

Dual-stage pulmonary nodule detection in CT scans via cross-layer attention and adaptive multi-scale 3D CNN

Lixin Wang¹ , Xiaowen Lan¹, Kaikai Zhang¹, Yanhui Wang², Shaofeng Wang³ and Wenjing Liu³

Abstract

Background: Early diagnosis of pulmonary nodules is crucial for improving the survival rate of lung cancer patients. However, significant variability in nodule size, shape, and anatomical location presents ongoing challenges for automated detection systems, often resulting in high false-positive rates.

Objective: This study aims to develop a dual-stage pulmonary nodule detection framework based on cross-layer attention fusion, with the goal of improving sensitivity while reducing false positives in chest CT scans.

Methods: We propose a two-stage detection pipeline. In the candidate detection stage, we design an Attention-guided Spatial and Channel Residual Module that integrates multi-scale residual connections with cross-dimensional attention to enhance discriminative features while preserving spatial detail. For false positive reduction, we introduce a Multi-scale Progressive Perception Network, which processes candidates across three anatomical resolutions through parallel branches and integrates top-down semantic fusion with localized attention. The model is evaluated on the LUNA16 dataset.

Results: Experimental results demonstrate that the proposed method achieves a sensitivity of 90.0% at 0.55 false positives per scan on the LUNA16 dataset. Compared to state-of-the-art approaches, our framework provides a favorable balance between sensitivity and precision.

Conclusions: The proposed dual-stage detection framework effectively enhances the performance of pulmonary nodule detection by incorporating cross-layer attention mechanisms and multi-scale feature integration. These findings suggest its potential for clinical deployment in computer-aided lung cancer screening.

Keywords

Attention mechanisms, computer-aided diagnosis, deep convolutional networks, false positive reduction, lung nodule detection

Received: 22 July 2025; accepted: 13 January 2026

Introduction

Lung cancer remains one of the most formidable threats to global health, ranking highest in both incidence and mortality among all cancer types.¹ Despite recent advances in therapeutic strategies, most patients are diagnosed at advanced stages due to the paucity of overt symptoms in early disease, resulting in limited treatment efficacy and poor prognosis. Clinical evidence suggests that overcoming this diagnostic bottleneck hinges on early detection, precise diagnosis, and timely intervention—factors that are decisive for improving five-year survival rates and long-term

outcomes.² Central to early diagnosis is the accurate

¹School of Digital and Intelligent Industry, Inner Mongolia University of Science and Technology, Baotou, China

²The First Affiliated Hospital of Baotou Medical College, Baotou, China

³School of Mechanical Engineering, Inner Mongolia University of Science and Technology, Baotou, China

Corresponding author:

Xiaowen Lan, School of Digital and Intelligent Industry, Inner Mongolia University of Science and Technology, Baotou, China.

Email: lanxiaowenning@163.com



interpretation of medical imaging, particularly the effective identification and localization of pulmonary nodules, which constitute the critical control point in arresting disease progression.³ Chest computed tomography (CT) stands as the most sensitive and widely adopted modality for nodule detection. However, conventional image reading relies heavily on the expertise and subjective judgment of radiologists, making it vulnerable to inter-observer variability and prone to missed or false diagnoses. The enormous volume and complex information content of CT datasets further exacerbate this challenge, as rapid yet accurate analysis of hundreds of slices places substantial cognitive demands on clinicians.⁴ Consequently, there is a pressing need for computer-aided diagnosis (CAD) systems to assist radiologists by first flagging potential nodules with high sensitivity, and then suppressing false positives to streamline decision-making and enhance diagnostic accuracy.⁵

With the rise of computer vision and artificial intelligence, deep learning-based CAD for automated pulmonary nodule detection has emerged as a major research focus. By training on large, annotated CT datasets, these systems learn to extract discriminative features associated with nodules, thereby achieving superior sensitivity and specificity compared to traditional, hand-crafted feature approaches.^{6–8} Numerous CAD solutions have been proposed to support radiological workflows; for example, Lu et al.⁹ and Gu et al.¹⁰ demonstrated conventional pipelines based on manually engineered features. Conventional methods often fall short in detecting pulmonary nodules with diverse shapes, sizes, textures, and anatomical locations, resulting in suboptimal detection performance. In contrast, deep learning approaches have demonstrated superior capability in capturing complex nodule characteristics compared to handcrafted features.¹¹ For instance, Setio et al.,¹² Jiang et al.,¹³ Zuo et al.,¹⁴ and Xie et al.¹⁵ proposed two-dimensional convolutional neural network (2D CNN)-based frameworks to extract nodule features. However, these 2D architectures inherently lack the ability to leverage the rich spatial context embedded in three-dimensional CT volumes. To address this limitation, Huang et al.,¹⁶ Wang et al.,¹⁷ Cao et al.,¹⁸ and Zhu et al.¹⁹ have introduced 3D CNN-based detection models, achieving improved accuracy and robustness in pulmonary nodule identification.

Despite these advances, high false positive rates persist due to the extreme variability in nodule size, morphology, and anatomical context. To mitigate this issue, we propose a two-stage detection framework driven by cross-layer attention fusion, designed to enhance detection precision while effectively suppressing false positives. In the first stage, multi-layer feature fusion extracts local morphological details of candidate nodules, augmented by a dual-dimensional channel-spatial attention mechanism that strengthens global spatial awareness and positional encoding for precise coarse localization. The second stage employs an adaptive residual 3D CNN architecture, which

captures hierarchical contextual features through multi-scale volumetric convolutions and further refines detection via an attention-guided fusion strategy to reduce false positives.

Extensive experiments on the LUNA16 dataset validate the effectiveness and competitiveness of our approach. The principal contributions of this work are as follows:

- We design a multi-level feature fusion anchor prediction architecture that integrates intermediate semantic representations with upsampled high-resolution features, achieving precise characterization of small and morphologically complex nodules.
- We introduce the Attention-guided Spatial and Channel Residual Module (ASCRM), which simultaneously captures multi-scale features via parallel spatial-attention and channel-recalibration units, preserving spatial integrity while amplifying responses to subtle lesions.
- We develop an adaptive false-positive suppression network that utilizes cross-scale feature fusion to integrate hierarchical context, resulting in a reduction of false positives while preserving sensitivity.

Our framework demonstrates a sensitivity of 90.0% at just 0.55 false positives per scan on LUNA16, indicating its strong potential for clinical deployment.

Methods

The proposed pulmonary nodule detection framework, shown in Figure 1, comprises two sequential stages: candidate generation and false-positive reduction. In the first stage, a U-shaped encoder-decoder network employs cascaded feature enhancement via the ASCRM module, fused with spatial coordinate priors, to preserve fine-grained CT details while leveraging multi-scale feature aggregation to boost candidate recall. To address the challenge of nodules adherent to the pleura or vasculature, as well as those exhibiting atypical morphology or diminutive size, we incorporate online hard-example mining during training to assign greater weight to these difficult cases—thereby enhancing sensitivity at the cost of an elevated false-positive rate. In the second stage, we introduce a three-branch progressive classifier in which each branch processes a specific scale of input; a cross-branch feature interaction mechanism enables bidirectional information flow and fusion across resolutions, enhancing the discrimination of true nodules from imaging artefacts.

