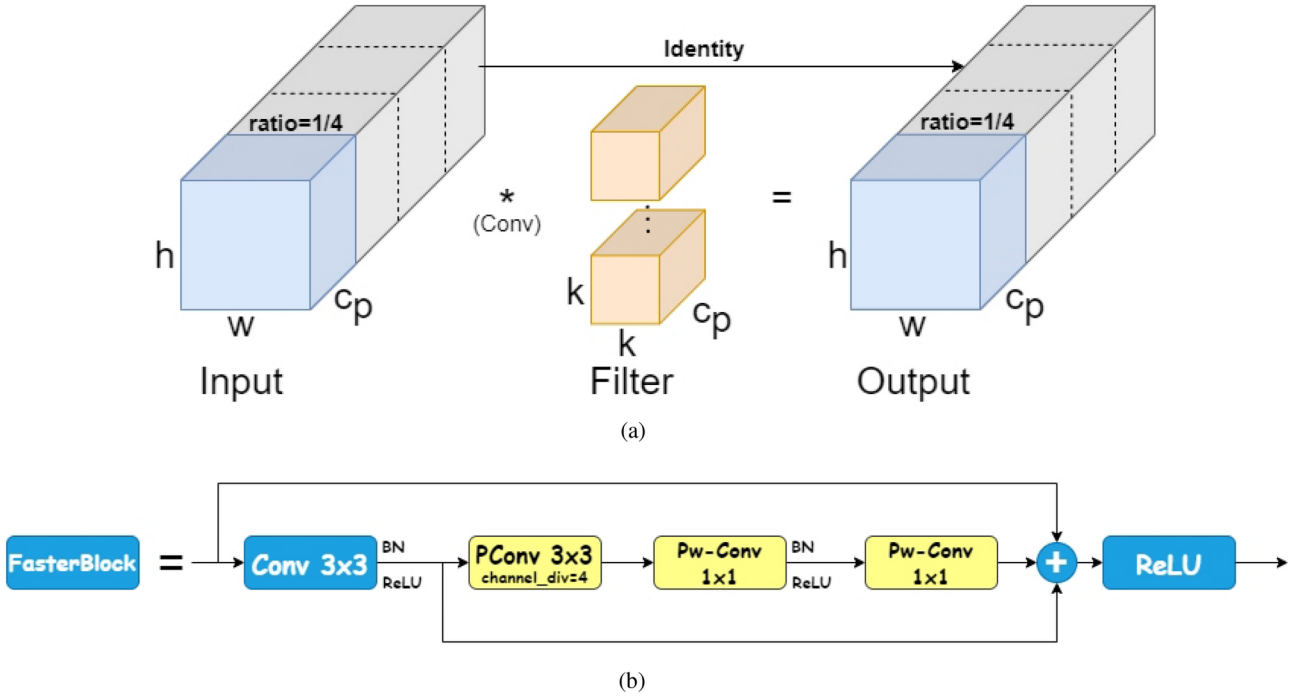


Figure 2. The structure of the FasterBlock. In the architecture diagram, the yellow parts originate from the original FasterNet like pointwise convolutions, while the blue parts represent the proposed modules. These modules further enhance its feature extraction capability for lung nodules.

is improved. This paper designs a fusion module, DSDF, that can merge both shallow and deep detail features and introduces it into the network to add an additional feature fusion layer for lung nodules. Its structure is shown in Fig. 3.

The DSDF module consists of three main components:

The first step is to further extract feature information from the image using group convolution[29] and channel transformation techniques[30], resulting in the extracted deep image features. First, group convolution is utilized for its small parameter size and computational load, as well as its ability to prevent overfitting, ensuring robustness and generalization during the feature extraction process. Then, channel transformation is applied to facilitate the exchange of feature information within groups. The feature extraction process is shown in Eqs. (3) and (4), where F_i represents the input feature information, F_d^i represents the deep feature information, $\oplus$ denotes element-wise summation, $PwConv^2$ represents 2 consecutive pointwise convolutions, *Cat* represents concatenation along the channel dimension, and *CS* and *GConv* denote channel transformation and group convolution, respectively.

$$F_c = Cat(CS(GConv(Conv(F_i))), Conv(F_i)), \tag{3}$$

$$F_d^i = F_c \oplus PwConv^2(F_c), \tag{4}$$

The second step is to further enhance the detection of lung nodules by strengthening the shallow and deep image features. In the channel dimension, the deep image features are concatenated with the shallow image features obtained from the backbone network. These concatenated features are then passed into a channel attention module consisting of convolution and pooling operations to generate attention weights. These weights are then applied to the original features via element-wise multiplication, and the resulting features are added to the original features from another branch to enhance their information representation ability[31]. The feature enhancement process is shown in Eqs. (5) and (6), where $\hat{F}_d^i$ and $\hat{F}_s^i$ represent the enhanced features, $\otimes$ denotes element-wise multiplication, F_s^i represents the shallow image features, and δ and G represent the sigmoid function and global average pooling layer, respectively.

$$\hat{F}_d^i = F_d^i \oplus (Conv(F_s^i) \otimes \delta(PwConv^2(GAP(Cat(F_d^i, Conv(F_s^i)))))), \tag{5}$$

$$\hat{F}_s^i = Conv(F_s^i) \oplus (F_d^i \otimes \delta(PwConv^2(GAP(Cat(F_d^i, Conv(F_s^i)))))), \tag{6}$$

The third step is to concatenate the enhanced shallow and deep image features along the channel dimension and input them into parallel channel attention modules and spatial attention modules for the final weight fusion. The weight fusion process can be expressed by Eqs. (7), (8), and (9), where A_C^i and A_S^i represent the channel attention and spatial attention, W^i resprent the final fusion weights:

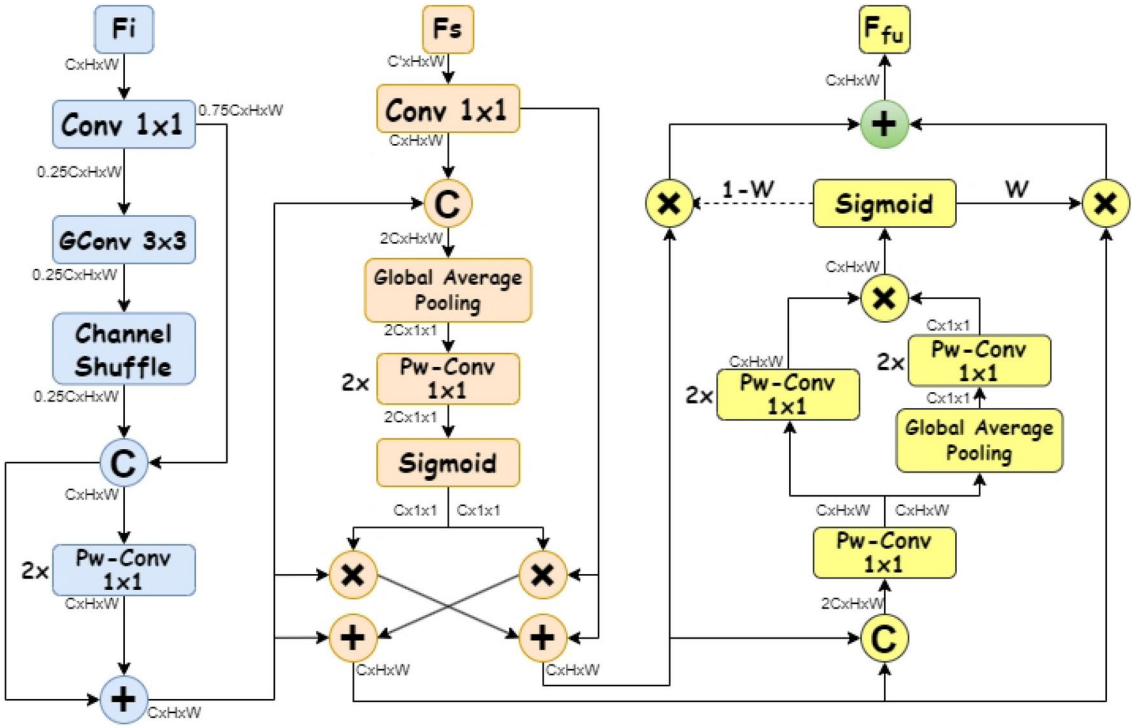


Figure 3. The structure of the DSDF. Deep-level image features are extracted and then feature enhancement is applied to both deep and shallow image features. Finally, weight fusion of the deep and shallow image features is performed.

