

Lightweight Pure 3D Lung Nodule Detection and Segmentation Network with Asymmetric Feature Interaction

Dongjie Jiao¹

Luohe Medical College

Yongchao Ma¹

13839527381@163.com

Luohe Medical College

Minli Wang

Second Affiliated Hospital of Luohe Medical College

Ming Wu

Luohe Medical College

Research Article

Keywords: medical image processing, pulmonary nodule detection and segmentation, three-dimensional convolutional neural network, asymmetric depthwise separable convolution, bidirectional feature interaction, edge computing

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11161081/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Lightweight Pure 3D Lung Nodule Detection and Segmentation Network with Asymmetric Feature Interaction

Dongjie Jiao¹, Yongchao Ma^{1*}, Minli Wang² and Ming Wu¹

¹ School of Public Health and Health Nutrition, Luohe Medical College, Luohe 462000, China

² The Second Affiliated Hospital of Luohe Medical College, Luohe 462300, China

*Correspondence: Yongchao Ma, E-mail: 13839527381@163.com

Dongjie Jiao. E-mail: jiaodongjie1999@163.com. ORCID iD: 0009-0002-8758-9980

Abstract

Three-dimensional convolutional neural networks (3D CNNs) for lung nodule analysis face a tension between accuracy and computational cost: isotropic cubic kernels match anisotropic CT data poorly, detection and segmentation are modeled in isolation, and parameter-constrained networks collapse to the trivial all-background solution under class imbalance. This study proposes Light3D-NoduleNet, a lightweight pure 3D joint detection and segmentation network for edge deployment. An asymmetric 3D depthwise separable convolution factorizes the cubic kernel into in-plane and inter-slice branches, cutting kernel parameters by about 56%; a scale-aware efficient channel attention adapts its kernel size to the channel count; and a bidirectional feature interaction module exchanges a spatial attention map and a channel attention vector between the two branches, closing the loop between tasks. Joint training on LUNA16 and LIDC-IDRI uses dynamic loss masking. On LIDC-IDRI, the model reaches a Dice coefficient of 0.5233 (median 95% Hausdorff distance 6.00 mm) with 5.24 M parameters and 178.3 G FLOPs, exceeding reproduced 3D U-Net and V-Net by 0.1545 and 0.0701. Three variants lacking bidirectional interaction all stagnate at 0.2302 after 100 epochs, whereas the full model converges to 0.5233. On a 15 W Jetson Orin NX it runs at 3.0 FPS with 435 MB VRAM, indicating that bidirectional feature interaction is a necessary architectural condition for lightweight 3D networks to escape the trivial solution under extreme foreground-background imbalance.

Keywords medical image processing, pulmonary nodule detection and segmentation, three-dimensional convolutional neural network, asymmetric depthwise separable convolution, bidirectional feature interaction, edge computing

1 Introduction

1.1 Research background and technical challenges

Lung cancer has the highest incidence and mortality of all malignancies worldwide [1, 2]. Prognosis depends strongly on the stage at diagnosis: the five-

year survival rate for stage I disease reaches 60%–80%, whereas for stage IV it falls below 5% [3]. Low-dose computed tomography (LDCT) screening reduces lung cancer mortality by approximately 20%–26% and has become an effective means of early nodule detection [4, 5]. Large-scale screening, however, generates enormous volumetric imaging data—a single CT examination contains 200–500 slices—and manual slice-by-slice interpretation is time-consuming and prone to visual fatigue, with inter-reader agreement for nodule assessment showing Kappa coefficients of only 0.4–0.6 [6, 7]. High-performance computer-aided detection and diagnosis (CAD) systems therefore carry considerable clinical value for improving screening efficiency and reading consistency.

Deep convolutional neural networks have become the mainstream approach to pulmonary nodule CAD. However, 3D convolutional networks face three coupled technical challenges in nodule analysis.

(1) anisotropic data versus isotropic convolution kernels. In-plane pixel spacing in CT typically ranges from 0.5 to 0.8 mm, whereas inter-slice spacing ranges from 1.0 to 2.5 mm—a three- to five-fold anisotropy. Standard 3D convolution applies an isotropic cubic kernel ($3 \times 3 \times 3$) that treats the three spatial dimensions equally. This produces substantial redundant computation along the slice direction, where information density is lower, and leads to inefficient use of parameters.