To characterize the computational cost of the two-stage framework, Table 1 reports the training time, parameter count, average computation time, and floating-point operations (FLOPs) for each stage. The training time is measured per fold. Although the first-stage candidate detector has fewer parameters, GPU memory constraints require each CT

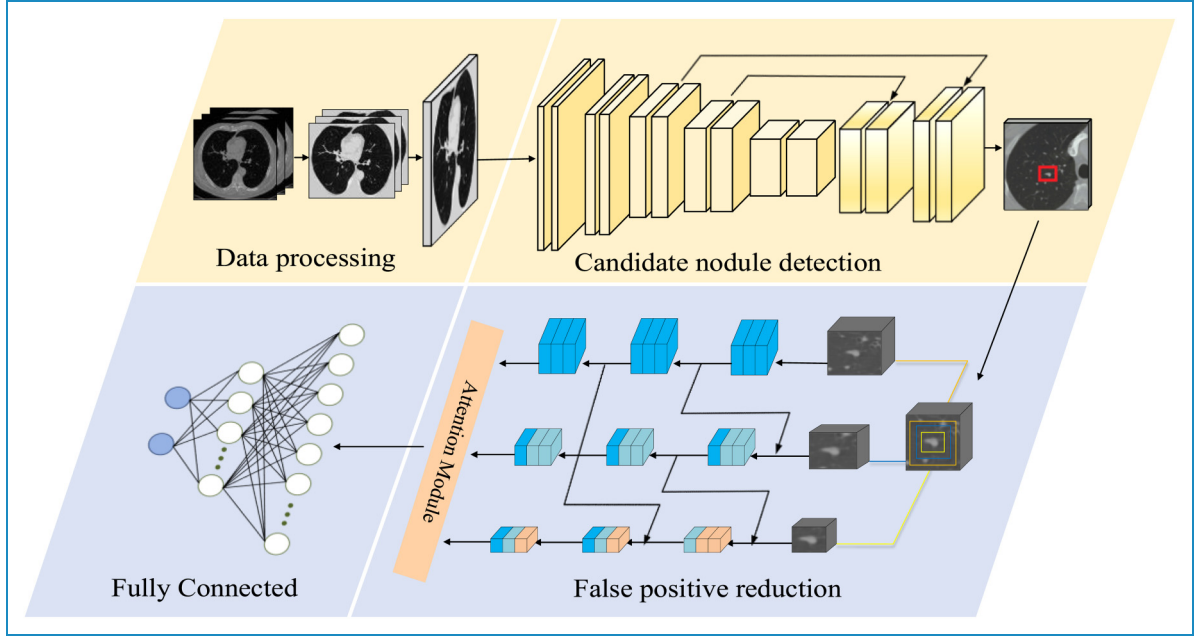


Figure 1. Workflow of the proposed pulmonary nodule detection framework.

scan to be processed in multiple $128 \times 128 \times 128$ sub-volumes, followed by block-level reconstruction during training and validation, leading to longer runtime. The larger 3D input blocks in this stage also result in substantially higher FLOPs than those of the second stage. In contrast, the second-stage false positive reduction network has more parameters but operates only on $32 \times 48 \times 48$ candidate nodule patches, which greatly reduces the computation scope and yields much lower FLOPs and faster training/inference.

Architecture of the cross-layer feature-fusion pulmonary nodule candidate detection network

In this section, we present a three-dimensional convolutional neural network (3D CNN) based on cross-layer feature fusion for the detection of pulmonary nodule candidates. The network takes a CT volume of size $128 \times 128 \times 128$ voxels as input and adopts an encoder-decoder architecture as its backbone (Figure 2). To effectively capture the diverse

morphological characteristics and complex spatial distribution of nodules, the model must possess strong feature representation capabilities, balancing global semantic understanding with fine-grained spatial detail preservation.

However, conventional convolutional modules exhibit limitations in modeling spatial and channel-wise attention. While they are proficient in extracting local spatial features, they lack explicit modeling of spatial positional information and offer limited interaction across channels. This hampers the network's ability to capture semantic dependencies and spatial regularities across scales—particularly problematic for nodules that are morphologically heterogeneous, diminutive, or ambiguously bounded. Furthermore, traditional convolutional operations are often deficient in integrating spatial and channel attention mechanisms, making it difficult to selectively enhance salient features, ultimately impairing detection sensitivity and accuracy.

To address these challenges, we introduce an ASCRM to enhance the network's representational and semantic modeling capacity. The encoder path incorporates five stacked ASCRM blocks as the core feature extraction units. Each ASCRM integrates spatial and channel attention to dynamically recalibrate feature responses, while employing multi-scale residual structures to facilitate cross-layer information exchange. This design enables the effective encoding of fine-grained semantic features critical to nodule detection. A detailed description of the ASCRM's architecture and fusion strategy is provided in the next section.

In parallel, the decoder path employs multi-level skip connections to fuse high-resolution structural cues (e.g. edges and textures) from shallow layers with upsampled deep features. This strategy mitigates spatial information loss during feature reconstruction, enhancing the delineation and fidelity of candidate nodule regions.

Table 1. Key performance metrics for two stages of pulmonary nodule detection.

Stage	Training time (per fold)	Params (M)	Average computation time per case	FLOPs (G)
1	16.7h	1.77	4.17 s (per CT scan)	193.39
2	7.5h	30.12	66.0 ms (per nodule patch)	0.954

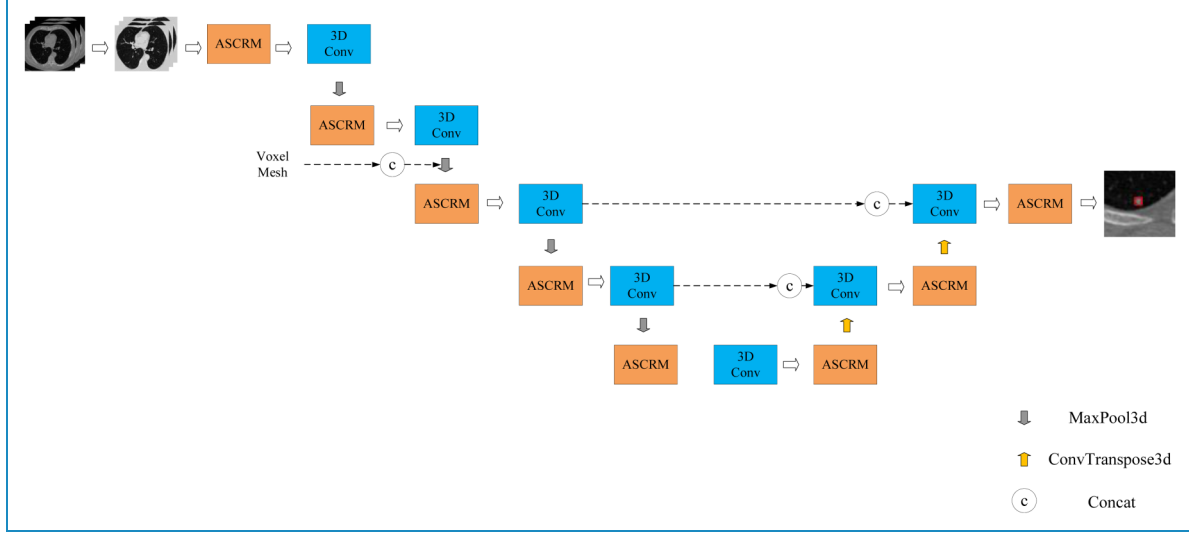


Figure 2. Architecture of the proposed candidate nodule detection network.

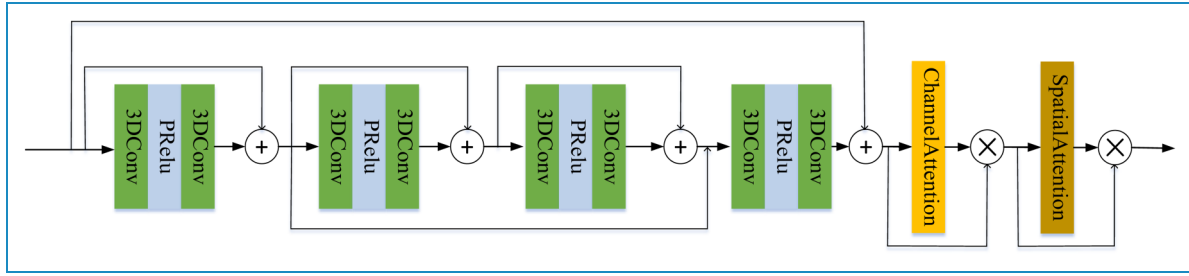


Figure 3. The proposed ASCRM module.

To further improve the network’s spatial awareness, we introduce voxel-wise spatial coordinate maps during the downsampling stage. These coordinate tensors are concatenated as additional input channels, providing spatial priors that guide the model in learning anatomical layouts of the lung fields. This approach suppresses false positives outside lung regions and improves detection accuracy and generalization—especially in cases involving blurred boundaries or structurally complex nodules.