$$A_C^i = PwConv^2(GAP(PwConv(Cat(\hat{F}_d^i, \hat{F}_s^i)))), \tag{7}$$

$$A_S^i = PwConv^2(PwConv(Cat(\hat{F}_d^i, \hat{F}_s^i))), \tag{8}$$

$$W^i = \delta\ (A_C^i \otimes A_S^i), \tag{9}$$

Since shallow and deep features are complementary, the weight of one feature can be represented as W^i, and the weight of the other feature can be represented as $1 - W^i$. Therefore, the feature fusion process can be expressed by Eq. (10):

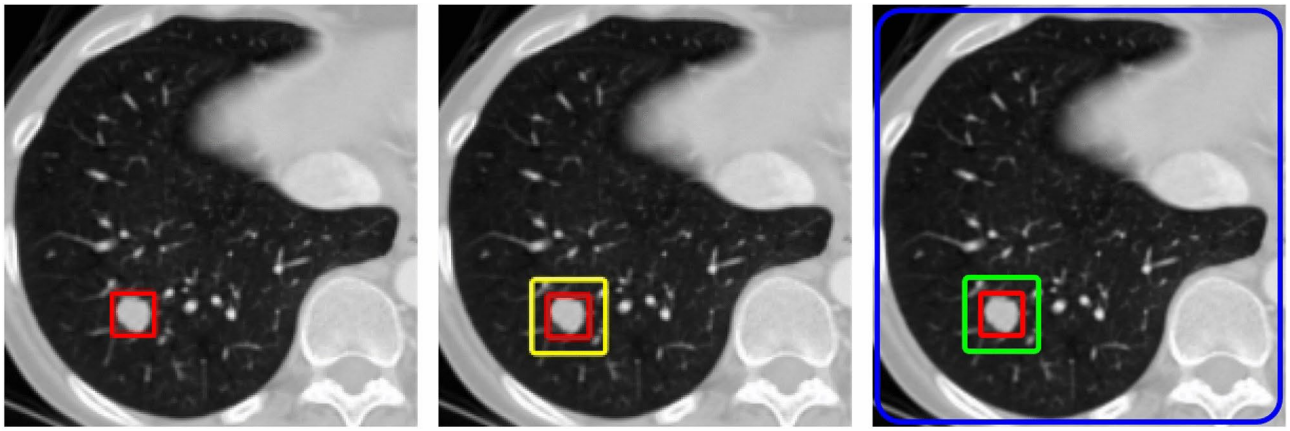
$$F_{fu}^i = (W^i \otimes \hat{F}_d^i) \oplus ((1 - W^i) \otimes \hat{F}_s^i). \tag{10}$$

The DSDF module, while balancing model complexity and performance, enhances the utilization of shallow image features obtained from the backbone network. It enables complementary interactions between abstract feature information at different levels, providing the model with feature inputs that are both comprehensive and accurate for subsequent tasks.

Joint context feature module

Downsampling reduces the resolution of feature maps, shrinking the spatial dimensions of the input image. For small targets such as lung nodules, their features are prone to being overly compressed during the downsampling process, leading to partial loss of information. Given the limited number of pixels representing lung nodules, multiple rounds of downsampling may leave only a minimal number of feature points, making it difficult for the model to accurately recognize the nodules.

The downsampling module in RT-DETR employs computationally efficient 3×3 convolutional blocks, which effectively reduce the number of model parameters and improve computational efficiency while maintaining a certain level of detection accuracy. However, its performance in lung nodule detection tasks remains suboptimal (details in Table 5). To address this, the Joint Context Feature (JCF) module is introduced to enhance the network's downsampling capabilities. We designed the JCF module to mimic the human visual system by leveraging surrounding contextual information to enhance detection. As illustrated in Fig. 4a, if the human visual system tries to locate the red region by focusing solely on the target itself, it becomes challenging because the target occupies only a small portion of the overall area. In Fig. 4b, defining the white region as the surrounding environment of the red box and utilizing information from the white box makes it easier to detect the target. For Fig. 4c, incorporating information from the entire scene (green box) in conjunction with the red and white regions further reduces the difficulty of target detection.



(a) (b) (c)

Figure 4. The illustration of target detection by leveraging surrounding contextual information. The JCF module mimics the human visual system by leveraging surrounding contextual information to enhance detection.

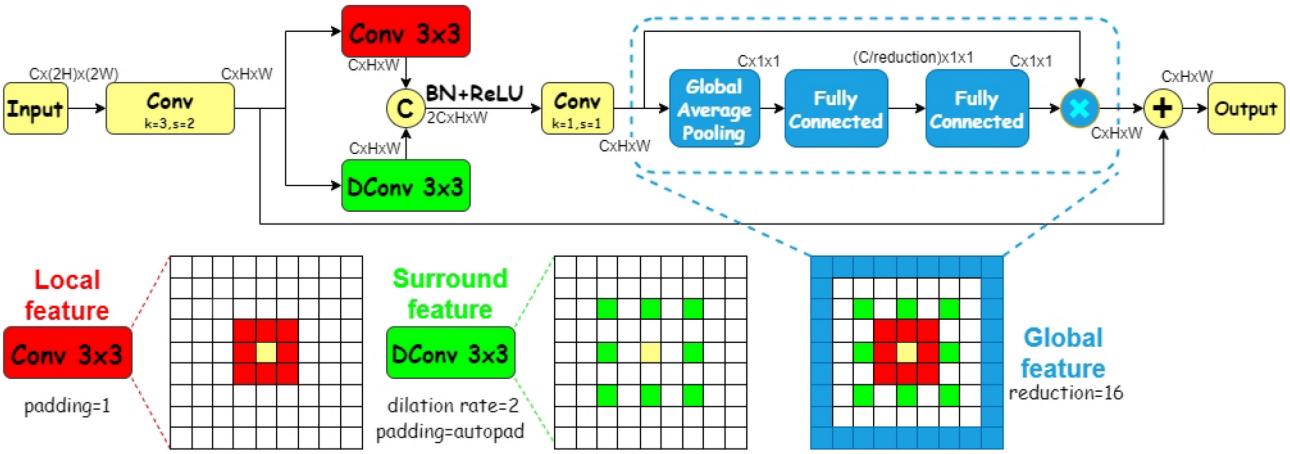


Figure 5. The structure of the JCF. JCF employs different convolutional modules to extract local and surrounding features, followed by fully connected layers to capture global features. Finally, a residual block with element-wise multiplication re-weights these features, enhancing lung nodule detection.