(2) isolated modelling of detection and segmentation. Detection answers “where is the nodule”; segmentation answers “how large is the nodule”. The two tasks rely on highly overlapping feature representations. Yet existing methods mostly model them independently, requiring separate models and inference pipelines, with task heads that share only backbone features and exchange no explicit information. This precludes synergistic gains.

(3) convergence difficulty under extreme class imbalance. In a $64 \times 64 \times 64$ voxel lung nodule patch, foreground (nodule) voxels typically account for less than 1% of the total. Under such extreme imbalance, parameter-constrained lightweight 3D networks readily converge to the trivial solution that classifies every voxel as background—a local minimum from which standard supervised training struggles to escape. Which architectural conditions can break this convergence barrier remains an open question in lightweight 3D medical image analysis.

These three challenges are interrelated: anisotropic convolution redundancy limits parameter efficiency, task isolation weakens feature discriminability, and the combination of restricted parameters with extreme class imbalance directly

causes training failure. Addressing all three within a unified lightweight framework is the core motivation of this work.

1.2 Three key gaps in existing methods

(1) no lightweight pure 3D joint detection-segmentation solution exists.

Detection and segmentation are mostly treated as independent tasks requiring separate models and inference pipelines. Joint frameworks such as Mask R-CNN have succeeded on 2D natural images, yet a lightweight realization for 3D medical imaging is still lacking. Integrating both tasks into a single end-to-end lightweight network would reduce redundant computation and exploit feature complementarity through multi-task learning.

(2) the interaction between attention mechanisms and parameter-efficient convolution in pure 3D networks has not been systematically studied. The inherent channel isolation of depthwise separable convolution limits cross-channel feature interaction—a limitation that matters particularly for medical images, whose discriminative features often span multiple channels. ECA-Net achieves efficient channel attention at negligible parameter cost, but to our knowledge neither its scale-aware extension to pure 3D networks nor its co-design with asymmetric depthwise separable convolution has been systematically investigated.

(3) the architectural conditions under which lightweight 3D networks converge successfully under extreme class imbalance have not been systematically studied. Existing work focuses on training techniques such as loss re-weighting and hard example mining, but provides no systematic analysis of how network architecture itself—in particular the construction of inter-task information pathways—affects convergence. Which missing architectural components cause complete training failure, and which can break the convergence barrier, remain unanswered. This work gives a quantitative answer through controlled ablation experiments.

1.3 Main contributions

To address the three gaps above, we propose Light3D-NoduleNet. The main contributions are as follows.

(1) An asymmetric 3D depthwise separable convolution tailored to the anisotropic properties of CT data, reducing convolutional kernel parameters by approximately 56%.

(2) A 3D scale-aware efficient channel attention with adaptive kernel size for cross-level receptive field modulation.

(3) A bidirectional feature interaction joint decoder that establishes closed-loop synergistic optimization between detection and segmentation.

(4) A three-stage progressive fine-tuning strategy that raises Dice from 0.424 to 0.523 (+23.4%).

(5) Through systematic ablation experiments, an architectural principle of general significance: bidirectional feature interaction is a necessary architectural condition for lightweight 3D networks to escape the trivial solution under extreme foreground-background imbalance. Three variants lacking this component (baseline, +Asym, +Asym+ECA) all stagnate at an identical Dice of 0.2302 after 100 epochs, whereas only the full model containing bidirectional interaction breaks through to 0.4240 in the main training stage and reaches 0.5233 after three-stage progressive fine-tuning. This convergence-versus-non-convergence dichotomy, rather than a quantitative difference, provides a reusable guiding principle for the architectural design of lightweight 3D medical imaging networks.

2 Related work

2.1 Pulmonary nodule detection and segmentation

Research on pulmonary nodule CAD dates back to the 1990s. Early work followed a cascade pipeline of candidate generation, handcrafted feature extraction and classifier discrimination [8]. For candidate generation, multi-threshold grey-level morphology [9], Hessian matrix eigenvalue analysis [10] and rule-based region growing [11] were commonly used to extract suspected nodule regions from lung parenchyma masks. For feature extraction, researchers designed handcrafted features covering grey-level statistics (mean, variance, skewness, kurtosis, histogram entropy), morphology (sphericity, compactness, spiculation index, lobulation index), texture (contrast, energy, homogeneity and correlation from the grey-level co-occurrence matrix; short- and long-run emphasis from the grey-level run-length matrix) and histogram of oriented gradients [12]. For classification, support vector machines, random forests and AdaBoost were widely adopted [13]. Handcrafted features, however, depend heavily on domain expertise, have limited representational capacity, and generalize poorly across scanning protocols and reconstruction kernels.