Finally, the network outputs a volumetric probability map of size $32 \times 32 \times 32$, where each voxel denotes the likelihood of being the center of a pulmonary nodule. This map is subsequently used for candidate proposal generation and passed on to the false positive reduction stage.

Attention-guided spatial and channel residual module

To effectively enhance the network’s capability of modeling fine-grained features during candidate nodule detection, we introduce the ASCRM as the core feature extraction unit. The architectural details of ASCRM are illustrated in Figure 3. Building upon conventional residual

structures, ASCRM integrates dense skip connections and the Convolutional Block Attention Module (CBAM)²⁰ to strengthen semantic representation across multiple scales and improve the network’s focus on critical spatial regions.

Specifically, the ASCRM module constructs a deep residual path by stacking multiple $3 \times 3 \times 3$ convolution layers, while introducing long-range connections to progressively propagate high-resolution structural features from early shallow layers to deeper stages. This design effectively preserves fine-grained textures and boundary details of nodules. Within the residual path, each layer follows the formulation:

Let $F \in \mathbb{R}^{C \times D \times H \times W}$ denote the input 3D feature map to the ASCRM module, where C is the number of channels and D, H, W represent the depth, height, and width of the volumetric feature space, respectively.

$$F_k = F_{k-1} + \mathcal{R}_k(F_{k-1}) \quad (1)$$

where $\mathcal{R}(\cdot)$ denotes a residual transformation consisting of two successive $3 \times 3 \times 3$ 3D convolutional layers with an intermediate PReLU activation, F_k denotes the intermediate feature map obtained after the k -th residual update.

To further facilitate information exchange across different feature hierarchies, a cross-layer residual fusion mechanism is introduced into the ASCRM module. This mechanism integrates feature representations F_i and F_j originating from different depths of the network; F_i typically retains higher-resolution texture details, whereas F_j conveys richer semantic information. To achieve more effective feature integration, a nonlinear residual refinement is applied to the summed features, yielding an enhanced aggregated representation:

$$F_{\text{agg}} = F_i + F_j + R(F_i + F_j) \quad (2)$$

This design strengthens the expressive capacity of the fused features and ensures stable propagation of cross-layer information. By mitigating representational degradation that can occur during multi-scale feature integration, the residual fusion structure enhances the joint modeling of fine-grained structural cues and high-level semantics, thereby providing a more discriminative feature foundation for the subsequent attention modules.

To improve feature discriminability, the ASCRM module integrates the CBAM, constructing a dual attention pathway across spatial and channel dimensions. The channel attention branch utilizes both average and max pooling operations to capture global response features, generating a channel-wise attention map. The spatial attention branch aggregates pooled descriptors to produce a spatial response map that enhances critical regions. These attention mechanisms are defined as:

$$A_{\text{spatial}} = \sigma(\text{Conv3D}([\text{AvgPool}_c(F), \text{MaxPool}_c(F)])) \quad (3)$$

$$A_{\text{channel}} = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (4)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. $\text{AvgPool}_c(\cdot)$ and $\text{MaxPool}_c(\cdot)$ denote pooling operations performed along the channel dimension, yielding two single-channel spatial descriptors that summarize the average and salient responses at each voxel location. The concatenated descriptors are mapped by a 3D convolution to capture local spatial dependencies, and the resulting A_{spatial} provides location-wise weights to emphasize informative regions while suppressing irrelevant background responses. $\text{AvgPool}(\cdot)$ and $\text{MaxPool}(\cdot)$ denote global pooling operations that aggregate each channel into a compact descriptor. $\text{MLP}(\cdot)$ denotes a multilayer perceptron, implemented as a lightweight learnable gating function shared by the two pooled descriptors to generate channel-wise weights. The resulting A_{channel} highlights discriminative feature channels and attenuates less informative ones. Through the joint use of A_{spatial} and A_{channel} , the module strengthens feature responses that are most relevant to pulmonary nodule candidates from both spatial and channel perspectives. The final output of the ASCRM module is computed as:

$$F_{\text{out}} = F + A_{\text{spatial}} \odot (A_{\text{channel}} \odot (F_{\text{agg}} + F)) \quad (5)$$

Where $\odot$ is the element-wise multiplication. This formulation facilitates the fusion of shallow and deep representations

and enhances feature responses along spatial and channel dimensions. When applied in a stacked manner within the network, ASCRM supports stable pulmonary nodule candidate detection, including cases with complex morphology.

Multi-scale progressive inference architecture

In the task of false positive reduction, the heterogeneous nature of pulmonary nodules—manifested in their diverse morphologies, variable densities, and particularly wide size range (Figure 4)—poses a significant challenge. Conventional fixed-scale detection strategies often struggle to concurrently capture the global structural patterns of large nodules and the subtle textural cues of smaller ones, leading to frequent misses of tiny nodules or misclassification of non-nodular tissues. To address these limitations, we propose a Multi-scale Progressive Perception Network (MPPN) that integrates a hierarchical context fusion scheme with attention-enhanced residual modules to enable precise classification of candidate nodules across scales (Figure 5).

As illustrated in Figure 5, MPPN stratifies candidates into three-scale categories based on nodule diameter, and accordingly extracts 3D voxel patches of size $32 \times 48 \times 48$, $16 \times 24 \times 24$, and $8 \times 12 \times 12$, forming an adaptive multi-scale input representation. This design avoids spatial distortions and semantic degradation typically introduced by uniform resampling, and improves the network's ability to model the geometry and context of nodules at different scales. Each scale-specific input is processed by an independent 3D convolutional branch, constructing a coarse-to-fine multi-scale feature extraction pathway. To alleviate semantic isolation between scales, a top-down feature injection mechanism is introduced in the medium and small-scale branches, allowing high-level semantics from larger scales to guide the detection of small nodules. Additionally, each branch incorporates a CBAM to enhance the model's focus on discriminative regions. The high-level semantic features from all branches are concatenated along the channel dimension, followed by global average pooling to obtain a compact global representation. This representation is then recalibrated via attention and passed through a fully connected layer to output the probability of each candidate being a true nodule.

Experiment

Dataset and preprocessing

This study employs the LUNA16²¹ benchmark dataset for the lung nodule detection task. The dataset is derived from the LIDC-IDRI database, a multi-center collaborative project initiated by the National Cancer Institute (NCI) and supported by the Foundation for the National Institutes of Health (FNIH), which includes chest CT scans annotated by multiple radiologists.²²

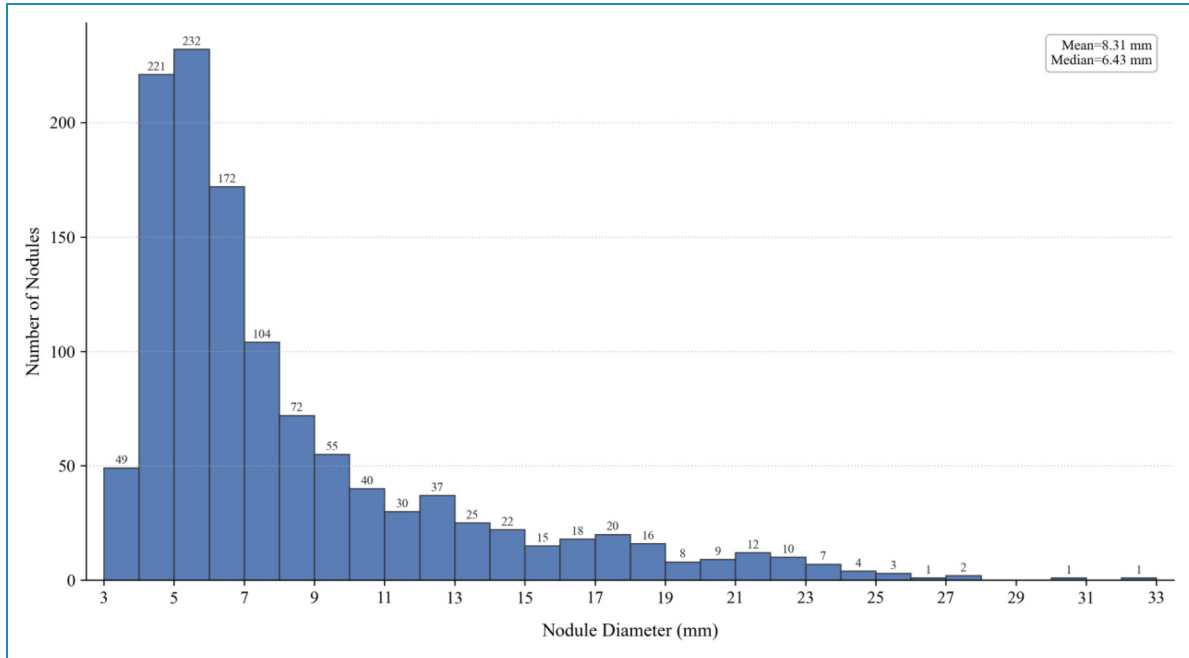


Figure 4. Distribution of lung nodule sizes in the LUNA16 dataset.