Building on the above concept, the JCF is introduced in the downsampling process to leverage contextual information for effective lung nodule detection. Initially, the dimensions of the feature map are reduced by a 3x3 convolutional block with a stride of 2. Second, conventional convolution is used to learn local features from the 8 neighboring feature vectors of the target. Since dilated convolution (DConv) possesses a relatively larger receptive field, it can effectively learn contextual information surrounding the target[32]. These two parts of features are then fused to obtain joint features. We simply design the feature fusion as a concatenation layer followed by the Batch Normalization(BN) and the Rectified Linear Unit(ReLU). Subsequently, we employ a bottleneck layer to compress the dimensions of the fused features, extracting the most representative information. In the next step, a global context feature extractor is constructed using global average pooling (GAP) and fully connected layers. This extractor multiplies the joint features element-wise within a residual block to re-weight them, enhance the weights of target features while suppressing the weights of other information, further optimizing the joint features. Finally, the input is concatenated with the optimized joint features for a global residual learning process, which mitigates the gradient vanishing issue caused by the element-wise multiplication residual block. The structure of the JCF module is illustrated in Fig. 5. The process is shown in Eqs. (11), (12) and (13), where F_i represents the input feature information, F_{down} represents the feature map after its dimensions have been reduced, F_{joi} represents joint features, F_o represents the output feature information, *DConv* represents dilated convolution, *Cat* represents concatenation along the channel dimension, $\oplus$ denotes element-wise summation, $\otimes$ denotes element-wise multiplication, *GAP* represents the global average pooling layer, $\hat{FC}$ and $\check{FC}$ represent the fully connected layers that reduce and restore the dimensionality of the features, respectively.

$$F_{down} = Conv^{3x3}(F_i), \tag{11}$$

$$F_{joi} = Conv^{1x1}(Cat(Conv^{3x3}(F_{down}), DConv(F_{down}))), \tag{12}$$

$$F_o = \hat{FC}(\check{FC}(GAP(F_{joi}))) \otimes F_{joi} \oplus F_{down}. \tag{13}$$

Results

Datasets introduction

Medical imaging datasets in the field of lung nodule detection are extremely scarce. This study utilizes the Lung Nodule Analysis 16(LUNA16) dataset for evaluation. The dataset originates from the Lung Image Database Consortium (LIDC-IDRI), which comprises 1,018 low-dose lung CT scans and is currently the largest and most widely used public dataset in this domain. LUNA16 is a subset of LIDC-IDRI. It is constructed by removing CT scans with a slice thickness greater than 3 mm and lung nodules smaller than 3 mm from LIDC-IDRI, resulting in 888 low-dose lung CT scans in MHD format.

Each CT scan in the LUNA16 dataset includes a series of axial slices of the thorax, with the number of slices varying depending on the scanning equipment, slice thickness, and individual patients. Since the CT images in LUNA16 cannot be directly used for training neural networks, this study preprocesses the original dataset by converting it into PNG format. Subsequently, a deep learning-based method is employed to segment the lung parenchyma in the PNG images, excluding surrounding bones and muscles to minimize interference and enhance detection accuracy and recognition rates. Examples of the PNG images and their segmented counterparts are shown in Fig. 6.

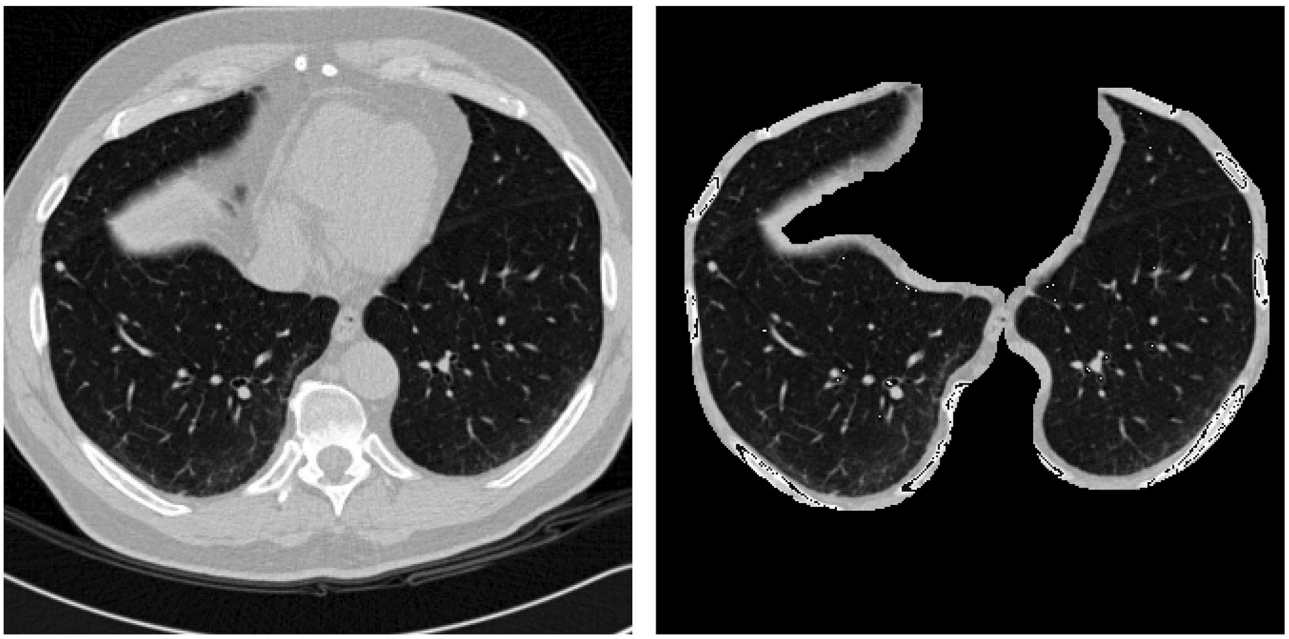
For deep learning, more training data generally enhances model detection performance. However, large and diverse medical image datasets are rare[33], making data augmentation necessary. In this study, 1,186 lung nodule images were expanded to 3,813 images through operations such as flipping, cropping, mirroring, rotating, and noise addition. Subsequently, 70% of the images were randomly selected as the training set, 10% as the validation set, and 20% as the test set[28].

Experimental settings

This study utilized an Intel Xeon CPU E5-2650 v3 2.30GHz as the CPU processor, with 16GB of RAM, a 1TB hard drive, and an RTX 3060 GPU with 12GB of memory, running on the Windows 10 operating system. The network was implemented using PyTorch, with GPU acceleration provided by CUDA v11.8 and CuDNN v8.9. The training parameters were set as follows: a batch size of 8, an initial learning rate of 0.0001, a final learning rate of 1.0, momentum of 0.9, and 300 epochs. No pre-trained weights were used during training.

Evaluation metrics

In this study, the experimental results were evaluated using the following metrics: Precision (P), Recall (R), mean Average Precision at Intersection over Union (IoU) is 0.5 (mAP@.5), mean Average Precision at IoU from 0.5 to 0.95 with a step size of 0.05(mAP@.5:.95), total number of parameters, Floating-point Operations



(a) (b)

Figure 6. Original images and segmented images.

(FLOPs), and Frames Per Second (FPS). Among these, IoU is the ratio of the intersection area to the union area between a predicted bounding box and a ground truth bounding box. mAP@.5 and mAP@.5:.95 signify detection capabilities across various IoU thresholds[34]. The number of parameters reflects the size and complexity of the model, while FLOPs represent the computational cost, both of which are used to assess the complexity of the network. FPS is used to measure the detection speed of the network. The formulas for the remaining metrics are as follows:

$$P = TP/(TP + FP), \tag{14}$$

$$R = TP/(TP + FN), \tag{15}$$

$$AP = \int_0^1 P(R)dR, \tag{16}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i, \tag{17}$$

$$IoU = \frac{A_{Intersection}}{A_{Union}}. \tag{18}$$

True Positive (TP) represents the number of lung nodule correctly detected, False Positive (FP) represents the number of incorrectly detected lung nodule, and False Negative (FN) represents the number of missed defects. n is the number of defect categories. In the LUNA16 dataset, n is set to 1. $A_{Intersection}$ is the intersection area between a predicted bounding box and a ground truth bounding box. A_{Union} is the union area.

Comparison with state-of-the-arts

To further validate the superiority of the LN-DETR network proposed in this paper, a series of comparative experiments were conducted between various mainstream object detection networks and LN-DETR. All experiments were performed under the same hardware, software, and parameter settings, using the same dataset. The results are shown in Table 1.

The YOLO series models are well-known for their high accuracy and fast speed. However, as shown in the experimental results in Table 1, LN-DETR does not show a significant difference in parameter count and computational load compared to YOLO, while demonstrating a clear advantage in detection accuracy. The mAP@0.5 and mAP@0.5:0.95 of the LN-DETR is 5.3% and 3.6% higher than the latest YOLOv9.