After 2015, deep convolutional neural networks rapidly displaced traditional machine learning approaches to CAD. Setio et al. [14] systematically reviewed deep learning for pulmonary nodule detection and classified mainstream methods into 2D CNN, 2.5D multi-stream fusion CNN and 3D CNN. In terms of detection performance, the 3D variant of Faster R-CNN proposed by Ding et al. [15] achieved a competition performance metric (CPM) of 0.891 on LUNA16, while Liao et al. [16] introduced a 3D region proposal network with a false-

positive reduction module, raising CPM to 0.947. For segmentation, 3D U-Net [17] and its variants (V-Net [18], Attention U-Net [19], nnU-Net [20]) constitute the mainstream route. The parameter counts and computational costs of these high-accuracy models, however, impede their application in resource-constrained settings—precisely the problem this work addresses.

For false-positive suppression in nodule detection, mainstream methods adopt a two-stage paradigm of candidate generation followed by false positive reduction (FPR). Setio et al. [14] systematically compared FPR strategies in the LUNA16 challenge and found that candidate classifiers based on multi-view 3D CNNs substantially reduce false positive rates; subsequent work further introduced 3D region proposal networks [16], multi-scale feature fusion and attention mechanisms to sharpen candidate discrimination. These methods share a common characteristic: a high-recall detector first generates numerous candidates, and an independent FPR network then performs binary classification, maintaining high sensitivity while suppressing false alarms. Introducing an independent FPR module, however, incurs additional parameters and computation, creating tension with lightweight design objectives. This work therefore examines the feasibility and gain of a compact FPR module within a lightweight joint network (Section 4.3.2).

Recent years have seen sustained attention on lightweight 3D medical image analysis for resource-constrained settings. One line of work starts from convolution factorization and structural redesign, compressing model size through depthwise separable convolution, inverted residual structures or grouped convolution, such as lightweight encoder designs for 3D medical segmentation [42, 43]; another introduces vision transformers or hybrid architectures to enhance long-range dependency modelling, although their parameter counts and computational costs typically exceed those of convolutional networks of comparable scale [44]. Several explorations of joint nodule detection and segmentation adopt a shared backbone with dual-branch outputs [35], but explicit inter-branch information interaction mechanisms are generally lacking. Overall, existing lightweight approaches have made substantial progress in parameter compression, yet two issues still lack systematic study and clear architectural guidance: training stability under extreme class imbalance, and how to construct effective information pathways between tasks. This work is situated in this context.

Regarding 3D versus 2.5D methods, pure 3D approaches, while somewhat higher in FLOPs (178 G vs. 50–100 G), possess an intrinsic advantage in Z-axis boundary accuracy: 3D convolution can directly perceive grey-level variation

trends across adjacent slices and naturally discriminates between vessel cross-sections extending along the Z axis and spheroidal nodules. By contrast, 2.5D methods often render obliquely oriented vessels and bronchial walls as quasi-circular cross-sections in orthogonal planes, increasing false positive rates. The pure 3D method proposed here achieves a median HD95 of 6.00 mm on the validation set, with no evident degradation of boundary accuracy arising from missing inter-slice information.

2.2 Lightweight network design

To reduce the storage and computational overhead of deep neural networks, researchers have explored several directions. The MobileNet series [21, 22] pioneered the decomposition of standard convolution into depthwise and pointwise convolution, reducing both computation and parameters to approximately one-ninth of standard convolution. The ShuffleNet series [23, 24] addressed the impeded inter-group information flow in depthwise separable convolution through channel shuffling. In 3D medical imaging, Lachinov et al. [25] transferred the inverted residual bottleneck of MobileNetV2 to a 3D U-Net encoder, compressing parameters to 30% of the baseline for brain tumour segmentation. Heinrich et al. [26] proposed a lightweight registration network combining a MobileNet encoder with a 3D decoder. More recently, structural re-parameterization [27], neural architecture search [28] and knowledge distillation [29] have also been applied to design efficient medical segmentation networks. Directly transferring 2D lightweight design patterns developed for natural images (such as depthwise separable convolution) to 3D medical imaging, however, does not adequately account for the anisotropic physical properties of CT data—in-plane pixel spacing of 0.5–0.8 mm versus inter-slice spacing of 1.0–2.5 mm, a three- to five-fold difference. The proposed asymmetric 3D-DSC is a customized design targeting precisely this physical characteristic.