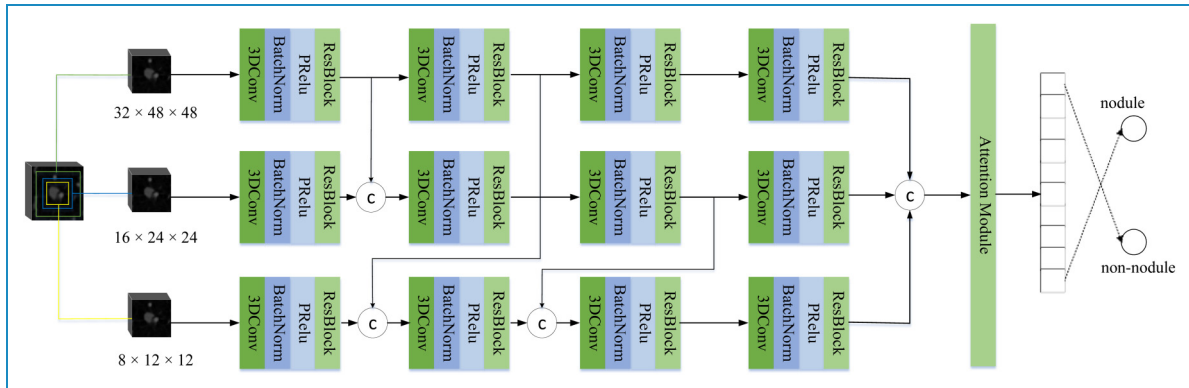


Figure 5. The proposed false positive reduction model.

The LUNA16 dataset was rigorously selected from the LIDC-IDRI database, with inclusion criteria that primarily exclude cases with missing slices, inconsistent pixel spacing, or slice thicknesses greater than 3 mm. The final dataset consists of 888 high-quality CT scans, all provided in MHD/RAW format. In terms of nodule annotation, the LUNA16 dataset follows a strict expert consensus standard, including only nodules with a diameter of ≥ 3 mm, which were confirmed by at least three of the four radiologists. Based on this criterion, the dataset contains 1186 lung nodule samples, each accompanied by detailed center coordinates and scaling information. Additionally, segmentation images of the lung regions for all CT scans are included.

To enhance the model's generalizability and robustness, a binary segmentation mask was first applied to precisely extract the lung parenchyma region from each input image (Figure 6). Each CT slice underwent preliminary denoising via a Gaussian filter to mitigate errors caused by blurred tissue boundaries. Subsequently, a fixed intensity threshold of -400 HU was employed to binarize the image, enabling initial separation of lung tissue. Considering potential discontinuities and small holes at the lung parenchyma edges, morphological closing operations were further applied to repair the binary images and fill residual small cavities. After obtaining closed lung parenchyma candidate regions, connected component analysis was used to retain the two largest connected areas, thereby excluding non-pulmonary

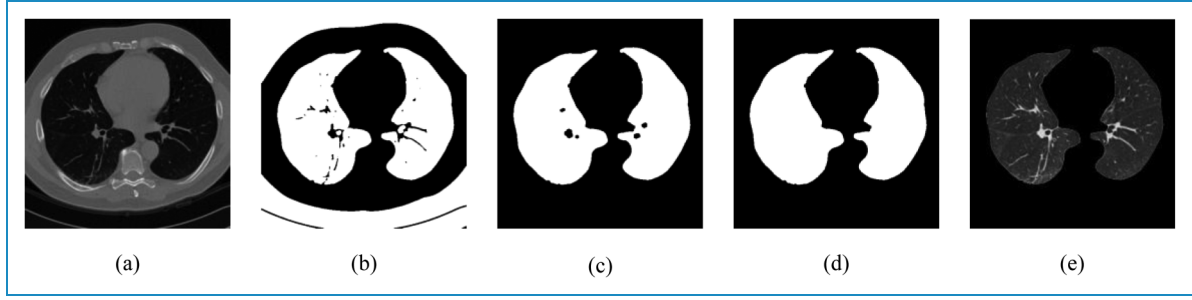


Figure 6. (a) Original image; (b) binary lung region mask obtained after thresholding; (c) initial extracted lung parenchyma mask; (d) lung parenchyma mask after morphological closing operation; (e) final segmented lung parenchyma.

Table 2. Results of six-fold cross-validation on the LUNA16 dataset.

Subset	0	1	2	3	4	5	Avg
Sensitivity	95.5	94.4	94.3	96.1	94.1	95.0	94.9
Average number of candidates /scan	3.1	3.6	5.9	4.0	4.1	4.4	4.2
CPM	86.9	87.6	86.4	91.4	87.6	87.2	87.8
FPs/Scan @ 90% Sensitivity	0.44	0.55	0.75	0.29	0.47	0.79	0.55

Table 3. Performance metrics for the false positive reduction network in the second stage.

Subest	0	1	2	3	4	5	Avg
Recall	0.966	0.957	0.944	0.936	0.924	0.951	0.949
Auc	0.974	0.988	0.971	0.983	0.971	0.992	0.980
F1-score	0.963	0.961	0.951	0.951	0.957	0.971	0.959
Precision	0.960	0.965	0.959	0.973	0.992	0.993	0.975

structures such as the trachea or external noise. Remaining internal holes were filled using a morphological binary fill operation, yielding complete masks for both left and right lungs.

To enhance contrast of pulmonary structures, CT values were linearly mapped within the lung window range of -1200 to 600 HU. Moreover, to address variations in spatial resolution arising from different manufacturers and scanner models, B-spline interpolation resampling was performed to achieve isotropic voxel spacing of 1 mm^3 , ensuring spatial consistency and comparability across all images.

Training process

The training process is divided into two stages: candidate nodule detection and false positive reduction. During the

candidate detection stage, considering the large volume of the entire CT scans, directly inputting the full volume into the network would cause GPU memory bottlenecks. Therefore, the images are partitioned into three-dimensional voxel blocks of size $128 \times 128 \times 128$ to alleviate computational resource constraints. To enhance model robustness and generalization, data augmentation techniques such as random flipping and scaling transformations are applied during training. For generating training anchors, boxes with an Intersection over Union (IoU) greater than 0.5 are labeled as positive samples, while those with an IoU less than 0.02 are labeled as negative samples. Three anchor sizes of 5 , 10 , and 20 voxels are employed. The proposed loss function comprises two components: classification loss and regression loss, optimized jointly via a dynamic weighting mechanism. For classification, an improved

Table 4. Validation results of the first-stage candidate detection network on the LIDC-IDRI dataset.