We compared LN-DETR with LSKNet, which is capable of dynamically adjusting the receptive field, PKINet that employs parallel depth-wise convolution kernels to capture dense texture features, DEANet that utilizes detail-enhanced attention blocks to bolster the model's feature learning capabilities, and ConvNeXt, a CNN model inspired by the Transformer architecture. The LN-DETR surpasses the aforementioned algorithms across all four metrics: Precision, Recall, mAP@0.5, mAP@0.5:0.95. Among them, LN-DETR outperforms the other algorithms in metric mAP@0.5 by 8.1%, 6.6%, 4.2% and 3.9%, respectively. This performance meets the precision requirements of the lung nodule detection task. Additionally, this paper also replicated and compared recent improvements to Vision Transformer(ViT) and DETR networks by other researchers. Although these models have optimized and enhanced the model in various aspects, they have not specifically considered small object detection tasks like lung nodule detection. Compared to these models, LN-DETR still achieves satisfactory results. On metric mAP@0.5, it outperforms other models by at least 4% (CrossViT) and up to 6.4% (RT-DETR), demonstrating that the algorithm exhibits higher accuracy and reliability in the lung nodule detection task.

Modle	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	FLOPs/G	FPS
YOLOv5-m	70.3	72.0	70.0	29.2	21.2	**49**	**75**
YOLOv8-m	77.8	63.0	69.5	29.1	25.8	78.7	66
YOLOv9-m[35]	80.5	74.8	78.4	32.7	20	76.3	74
LSKNet[36]	79.6	75.4	75.6	34.3	31.0	90.7	54
PKINet-s[37]	82.3	71.8	77.1	33.8	30.8	90.7	66
DEANet[38]	82.1	75.2	79.5	34.8	20.4	59.2	64
ConvNeXt-t[39]	80.7	74.8	79.8	34.6	47.8	262	26
(ViT)DaViT-t[40]	83.1	76.7	78.2	34.4	38.5	244	23
(ViT)CrossViT[41]	83.8	75.6	79.7	35.5	44.9	56.6	57
(ViT)RMT-t[42]	79.6	76.1	79.2	35.7	23	199	20
(ViT)TransNeXt-t[43]	81.9	74.8	79.0	34.9	47.9	273	22
MFDS-DETR[44]	83.5	77.7	79.6	35.8	**18.1**	53.3	72
RT-DETR	82.2	74.2	77.3	33.4	19.97	57.3	71
LN-DETR(ours)	**85.7**	**82.4**	**83.7**	**36.3**	19.18	51.2	66

Table 1. Result of comparative experiments.

In terms of computational cost, the proposed algorithm also significantly outperforms other algorithms. It reduces FLOPs by 2.1G compared to MFDS-DETR, which has the lowest computational cost among the compared methods, and by 6.1G compared to the baseline model RT-DETR. This demonstrates the feasibility of the optimizations aimed at reducing computational burden. These optimizations enhance the model's adaptability to devices with limited computational resources.

In conclusion, the experimental results demonstrate that the LN-DETR network proposed for lung nodule detection outperforms these improved networks on the LUNA16 dataset.

Ablation studies and module comparisons

Ablation studies

To verify the performance of each module, detailed ablation studies were conducted using the LUNA16 dataset. The RT-DETR model was used as the baseline, and experiments were carried out under the same environment and parameter settings. In each experiment, different modules were added to evaluate the impact of each module on detection performance. The specific results are shown in Table 2.

Experiment 1 demonstrates the performance of the RT-DETR algorithm in the lung nodule detection task, providing a reference for evaluating the effects of the improvements introduced by each module.

Experiments 2, 3, and 4 demonstrate the individual performance of each module within the RT-DETR network. Experiment 2 reconstructs the feature extraction module in the backbone network with FasterBlock. This module not only inherits the lightweight and fast advantages of FasterNet but also improves detection accuracy. Experiment 3 introduces the DSDF module to enhance the feature fusion layer for lung nodule detection. Compared to Experiment 1, DSDF significantly improves the mAP while keeping the model size almost unchanged, and also slightly boosts both Precision and Recall. This suggests that the module effectively utilizes the semantic information from both shallow and deep layers to enhance the model's ability to detect lung nodules. Experiment 4 replaces the downsampling module with JCF, and the results show a significant improvement in detection accuracy, indicating that the module effectively addresses the issue of feature loss during downsampling, thereby improving precision.

Experiments 5, 6, and 7 demonstrate the combined performance of various modules within the RT-DETR network. Experiment 5 shows the results of combining FasterBlock with DSDF in RT-DETR. This combination leads to an overall improvement in detection accuracy, while simultaneously reducing both the parameter count and computational load. This demonstrates that DSDF effectively compensates for the detection accuracy limitations of FasterBlock by utilizing the features extracted from the backbone network. Experiment 6 examines the combination of FasterBlock and JCF. The results of this combination are slightly lower than those in Experiment 4, but still show an overall improvement compared to Experiment 1. This indicates that FasterBlock can work synergistically with JCF without causing any significant negative impact. Experiment 7 presents the combination of DSDF and JCF. Compared to Experiment 1, this combination shows a significant improvement in detection accuracy. However, it performs less favorably in terms of parameter count and computational load. In comparison to Experiment 4, the combination results in a noticeable reduction in these aspects. This suggests that DSDF and JCF complement each other well in detection tasks, while FasterBlock effectively balances the computational drawbacks of these two modules.

Experiment 8 represents the LN-DETR network proposed in this paper. It can be observed that this network integrates the advantages of each module, leveraging their strengths to compensate for weaknesses, significantly improving the network's ability to detect lung nodule targets. Compared to Experiment 1, precision increased by 3.5%, recall improved by 8.2%, mAP@.5 rose by 6.4%, and mAP@.5:.95 enhanced by 2.9%. Additionally, both parameter count and computational load were slightly reduced. This demonstrates that the LN-DETR network achieves a good balance between detection accuracy and computational efficiency.

Module comparisons

To demonstrate the light-weight advantages of FasterBlock over the original feature extraction module, BasicBlock, and other feature extraction modules, this paper compares the computational load and the number of parameters of these modules. The comparison results are shown in Table 3. Compared to other modules, FasterBlock has significant advantages in both computational load and parameter count, indicating that FasterBlock has lower computational overhead during feature extraction, which can effectively reduce the overall size of the network.

Exp	Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	FLOPs/G	FPS
1	RT-DETR	82.2	74.2	77.3	33.4	19.97	57.3	71
2	FasterBlock	82.7	73.9	78.9	34.7	16.89	49.8	**75**
3	DSDF	82.5	77.2	81.2	34.5	19.82	53.9	70
4	JCF	83.8	76.2	81.1	34.7	22.43	62.1	72
5	FasterBlock+DSDF	84.5	80.1	82.8	37.0	**16.73**	**46.5**	67
6	FasterBlock+JCF	85.1	72.9	80.7	34.1	19.34	54.6	70
7	DSDF+JCF	85.1	78.6	80.9	35.3	22.17	58.3	67
8	LN-DETR(ours)	**85.7**	**82.4**	**83.7**	**36.3**	19.18	51.2	66

Table 2. Result of ablation studies.

Feature Extraction Module	FLOPs/G	Params/M
BasicBlock	7.82	15.2
StarBlock[45]	8.49	16.5
Conv3XCBlock[46]	10.36	20.2
FasterBlock(ours)	**5.99**	**11.6**

Table 3. Comparison of Computational Overhead.

Module	Precision/%	Recall/%	mAP@.5/%	mAP@.5:.95/%
CCFM(base)	82.2	74.2	77.3	33.4
ASSF[47]	81.1	73.9	76.3	34.8
CAFM[48]	80.2	73.9	76.2	33.9
CSFCN[49]	80.6	76.1	77.2	**34.9**
SlimNeck[50]	81.6	76.1	80.9	34.7
DSDF(ours)	**82.5**	**77.2**	**81.2**	34.5

Table 4. Comparison of feature fusion modules.

Module	Precision/%	Recall/%	mAP@.5/%	mAP@.5:.95/%
Baseline	82.2	74.2	77.3	33.4
HaarWavelet[51]	75.7	73.2	74.2	32.9
WaveletPooling[52]	80.3	76.1	77.9	34.6
RobustFeature[53]	80.1	76.0	78.2	32.0
JCF(ours)	**83.8**	**76.2**	**81.1**	**34.7**

Table 5. Comparison of downsampling modules.