2.3 Attention mechanisms

Attention mechanisms have become a key technique for enhancing representational capacity in deep learning. SENet [30] learns non-linear inter-channel dependencies through global average pooling followed by a two-layer fully connected network. CBAM [31] further introduces a spatial attention branch complementary to channel attention. The dimensionality reduction in SENet, however, inevitably compresses channel information. ECA-Net [32] showed that appropriate local cross-channel interaction can match or surpass SENet while avoiding dimensionality reduction, at minimal parameter cost (only k parameters). Building on ECA-Net, this work extends it to 3D and introduces a scale-aware kernel size adaptation mechanism that automatically adjusts the attention

receptive field according to network depth, better accommodating the cross-scale feature requirements of medical images ranging from small nodules to global anatomical structures.

2.4 Multi-task joint learning

Object detection and semantic segmentation are two highly related pixel-level prediction tasks. Mask R-CNN [33] achieved simultaneous output of detection boxes and instance masks by adding a parallel fully convolutional segmentation branch to Faster R-CNN. In medical imaging, joint detection and segmentation can supply both “where” (localization) and “how large” (contour) information about nodules. Most existing multi-task networks, however, reuse shallow features merely by sharing a backbone feature extractor, with no explicit information interaction between task heads [34, 35]. YOLACT [36] realizes real-time instance segmentation by linearly combining mask coefficients with prototype masks, but its inter-task interaction remains limited. The proposed bidirectional feature interaction mechanism establishes a genuine two-way information channel between detection and segmentation through two complementary gating mechanisms: spatial attention ($D \rightarrow S$) and channel attention ($S \rightarrow D$).

2.5 Edge computing and deployment

Edge computing aims to move data processing and model inference from cloud data centers to network edge nodes close to the data source [37]. In medical imaging, edge inference offers notable advantages: protection of patient privacy (avoiding upload to the cloud), reduced network latency (real-time diagnostic feedback) and suitability for field or mobile scenarios (offline operation of portable devices). The NVIDIA Jetson series (Nano, TX2, Xavier NX, Orin) is currently the most widely used edge AI computing platform, with power consumption ranging from 5 W to 30 W. Because the computational capability and storage of edge devices fall far short of desktop GPUs, model compression and acceleration techniques—including quantization [38], pruning [39], knowledge distillation [29] and lightweight architecture design [21–24]—are prerequisites for deploying deep learning models at the edge. Starting from lightweight architecture design, this work constructs an efficient model during training through asymmetric convolution factorization and channel attention, satisfying edge deployment constraints without post-hoc compression.

3 Network architecture

3.1 Overall architecture

The overall architecture of Light3D-NoduleNet follows the design philosophy of “lightweight encoding → deep decoding → task synergy”, as illustrated in Fig. 1. The network input is a preprocessed and cropped $Z \times Y \times X = 64 \times 64 \times 64$ volumetric CT cube. The output consists of four components: (1) a center-point heatmap $H \in [0,1]^{64 \times 64 \times 64}$ indicating the probability of a nodule center at each voxel; (2) an offset regression map $O \in \mathbb{R}^{3 \times 64 \times 64 \times 64}$ correcting sub-voxel center displacement introduced by downsampling; (3) a size regression map $S \in \mathbb{R}_+^{1 \times 64 \times 64 \times 64}$ estimating nodule diameter; and (4) a segmentation mask $M \in [0,1]^{1 \times 64 \times 64 \times 64}$ predicting the nodule region voxel-wise.