Metric	LIDC-IDRI (6-model ensemble)
Sensitivity	92.8
Average number of candidates /scan	4.7
CPM	85.2
FPS/Scan @ 90% Sensitivity	0.63

binary focal loss (Binary Focal Loss) is adopted, whose core formulation is defined as follows:

$$\mathcal{L}_{cls} = -\frac{1}{N_{pos} + N_{neg}} \sum_i \alpha(1 - p_i)^\gamma \log(p_i) \quad (6)$$

Here, N_{pos} and N_{neg} denote the numbers of positive and negative samples, respectively. For each sample i , $\hat{y}_i$ denotes the predicted logit and $\sigma(\cdot)$ is the Sigmoid function; p_i is defined as the predicted probability assigned to the ground-truth class, that is, $p_i = \sigma(\hat{y}_i)$ for positive samples and $p_i = 1 - \sigma(\hat{y}_i)$ for negative samples. The parameter α controls the overall scaling of the classification term, while γ down-weights well-classified samples to mitigate the impact of class imbalance. In our experiments, $\alpha=0.6$ and $\gamma=2$.

For the regression task, the Smooth L1 Loss is employed, defined as:

$$\mathcal{L}_{reg} = \sum_{k=1}^4 \begin{cases} 0.5(\hat{x}_k - x_k)^2, & \text{if } |\hat{x}_k - x_k| < 1 \\ |\hat{x}_k - x_k| - 0.5, & \text{otherwise} \end{cases} \quad (7)$$

Here, x_k and $\hat{x}_k$ denote the ground-truth and predicted box parameters, respectively. The regression loss is computed over positive samples, and the four terms correspond to the four box parameters.

The total loss function balances the two objectives—classification and regression—via an adaptive weighting mechanism, and is expressed as:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{cls} + \mathcal{L}_{reg}, \lambda = 5.0 \sqrt{\frac{N_{pos}}{N_{neg}}} \quad (8)$$

where the dynamic coefficient λ is adaptively adjusted based on the distribution of training samples to regulate the contribution of the classification loss. This design helps prevent the sparsity of regression samples from dominating the learning process, thereby enhancing classification accuracy while maintaining stable localization performance.

In the false positive reduction stage, the model is trained using image patches of three different sizes: $32 \times 48 \times 48$, $16 \times 24 \times 24$, and $8 \times 12 \times 12$ voxels. To address the

Table 5. Validation results of the second-stage false positive reduction network on the LIDC-IDRI dataset.

Metric	LIDC-IDRI (6-model ensemble)
Recall	0.932
Auc	0.968
F1-score	0.937
Precision	0.942

imbalance between true positive and false positive nodules, data augmentation techniques such as random flipping, translation, and rotation are applied. Additionally, the ratio of true positive to false positive samples is expanded to 1:3 to ensure a balanced training distribution. The model is optimized using the cross-entropy loss function, with an initial learning rate of 0.0001 and the Adam optimizer. All experiments are implemented using the PyTorch framework and trained on an Ubuntu server equipped with an NVIDIA GeForce RTX 3090 GPU.

Experimental results

To systematically evaluate the stability and generalizability of the proposed model under varying data splits, a six-fold cross-validation strategy was employed. The performance metrics across the six subsets are summarized in Table 2. On average, the model achieved a sensitivity of 94.9%, a Competition Performance Metric (CPM) score of 87.8, and generated only 4.2 candidate detections per scan. Notably, when the sensitivity was maintained at 90.0%, the average false positive rate was controlled within 0.55 per scan.

Building on the analysis of the system's overall performance, we further evaluated the independent classification capability of the second-stage false positive reduction network. As shown in Table 3, this network achieved an average AUC of 0.980, Precision of 0.975, and F1-score of 0.959 on the test set, while maintaining a recall of 0.949. These results indicate that the false positive reduction network demonstrates robust discriminative power in distinguishing true nodules from imaging artifacts, and its stable classification performance provides a reliable foundation for the system's low false positive rate.

To evaluate the generalizability of the proposed framework across different data sources, we conducted additional validation experiments on the LIDC-IDRI dataset. The six models obtained from the six-fold cross-validation on LUNA16 were applied to LIDC-IDRI for inference. Since LUNA16 is constructed from a subset of LIDC-IDRI, overlapping cases may introduce bias. Before running inference on LIDC-IDRI, we removed the CT scans that had been

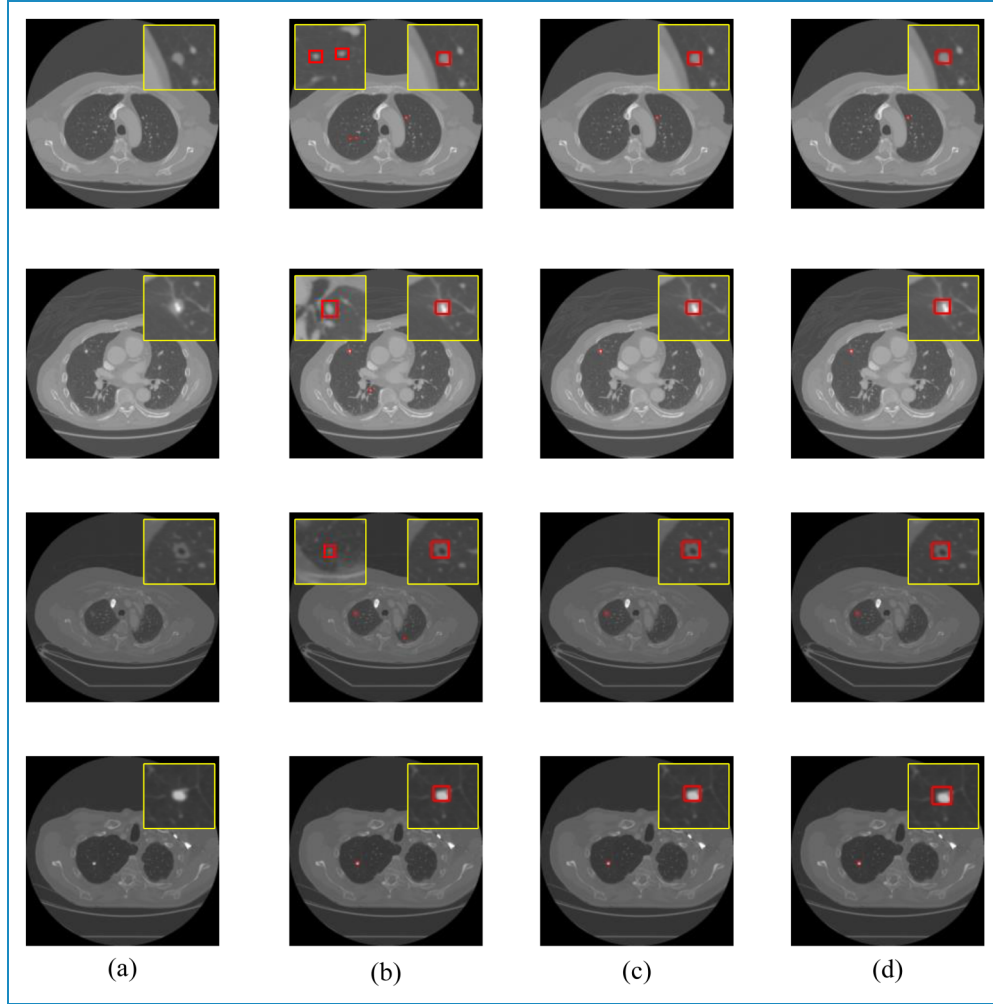


Figure 7. Representative intermediate results of the proposed framework. Column (a) shows the original chest CT images; column (b) displays the detected candidate nodules from the first-stage detection model; column (c) presents the refined outputs after false positive reduction; and column (d) illustrates the corresponding ground-truth annotations.

used to train each model, ensuring case-level independence between the training data and the LIDC-IDRI evaluation set. Performance metrics were computed for each of the six models and then averaged to obtain the final results. The LIDC-IDRI results of the first-stage candidate detection network and the second-stage false positive reduction network are summarized in Tables 4 and 5, respectively.

As shown in Table 4, the first-stage candidate detector maintains a high sensitivity of 92.8% on LIDC-IDRI and generates 4.7 candidates per scan on average. When the operating point is adjusted to achieve 90% sensitivity, the false positives are controlled at 0.63 FPs/scan, indicating stable candidate generation under different operating thresholds. Table 5 shows that the second-stage false positive reduction network preserves strong discriminative capability on LIDC-IDRI, achieving an AUC of 0.968 with a balanced precision–recall profile (Precision = 0.942,

Recall = 0.932). These results support the effectiveness of the overall pipeline in suppressing false positives on LIDC-IDRI.