To demonstrate the sensitivity to lung nodules, various feature fusion modules were applied to the lung nodule detection task, and the proposed DSDF module was compared with these modules without changing the other network structures. The experimental results are shown in Table 4.

Under the same architecture, the DSDF module achieved a 3.9% improvement in mAP@.5 and a 1.1% improvement in mAP@.5:.95 compared to the CCFM in the baseline module. Additionally, its Precision, Recall, and mAP were all higher than those of other modules. Although the mAP@.5:.95 indicator is not always the best, the gap is not significant. The comparative experimental results demonstrate that DSDF effectively enhances the network's detection accuracy for lung nodules, while also validating the rationality of its structural design.

To validate the effectiveness of the proposed JCF module in the lung nodule detection task, the performance of replacing the downsampling module of the baseline model with other downsampling modules was tested. The experimental results are shown in Table 5.

The comparison of experimental results reveals that JCF, by effectively utilizing contextual information, significantly outperforms the baseline model in terms of detection accuracy, particularly achieving a 3.8% improvement in mAP@.5. Additionally, compared to other downsampling modules, JCF also demonstrated excellent performance. These results confirm the effectiveness of the JCF module in the lung nodule detection task.

Statistical analysis

We randomly selected 100 different images from the LUNA16 dataset (50 for RT-DETR and 50 for LN-DETR) as the source for statistical analysis. We collected the precisions of RT-DETR and LN-DETR on their respective assigned 50 images. Subsequently, an independent samples t-test was conducted to determine whether there were significant differences between the two sets of precisions. The results are shown in Table 6. Two key metrics are presented: p and Cohen's d. The former indicates whether a significant difference exists, with a value less than 0.01 indicating significance; the latter represents the magnitude of the difference, where a larger value indicates a greater difference. As shown in the table, there is not only a significant difference between RT-DETR and the proposed LN-DETR algorithm, but the magnitude of the difference is also substantial. Additionally, the average accuracy of LN-DETR is higher than that of RT-DETR, demonstrating that LN-DETR outperforms the original algorithm in practical detecting.

Qualitative results

Figure 7 shows a comparison of the detection results on the LUNA16 dataset between multiple models and LN-DETR. Due to the small size of the targets, for easier observation, the images have been cropped. YOLOv9, as a general-purpose object detection model, exhibits insufficient adaptability in its anchor box design and feature

Model	Sample Capacity	mean	Standard deviation	Cohen's d	t	df	p
RT-DETR	50	0.69	0.11	0.922	4.611	87.713	p<0.01
LN-DETR(ours)	50	0.77	0.07				

Table 6. Independent samples t-test.

pyramid structure for small targets such as pulmonary nodules (3-30 mm in size). Pulmonary nodules in CT images are characterized by low contrast and diverse appearances, and the single-stage detection architecture of the YOLO series tends to suffer from missed detections due to inadequate granularity in feature extraction. While RT-DETR achieves real-time performance by compressing computational overhead, this comes at the cost of sacrificing the ability to perform refined multi-scale feature fusion. It can be observed that the baseline model RT-DETR and YOLOv9, encountered issues such as missed detections, false positives, and lower accuracy during the lung nodule detection process, as they were not specifically optimized for lung nodule detection. Furthermore, the results of LN-DETR demonstrate that the proposed DSDF and JCF modules optimize the model's feature fusion and downsampling architecture, significantly enhancing its detection sensitivity to pulmonary nodules. As a result, LN-DETR demonstrates excellent detection performance for lung nodules, outperforming other models in terms of both false positive rate and detection accuracy.

Figure 8 demonstrates the detection results of lung nodules by the proposed LN-DETR under varying levels of noise. In the experiments, different noise conditions were simulated by altering the standard deviation of Gaussian noise. The comparison reveals that the proposed algorithm is almost unaffected under noise conditions with a standard deviation(σ) of 25, with the accuracy loss fluctuating within a range of 0 to 4%. Under noise conditions with a σ of 50, the impact becomes more pronounced, with the maximum accuracy loss reaching 13% and the minimum being only 2%. The comparison indicates that while the proposed algorithm is somewhat affected by noise, it also exhibits a certain degree of resistance, demonstrating its strong feature recognition capability and robustness.

The top and bottom rows of images in Fig. 9 respectively illustrate the heatmap focus results of the RT-DETR and the LN-DETR proposed in this paper. As can be observed from the figure, RT-DETR exhibits insufficient attention to certain pulmonary features. In contrast, LN-DETR demonstrates a higher degree of focus on key target regions and possesses a superior capability for capturing the characteristics of minute nodules within the lungs.

Conclusion

To address the challenges of small target size and low detection accuracy in lung nodule detection tasks, this paper proposes a lung nodule detection model, LN-DETR, based on RT-DETR. The model improves the backbone network using the FasterBlock module, enhancing feature extraction capabilities while reducing floating-point operations and the number of parameters. A deep and shallow feature fusion module is added to the feature fusion network to improve the network's ability to aggregate cross-scale feature from both shallow and deep features. Additionally, the Contextual Joint Feature (JCF) module is employed to improve the detection performance for lung nodules by re-weighting the feature of target regions via local and global feature. Experimental results show that LN-DETR achieves significant improvements in detection performance, with precision, recall, mAP@.5, and mAP@.5:.90 increasing by 3.5%, 8.2%, 6.4%, and 2.9%, respectively, while slightly reducing the number of parameters and computation. Overall, LN-DETR demonstrates excellent detection performance.

Although the LN-DETR model proposed in this paper demonstrates good detection accuracy, there is still room for improvement. Second, LN-DETR performs poorly on the fps metric, indicating that the algorithm still has certain shortcomings in terms of real-time performance. Third, although LN-DETR has achieved some success in reducing the number of model parameters and computational complexity, the extent of optimization is relatively small and insufficient to bring about a qualitative improvement. Therefore, our next steps will focus on investigating detection accuracy, real-time performance, the number of parameters, and computational complexity. Additionally, more efforts will be made to lower the model's hardware requirements, enabling better deployment on hospital equipment.

While the proposed LN-DETR demonstrates significant improvements in accuracy and efficiency, it still has certain limitations. For example, different datasets may exhibit varying characteristics; thus, further testing on diverse datasets is necessary to validate the algorithm's robustness and adaptability in different clinical environments. Furthermore, this study only examined the performance of LN-DETR in pulmonary nodule detection. The performance of LN-DETR on other medical object detection tasks remains inconclusive. Therefore, further research is needed to expand the investigation of LN-DETR in the field of medical object detection. Additionally, despite efforts to reduce computational complexity, the RT-DETR-based model remains relatively resource-intensive in terms of both memory and processing time. Future work should prioritize further optimization of the network to achieve a more effective balance between performance and computational efficiency.