Figure 7 provides a visual comparison of intermediate detection results at different stages of the pipeline. Column (a) shows the original chest CT images, column (b) presents the candidate nodule detection outputs, column (c) displays the refined results after false positive suppression, and column (d) shows the corresponding clinical ground-truth annotations. The proposed two-stage detection framework adopts a cascaded architecture: the first-stage candidate detection network emphasizes sensitivity through rich feature extraction (Figure 7(b)), but inevitably introduces a considerable number of false positives; the second-stage false positive reduction model leverages deep feature discrimination to suppress misclassified nodules effectively (Figure 7(c)). Experimental results demonstrate that this

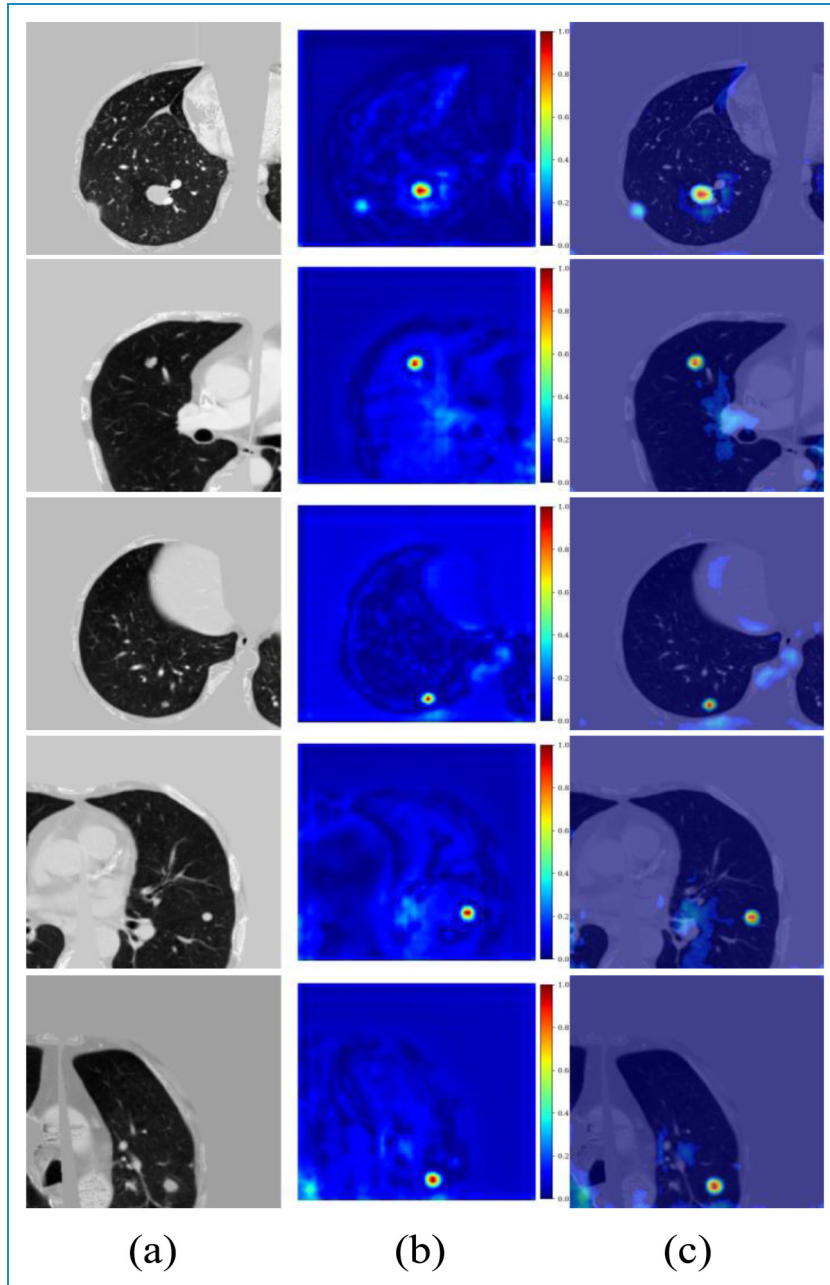


Figure 8. Grad-CAM visualization of first-stage candidate nodule detection. Column (a) shows the original chest CT images; column (b) displays the Grad-CAM heatmaps highlighting the regions of interest identified by the model; column (c) presents the overlay of the Grad-CAM heatmaps on the original images, showing the model's attention in context.

cascaded “detection-then-refinement” paradigm significantly enhances specificity while maintaining high sensitivity, thereby improving the overall robustness of pulmonary nodule detection.

Furthermore, to gain deeper insights into the decision-making mechanism of the first-stage candidate nodule detection model, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to visualize the spatial regions that contribute most to the model's predictions (Figure 8). The visualization results show that the model

produces strong and spatially concentrated activations in regions corresponding to pulmonary nodules. These activations are consistently localized around the nodule areas across different samples, indicating that the model relies primarily on nodule-relevant spatial features when generating detection responses. The observed activation patterns suggest that the learned feature representations are closely associated with the morphological and intensity characteristics of nodules in CT images. These CAM visualizations provide an interpretable view of the detection process, offering qualitative

Table 6. Performance comparison of candidate nodule detection methods.

Schemes	Year	Sensitivity	Average number of candidates /scan
Pereira et al. ²³	2021	98.2	49.67
Yuan et al. ²⁴	2021	94.0	50.25
Zhao et al. ²⁵	2023	94.7	14.10
Usman et al. ²⁶	2024	97.8	21.61
Xion et al. ²⁷	2024	97.4	9.67
Ours	2025	98.5	6.75

Table 7. Parameter comparison of candidate nodule detection methods.

Schemes	Year	Params (M)
Tang et al. ²⁸	2025	30.60
Almahasneh et al. ²⁹	2025	3.10
Huang et al. ³⁰	2025	6.64
Ours	2025	1.77

evidence of how the model focuses on nodule-related regions during candidate generation.

Comparison with other methods

In this study, we propose a two-stage pulmonary nodule detection framework that takes three-dimensional chest CT scans as input. The first stage employs a U-shaped network to generate a nodule probability map and identify initial candidate regions. Subsequently, a MPPN, specifically designed for this task, is introduced to further discriminate these candidates, effectively suppressing false positives and improving detection accuracy. By integrating coarse candidate generation and refined classification, the two-stage architecture achieves a notable enhancement in overall detection performance.

We conducted a comparative performance analysis of both the single-stage candidate detection model and the complete two-stage detection framework against several state-of-the-art methods. In the candidate detection stage, we benchmarked the sensitivity and the average number of candidates per scan (FPs/scan) of our method against those reported by Pereira et al. (2021),²³ Yuan et al. (2021),²⁴ Zhao et al. (2023),²⁵ Usman et al. (2024),²⁶ and

Xiong et al. (2024).²⁷ As shown in Table 6, our approach achieved a sensitivity of 98.5%, while maintaining the lowest average number of false positives per scan, demonstrating its ability to retain high sensitivity with superior false positive control. Beyond detection performance, model size is an important consideration for computational cost and practical deployment. Accordingly, Table 7 compares the parameter count of our candidate detection network with recent representative methods. Our first-stage detector contains 1.77 M parameters, which is smaller than the compared networks, indicating a relatively compact model design.

We further compare our method with several representative approaches in the false positive reduction phase, as summarized in Table 8. Our method demonstrates stable performance across operating points and achieves a CPM of 0.878, which is the highest among the compared methods. At operating points of 0.5 and 1 FPs/scan, our method attains sensitivities of 0.891 and 0.939, respectively, indicating strong detection sensitivity under moderate false-positive constraints. Overall, the results in Table 8 suggest that the proposed framework provides a balanced trade-off across operating points and yields favorable CPM performance.