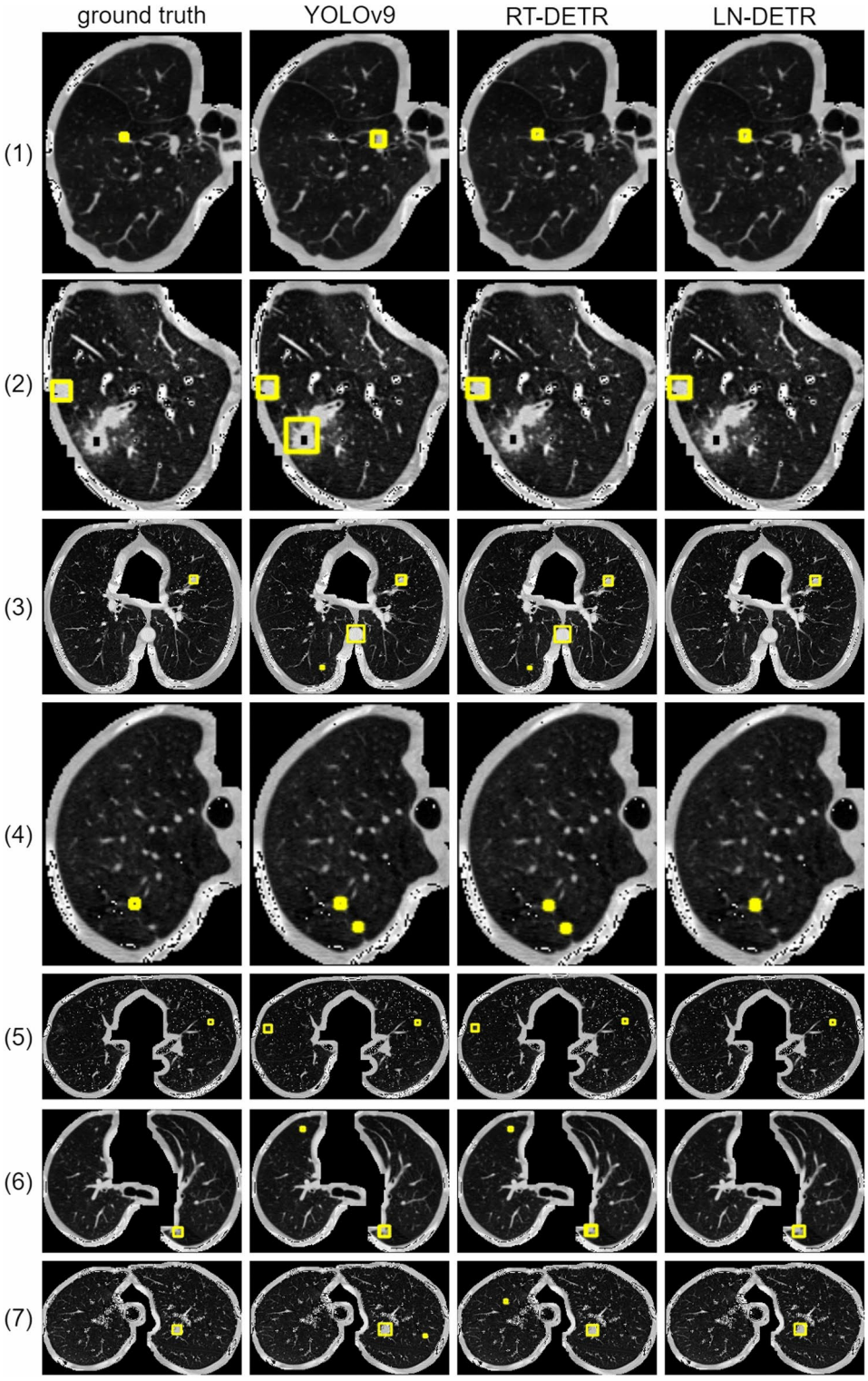


Figure 7. Visualization of experimental results of three methods. Labels represent the ground truth, LN-DETR represents our method.

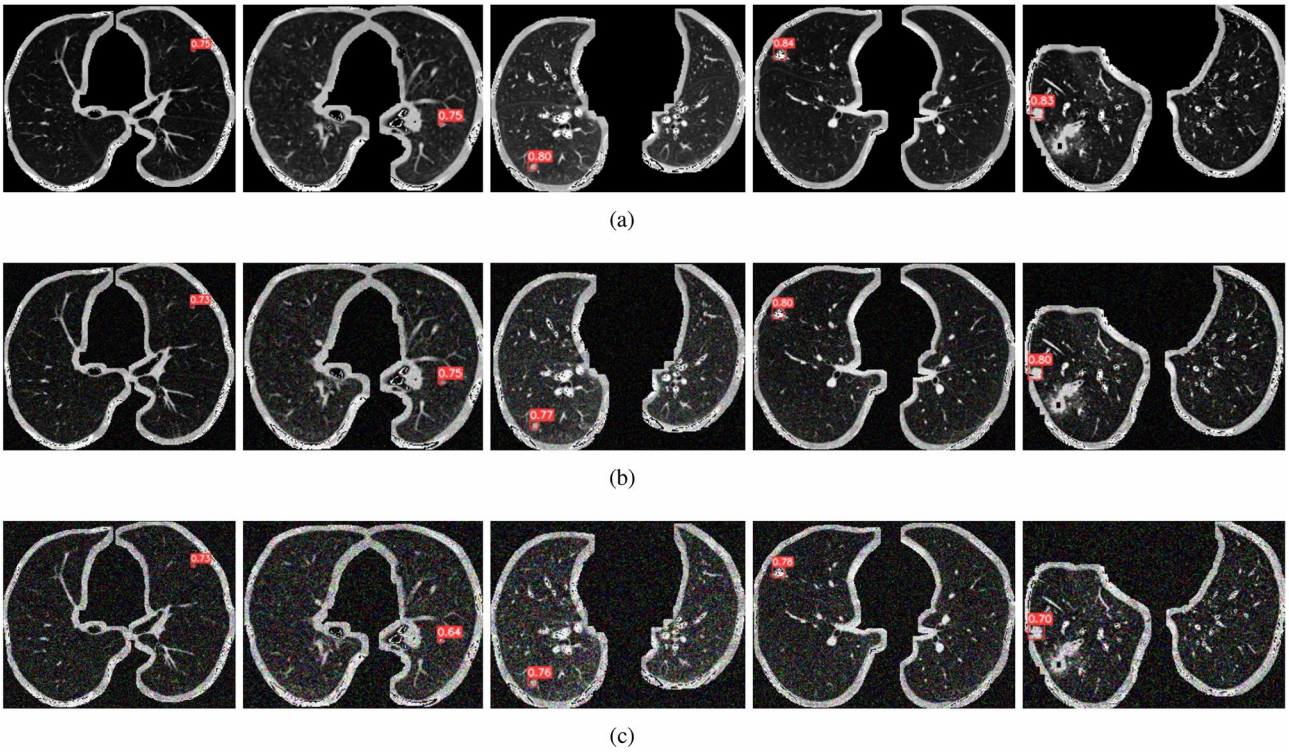


Figure 8. The response of LN-DETR to varying levels of noise. LN-DETR maintains a relatively stable accuracy across different noise levels.

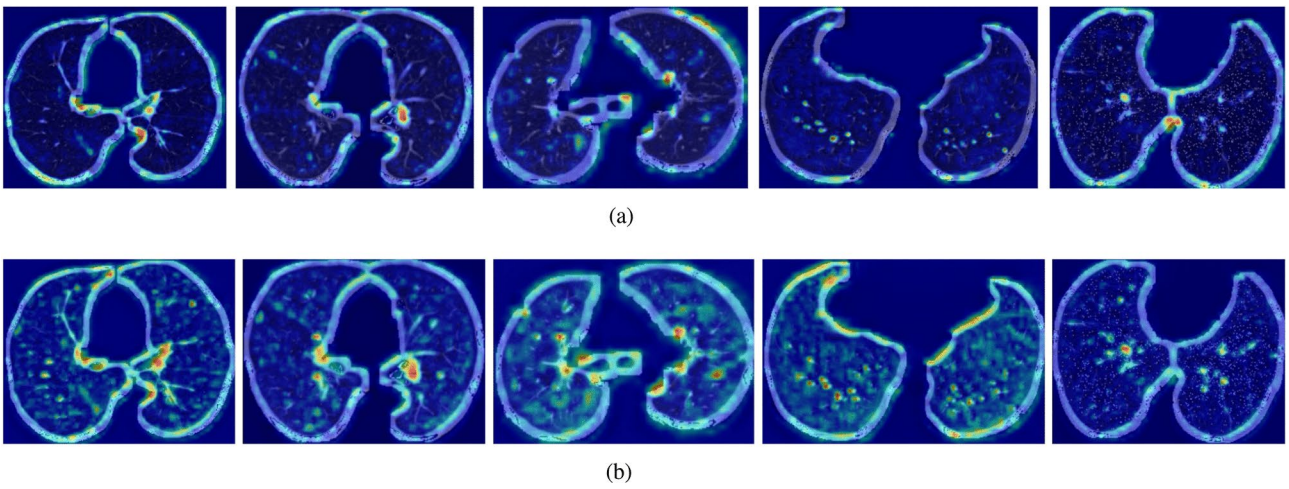


Figure 9. Grad-CAM visualization between RT-DETR and LN-DETR. From the heatmap, LN-DETR demonstrates a stronger activation of the target regions.

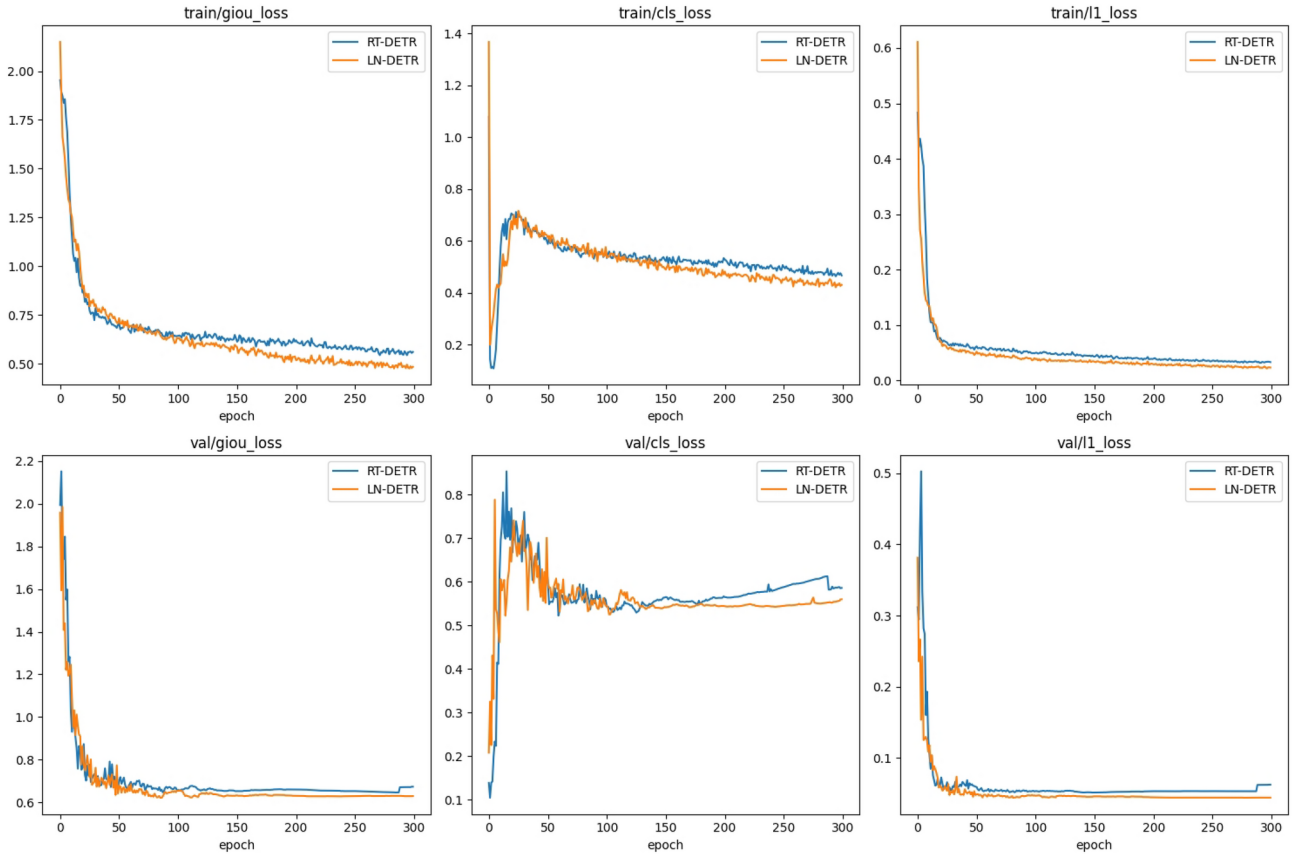


Figure 10. Comparison of train and validation curve variation.

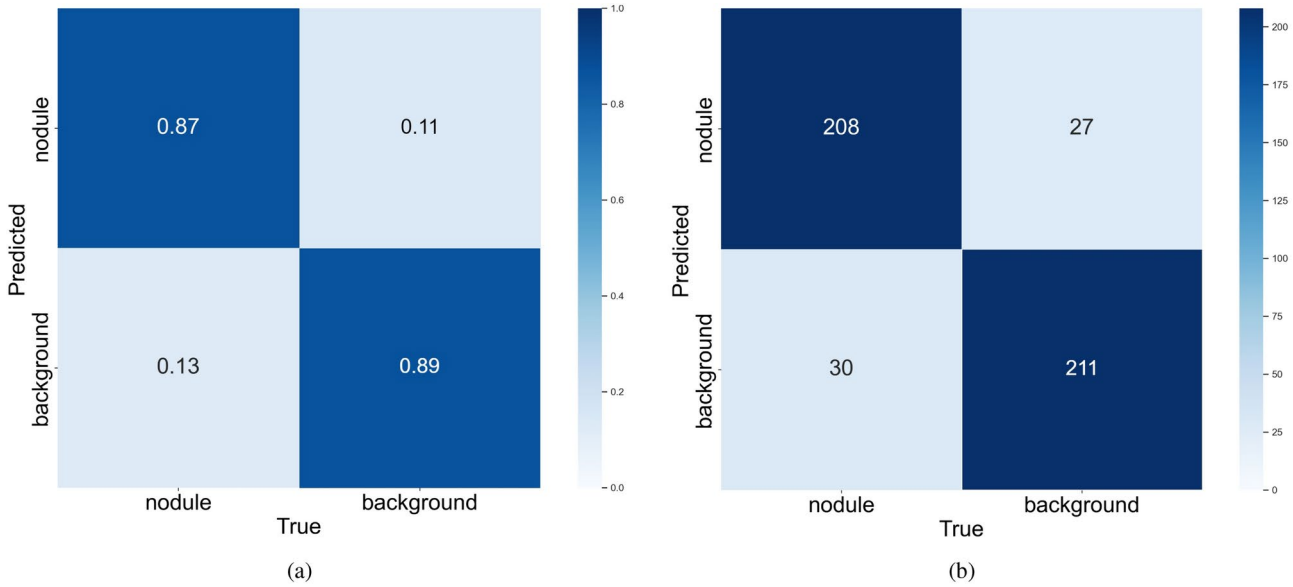


Figure 11. The confusion matrix of LN-DETR.

Data availability

We use the public dataset from LIDC-IDRI - The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net/collection/lidc-idri/) for algorithm evaluation.

Appendix

We utilized the RT-DETR and the proposed LN-DETR algorithms to plot Fig. 10 based on the data obtained from each training and validation epoch. This figure illustrates the comparative curves of three loss functions: GIoU Loss (Generalized Intersection over Union Loss), Cls Loss (Classification Loss) and L1 Loss (Mean Absolute Error Loss). These loss functions are used to optimize the size of bounding boxes, the position of bounding boxes, and the classification of targets, respectively. Lower loss values indicate closer alignment with the correct answers, reflecting better model performance. As shown in the figure, although the convergence speed of the proposed algorithm is similar to that of the original algorithm, its final results on both the training and validation sets outperform the original algorithm. This demonstrates that LN-DETR possesses excellent pulmonary nodule detection capabilities.

The confusion matrix, depicted in Fig. 11, offers a comprehensive analysis of the LN-DETR's classification accuracy. Figure 11 displays the normalized results of label classification and the specific numerical counts of label classifications, respectively. The data reveal that 87% of pulmonary nodules were accurately identified (true positives), whereas 13% of nodules were undetected and erroneously classified as background (false negatives). Additionally, Fig. 11b highlights 27 instances of background regions being incorrectly identified as pulmonary nodules (false positives). These results indicate that the model demonstrates high precision and recall, yet still exhibits non-negligible rates of missed detections and false alarms.