Figure 9 illustrates the FROC (Free-response Receiver Operating Characteristic) curve of our method on the LUNA16 dataset, which is used to assess detection performance under different false positive rates. As shown in the figure, even under a strict constraint of no more than 0.55 FPs/scan, our method still achieves a detection sensitivity of 0.9, clearly demonstrating its strong nodule recognition capability while maintaining a low false positive rate.

The superior performance of our approach is mainly attributed to the integration of multi-scale feature fusion and attention mechanisms, which enable the model to more accurately distinguish between nodules and non-nodules in the complex anatomical background of the lung. The combined results in Figure 9 and Table 8 indicate that our method achieves a favorable balance between high sensitivity and low false positives, showing solid potential for practical application—particularly in clinical lung nodule screening scenarios where strict control over the false positive rate is required.

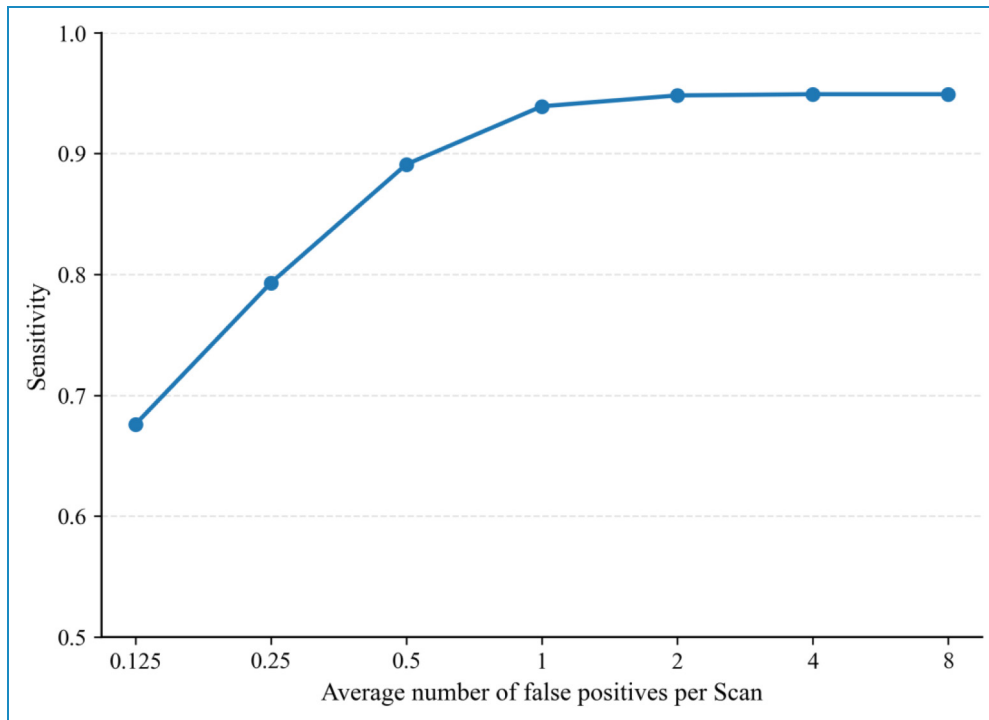
Discussions

The results of the comparative experiments demonstrate that the proposed two-stage detection framework performs competitively on the LUNA16 dataset, indicating that the framework design contributes to addressing the challenge of high false positive rates in pulmonary nodule detection. Traditional detection methods often face limitations when dealing with the wide variations in nodule size, shape, and anatomical location, particularly in cases where nodule morphology and anatomical background are complex. While methods based on 2D CNNs can effectively capture nodule features, they typically struggle to fully leverage the rich spatial context embedded in 3D CT data.

Table 8. Comparative performance evaluation in the false positive reduction phase.

Schemes	year	False positive per scan							CPM
		1/8	1/4	1/2	1	2	4	8	
Pezeshk et al. ³¹	2018	0.637	0.723	0.804	0.865	0.907	0.938	0.952	0.832
Huang et al. ³²	2019	0.817	0.851	0.869	0.883	0.891	0.907	0.914	0.876
Li et al. ³³	2020	0.739	0.803	0.858	0.888	0.907	0.916	0.920	0.862
Mei et al. ³⁴	2021	0.712	0.802	0.865	0.901	0.937	0.946	0.955	0.874
Zhao et al. ²⁵	2023	0.656	0.754	0.833	0.917	0.951	0.970	0.977	0.865
Tan et al. ³⁵	2023	0.728	0.799	0.860	0.901	0.926	0.941	0.956	0.873
Lu et al. ³⁶	2023	0.712	0.792	0.858	0.900	0.928	0.949	0.961	0.871
Almahasneh et al. ²⁹	2025	0.748	0.812	0.856	0.893	0.919	0.942	0.945	0.874
Ours	2025	0.676	0.793	0.891	0.939	0.948	0.949	0.949	0.878

Bold text indicates the false positive rate per scanning round.

**Figure 9.** Free-response receiver operating characteristic (FROC) curve.

The proposed cross-layer attention fusion two-stage framework combines attention mechanisms with multi-scale fusion, effectively reducing false positive rates while maintaining high sensitivity. The ASCRM module enhances feature representation capabilities, improving the model's accuracy in nodule detection, while the MPPN network effectively suppresses false positives

and enhances the fine discrimination of candidate nodules.

Ablation experiment

To validate the effectiveness of the ASCRM module and the false positive reduction network (MPPN), we conducted

Table 9. Results of four ablation experiments.

Model configuration	Sensitivity	Average number of candidates /scan	FPs/Scan @ 90% Sensitivity	P-value
Baseline model only	91.63 ± 1.35	3.81 ± 0.38	1.20 ± 0.3	0.0004
Baseline + CBAM attention mechanism	97.19 ± 0.92	14.26 ± 0.61	2.92 ± 1.19	0.0048
Baseline + ASCRM module	96.99 ± 1.8	10.48 ± 5.1	1.62 ± 0.55	0.0029
Baseline + ASCRM module + False Positive Reduction (MPPN)	94.9 ± 0.78	4.18 ± 0.95	0.55 ± 0.19	-

four ablation experiments: (i) baseline model only, (ii) baseline with CBAM-based attention mechanism, (iii) baseline with ASCRM module, and (iv) ASCRM module combined with the MPPN model. We validated the models on six subsets, and Table 9 provides the mean and standard deviation of the results from these six trials. The baseline model, without any enhancements, achieved a sensitivity of 91.63%. Incorporating the CBAM attention mechanism significantly improved sensitivity to 97.19%; however, it also dramatically increased the average number of detected candidates per scan from 3.81 to 14.26, and the false positive rate at 90.0% sensitivity rose from 1.20 to 2.92 per scan. This increase was primarily due to the detection of more true nodules at the cost of introducing excessive false positives under the same confidence threshold.

By integrating the ASCRM module, the model maintained high sensitivity while reducing the number of false positives compared to using CBAM alone. Furthermore, when the MPPN was added to the ASCRM-enhanced model, the false positive rate decreased to 4.18 per scan, with only a marginal drop in sensitivity. Overall, these results demonstrate that the ASCRM module and MPPN strategy effectively suppress false positives while preserving high detection sensitivity.

Additionally, we performed paired t-tests to calculate the significance of the differences between each model and the final two-stage model (Baseline + ASCRM + MPPN) for the FPs/Scan @ 90% Sensitivity metric. The results indicate that the p -values for all ablation models compared to the final two-stage model are less than 0.05, demonstrating statistically significant differences.

Conclusion

This paper proposes a two-stage pulmonary nodule detection method based on cross-layer attention fusion. In the candidate nodule detection stage, a multi-scale feature extraction and cross-layer attention fusion mechanism is introduced. The attention-enhanced module (ASCRM) effectively improves the model’s sensitivity to tiny nodules and enhances the recall rate of candidate regions. In the

false positive reduction stage, a MPPN is constructed to integrate semantic information across different spatial scales, thereby improving the model’s ability to distinguish true nodules from artifacts and significantly reducing the false positive rate.