Received: 27 December 2024; Accepted: 28 April 2025
Published online: 03 May 2025

References

1. Feng, R.-M., Zong, Y.-N., Cao, S.-M. & Xu, R.-H. Current cancer situation in china: good or bad news from the 2018 global cancer statistics?. *Cancer Commun.* **39**, 1–12 (2019).
2. Howlader, N. et al. The effect of advances in lung-cancer treatment on population mortality. *N. Engl. J. Med.* **383**, 640–649 (2020).
3. Siegel, R. L. et al. Colorectal cancer statistics, 2017. *CA Cancer J. Clin.* **67**, 177–193 (2017).
4. Henschke, C. I. et al. Early lung cancer action project: overall design and findings from baseline screening. *The Lancet* **354**, 99–105 (1999).
5. Jing, J. et al. Training low dose ct denoising network without high quality reference data. *Phys. Med. Biol.* **67**. https://doi.org/10.1088/1361-6560/ac5f70 (2022).
6. Oudkerk, M., Liu, S., Heuvelmans, M. A., Walter, J. E. & Field, J. K. Lung cancer ldct screening and mortality reduction-evidence, pitfalls and future perspectives. *Nat. Rev. Clin. Oncol.* **18**, 135–151 (2021).
7. Tong, J., Da-Zhe, Z., Ying, W., Xin-Hua, Z. & Xu, W. Computer-aided lung nodule detection based on ct images. In *2007 IEEE/ICME international conference on complex medical engineering*, 816–819 (IEEE, 2007).
8. Ge, Y. et al. Camouflaged object detection via location-awareness and feature fusion. *Image Vis. Comput.* 105339 (2024).
9. Haibo, L., Shanli, T., Shuang, S. & Haoran, L. An improved yolov3 algorithm for pulmonary nodule detection. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, vol. 4, 1068–1072 (IEEE, 2021).
10. Wang, Y. et al. Hybrid u-net-based deep learning model for volume segmentation of lung nodules in ct images. *Med. Phys.* **49**, 7287–7302 (2022).
11. Gu, Y. et al. Automatic lung nodule detection using a 3d deep convolutional neural network combined with a multi-scale prediction strategy in chest cts. *Comput. Biol. Med.* **103**, 220–231 (2018).
12. Song, H., Yuan, Y., Ouyang, Z., Yang, Y. & Xiang, H. Quantitative regularization in robust vision transformer for remote sensing image classification. *Photogram. Rec.* **39**, 340–372 (2024).
13. Tanwar, V., Sharma, B., Yadav, D. P. & Dwivedi, A. D. Enhancing blood cell diagnosis using hybrid residual and dual block transformer network. *Bioengineering* **12**, 98 (2025).
14. Song, H., Yuan, Y., Ouyang, Z., Yang, Y. & Xiang, H. Efficient knowledge distillation for hybrid models: A vision transformer-convolutional neural network to convolutional neural network approach for classifying remote sensing images. *IET Cyber-Syst. Robot.* **6**, e12120 (2024).
15. Song, H. et al. Qaga-net: enhanced vision transformer-based object detection for remote sensing images. *Int. J. Intell. Comput. Cybern.* (2024).
16. Dai, X. et al. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2988–2997 (2021).
17. Armato, S. G. III. et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Med. Phys.* **38**, 915–931 (2011).
18. Kumar, M. K. & Amalanathan, A. Advancements in optimization algorithms for lung nodule detection and classification: A review. In *2023 1st International Conference on Optimization Techniques for Learning (ICOTL)*, 1–6 (IEEE, 2023).
19. Yadav, D. P., Sharma, B., Webber, J. L., Mehbodniya, A. & Chauhan, S. Edtnet: A spatial aware attention-based transformer for the pulmonary nodule segmentation. *PLoS ONE* **19**, e0311080 (2024).
20. Kumar, M. V. et al. Detection of lung nodules using convolution neural network: a review. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 590–594 (IEEE, 2020).
21. Liu. *Research and implementation of lung nodule auxiliary detection method based on YOLOv4*. Master's thesis, Chongqing: Chongqing University of Posts and Telecommunications (2021).
22. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020).
23. Ruan, B.-K., Shuai, H.-H. & Cheng, W.-H. Vision transformers: state of the art and research challenges. arXiv preprint arXiv:2207.03041 (2022).
24. Han, K. et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 87–110 (2022).
25. Thisanke, H. et al. Semantic segmentation using vision transformers: A survey. *Eng. Appl. Artif. Intell.* **126**, 106669 (2023).
26. Fauzya, S. P., Ardiyanto, I. & Nugroho, H. A. A comparative study on lung nodule detection: 3d cnn vs vision transformer. In *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 417–422 (IEEE, 2024).
27. Zhao, Y. et al. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974 (2024).
28. Chen, J. et al. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12021–12031 (2023).

29. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25** (2012).
30. Ma, N., Zhang, X., Zheng, H.-T. & Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131 (2018).
31. Tang, L., Zhang, H., Xu, H. & Ma, J. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf. Fusion* **99**, 101870 (2023).
32. Wu, T., Tang, S., Zhang, R., Cao, J. & Cgnet, Y. Z. A light-weight context guided network for semantic segmentation., 2020, 30. https://doi.org/10.1109/TIP 1169–1179 (2020).
33. Orlando, N. et al. Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3d ultrasound. *Phys. Med. Biol.* **67**. https://doi.org/10.1088/1361-6560/ac5a93 (2022).
34. Gong, W. Lightweight object detection: A study based on yolov7 integrated with shufflenetv2 and vision transformer. arXiv preprint arXiv:2403.01736 (2024).
35. Wang, C.-Y., Yeh, I.-H. & Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*, 1–21 (Springer, 2025).
36. Li, Y. et al. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16794–16805 (2023).
37. Cai, X. et al. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27706–27716 (2024).
38. Chen, Z., He, Z. & Lu, Z.-M. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Trans. Image Process.* (2024).
39. Liu, Z. et al. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
40. Ding, M. et al. Davit: Dual attention vision transformers. In *European conference on computer vision*, 74–92 (Springer, 2022).
41. Chen, C.-F. R., Fan, Q. & Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 357–366 (2021).
42. Fan, Q., Huang, H., Chen, M., Liu, H. & He, R. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5641–5651 (2024).
43. Shi, D. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17773–17783 (2024).
44. Chen, Y. et al. Accurate leukocyte detection based on deformable-detr and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* **170**, 107917. https://doi.org/10.1016/j.compbiomed.2024.107917 (2024).
45. Ma, X., Dai, X., Bai, Y., Wang, Y. & Fu, Y. Rewrite the stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5694–5703 (2024).
46. Wan, C. et al. Swift parameter-free attention network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6246–6256 (2024).
47. Kang, M., Ting, C.-M., Ting, F. & Phan, R. Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **147**, 105057. https://doi.org/10.1016/j.imavis.2024.105057 (2024).
48. Hu, S., Gao, F., Zhou, X., Dong, J. & Du, Q. Hybrid convolutional and attention network for hyperspectral image denoising. *IEEE Geosci. Remote Sens. Lett.* (2024).
49. Kaige, L., Geng, Q., Wan, M., Cao, X. & Zhou, Z. Context and spatial feature calibration for real-time semantic segmentation. *IEEE Trans. Image Process.* (2023).
50. Li, H. et al. Slim-neck by gsconv: a lightweight-design for real-time detector architectures. *J. Real-Time Image Proc.* **21**, 62 (2024).
51. Xu, G. et al. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recogn.* **143**, 109819 (2023).
52. Williams, T. & Li, R. Wavelet pooling for convolutional neural networks. In *International conference on learning representations*, 0 (2018).
53. Lu, W., Chen, S.-B., Tang, J., Ding, C. H. & Luo, B. A robust feature downsampling module for remote-sensing visual tasks. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–12 (2023).

Author contributions

Dibin Zhou conceived the experiment(s). Honggang Xu conducted the experiment(s). Wenhao Liu and Fuchang Liu analysed the results. Fuchang Liu and Dibin Zhou wrote the main manuscirpt text. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025