We conducted comprehensive experimental evaluations on the LUNA16 dataset, and the results demonstrate that the proposed method achieves high recall while significantly improving detection accuracy. Compared with several state-of-the-art detection models, our method shows superior performance, validating its effectiveness and robustness. Future work will focus on enhancing the clinical interpretability and lightweight deployment of the model to promote its application in real-world CAD of lung cancer.

Nevertheless, despite the proposed method achieving certain results in experimental performance, several limitations deserve further attention. The training and validation of the model rely primarily on the LUNA16 dataset, whose homogeneous imaging conditions and single-source origin may restrict the model’s generalizability in real-world clinical environments. The lack of independent validation on real clinical data remains one of the major limitations of the current work. Moreover, the detection of nodules smaller than 3 mm or those closely attached to the pleura or vasculature remains challenging, as such nodules typically exhibit low contrast and irregular morphology, placing higher demands on feature representation. In terms of computational efficiency, although the two-stage architecture performs well overall, the first stage still requires dense feature extraction across the full 3D CT volume, resulting in considerable computational overhead that may hinder deployment on resource-constrained systems. In addition, although Grad-CAM visualizations provide some insight into the model’s behavior, the decision-making process retains a degree of “black-box” characteristics, indicating that further efforts are needed to enhance interpretability.

Future work will focus on improving the model’s generalizability to real clinical data, optimizing computational efficiency, strengthening interpretability, and validating the approach on larger and multi-center datasets.

Acknowledgements

The authors would like to thank the project team for providing access to computing resources.

ORCID iD

Lixin Wang  <https://orcid.org/0009-0003-7207-0024>

Author contributions

Lixin Wang contributed to methodology, investigation, formal analysis, and writing the original draft. Xiaowen Lan was responsible for conceptualization and writing—review & editing. Kaikai Zhang contributed to conceptualization and visualization, while Yanhui Wang was involved in formal analysis and visualization. Shaofeng Wang and Wenjing Liu managed project administration and participated in writing—review & editing. All authors have read and approved the final manuscript.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Natural Science Foundation of Inner Mongolia Autonomous Region (grant number 2025LHMS06016).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

The dataset utilized in this study is available at <https://luna16.grand-challenge.org/>.

References

- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74: 229–263.
- Halder A, Dey D and Sadhu AK. Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review. *J Digit Imaging* 2020; 33: 655–677.
- Nooreldeen R and Bach H. Current and future development in lung cancer diagnosis. *Int J Mol Sci* 2021; 22: 8661.
- Valente IRS, Cortez PC, Neto EC, et al. Automatic 3D pulmonary nodule detection in CT images: a survey. *Comput Methods Programs Biomed* 2016; 124: 91–107.
- Quanyang W, Yao H, Sicong W, et al. Artificial intelligence in lung cancer screening: detection, classification, prediction, and prognosis. *Cancer Med* 2024; 13: e7140.
- Deng K, Wang L, Liu Y, et al. A deep learning-based system for survival benefit prediction of tyrosine kinase inhibitors and immune checkpoint inhibitors in stage IV non-small cell lung cancer patients: A multicenter, prognostic study. *EClinicalMedicine* 2022; 51: 101541.
- Jiang B, Li N, Shi X, et al. Deep learning reconstruction shows better lung nodule detection for ultra-low-dose chest CT. *Radiology* 2022; 303: 202–212.
- Xu Y, Hosny A, Zeleznik R, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res* 2019; 25: 3266–3275.
- Lu L, Tan Y, Schwartz LH, et al. Hybrid detection of lung nodules on CT scan images. *Med Phys* 2015; 42: 5042–5054.
- Gu Y, Lu X, Zhang B, et al. Automatic lung nodule detection using multi-scale dot nodule-enhancement filter and weighted support vector machines in chest computed tomography. *PLoS One* 2019; 14: e0210551.
- Ali H, Mohsen F and Shah Z. Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review. *BMC Med Imaging* 2023; 23: 129.
- Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans Med Imaging* 2016; 35: 1160–1169.
- Jiang H, Ma H, Qian W, et al. An automatic detection system of lung nodule based on multigroup patch-based deep learning network. *IEEE J Biomed Health Inform* 2017; 22: 1227–1237.
- Zuo W, Zhou F, Li Z, et al. Multi-resolution CNN and knowledge transfer for candidate classification in lung nodule detection. *Ieee Access* 2019; 7: 32510–32521.
- Xie H, Yang D, Sun N, et al. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognit* 2019; 85: 109–119.
- Huang X, Shan J and Vaidya V. Lung nodule detection in CT using 3D convolutional neural networks. In: 2017 *IEEE 14th international symposium on biomedical imaging (ISBI 2017)* 2017, pp.379–383. IEEE.
- Wang J, Wang J, Wen Y, et al. Pulmonary nodule detection in volumetric chest CT scans using CNNs-based nodule-size-adaptive detection and classification. *IEEE Access* 2019; 7: 46033–46044.
- Cao H, Liu H, Song E, et al. A two-stage convolutional neural networks for lung nodule detection. *IEEE J Biomed Health Inform* 2020; 24: 2006–2015.
- Zhu W, Liu C, Fan W, et al. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In: 2018 *IEEE winter conference on applications of computer vision (WACV)* 2018, pp.673–681. IEEE.
- Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)* 2018, pp.3–19.
- Setio AAA, 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. *Med Image Anal* 2017; 42: 1–13.
- Armato SG III, McLennan G, Bidaut L, 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* 2011; 38: 915–931.

23. Pereira FR, De Andrade JMC, Escuissato DL, et al. Classifier ensemble based on computed tomography attenuation patterns for computer-aided detection system. *IEEE Access* 2021; 9: 123134–123145.
24. Yuan H, Wu Y, Cheng J, et al. Pulmonary nodule detection using 3-d residual u-net oriented context-guided attention and multi-branch classification network. *Ieee Access* 2021; 10: 82–98.
25. Zhao D, Liu Y, Yin H, et al. An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image. *Expert Syst Appl* 2023; 211: 118672.
26. Usman M, Rehman A, Shahid A, et al. Meds-net: multi-encoder based self-distilled network with bidirectional maximum intensity projections fusion for lung nodule detection. *Eng Appl Artif Intell* 2024; 129: 107597.
27. Xiong Y, Deng L and Wang Y. Pulmonary nodule detection based on model fusion and adaptive false positive reduction. *Expert Syst Appl* 2024; 238: 121890.
28. Tang C, Zhou F, Sun J, et al. Lung-YOLO: multiscale feature fusion attention and cross-layer aggregation for lung nodule detection. *Biomed Signal Process Control* 2025; 99: 106815.
29. Almahasneh M, Xie X and Paiement A. Attentnet: fully convolutional 3d attention for lung nodule detection. *SN Comp Sci* 2025; 6: 292.
30. Huang H-H, Zhao Y, Wei S-Y, et al. Lightweight lung-nodule detection model combined with multidimensional attention convolution. *Curr Med Imaging* 2025; 21: e15734056310722.
31. Pezeshk A, Hamidian S, Petrick N, et al. 3-D Convolutional neural networks for automatic detection of pulmonary nodules in chest CT. *IEEE J Biomed Health Inform* 2018; 23: 2080–2090.
32. Huang W, Xue Y and Wu Y. A CAD system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning. *Plos One* 2019; 14: e0219369.
33. Li Y and Fan Y. DeepSEED: 3D squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection. In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI)* 2020, pp.1866–1869. IEEE.
34. Mei J, Cheng M-M, Xu G, et al. SANet: a slice-aware network for pulmonary nodule detection. *IEEE Trans Pattern Anal Mach Intell* 2021; 44: 4374–4387.
35. Tan Y, Fu X, Zhu J, et al. A improved detection method for lung nodule based on multi-scale 3D convolutional neural network. *Concurr Comput Pract Exp* 2023; 35: e7034.
36. Lu X, Zeng N, Wang X, et al. Ffnet: an end-to-end framework based on feature pyramid network and filter network for pulmonary nodule detection. In: *2023 IEEE 20th international symposium on biomedical imaging (ISBI)* 2023, pp.1–6. IEEE.