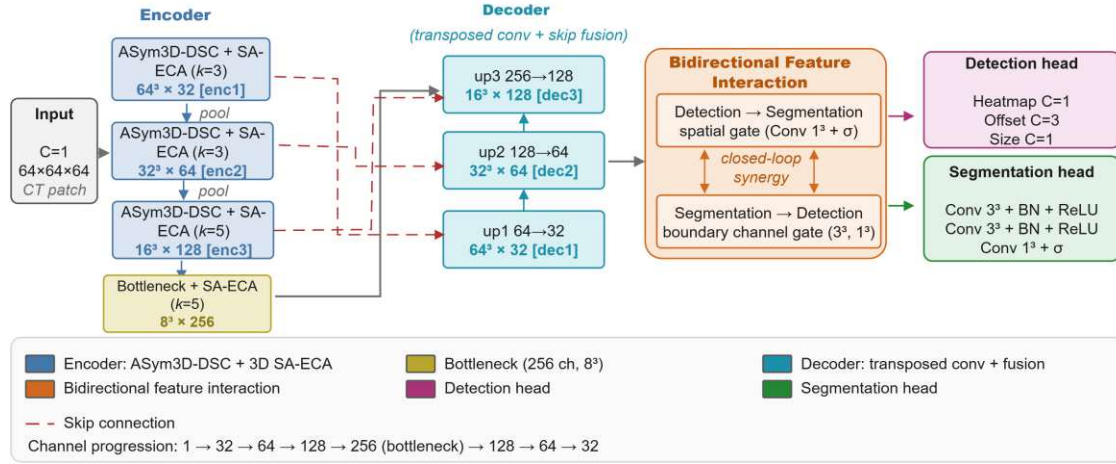


Fig. 1 Overall architecture of Light3D-NoduleNet. The single-channel CT patch ($64 \times 64 \times 64$) is processed by a stem convolution and three encoder levels, each stacking two asymmetric 3D depthwise separable convolutions (ASym3D-DSC) with 3D scale-aware efficient channel attention (SA-ECA, kernel size k adapts from 3 to 5 with depth); the bottleneck holds 256 channels at 8^3 . The decoder performs three transposed-convolution upsampling stages with skip connections from the corresponding encoder levels. At the decoder output, the bidirectional feature interaction module exchanges information between the two branches through a detection-to-segmentation spatial gate and a segmentation-to-detection boundary channel gate, forming closed-loop synergy. The detection head outputs a center heatmap, an offset map and a size map; the segmentation head outputs the nodule mask. Channel progression: $1 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32$.

The encoder comprises four levels of stacked asymmetric 3D-DSC modules, performing progressive spatial downsampling through stride-2 3D convolution. Each encoder level contains two consecutive AsymDSC-ECA basic modules (defined in Sections 3.2 and 3.3), with spatial dimensions halved and channel counts doubled at each level: $64^3 \times 32 \rightarrow 32^3 \times 64 \rightarrow 16^3 \times 128 \rightarrow 8^3 \times 256$. The bottleneck layer further compresses features to $8^3 \times 256$. The decoder is symmetric, progressively upsampling through stride-2 3D transposed convolution and fusing fine-grained spatial information from the corresponding encoder level through skip connections.

At the decoder output, the shared feature map flows into two parallel branches: the detection branch outputs heatmap, offset and size heads through $1 \times 1 \times 1$ convolution, while the segmentation branch outputs the segmentation mask through two $3 \times 3 \times 3$ convolutions and one $1 \times 1 \times 1$ convolution. Before output, the two branches exchange explicit cross-task information through the bidirectional feature interaction module (defined in Section 3.4).

The encoder-decoder architecture rests on the following design philosophy. The encoder extracts multi-scale semantic features through progressive downsampling of asymmetric 3D-DSC modules, with ECA attention adaptively adjusting inter-channel importance weights at every level. Bidirectional feature interaction at the bottleneck breaks the information barrier between detection and segmentation, allowing the two gradient flows to be optimized synergistically rather than interfering with one another. The decoder fuses fine-grained encoder spatial information through skip connections, progressively restoring the original resolution. The design as a whole balances parameter efficiency against representational capacity, achieving competitive detection and segmentation accuracy under a constraint of only 5.24 M parameters.

To alleviate channel distribution differences between encoder and decoder features, a lightweight channel adaptation module (CAM) is introduced in the skip connections, performing channel alignment of fused features through $1 \times 1 \times 1$ pointwise convolution, batch normalization and ReLU activation, adding only $C \times C$ parameters (e.g. only 4,096 parameters when $C = 64$).

3.2 Asymmetric depthwise separable convolution

The standard 3D convolution operation can be formulated as

$$Y_{(d,h,w)}^{C_{out}} = \sum_{C_{in}=1}^{C_{in}} \sum_{i=-1}^1 \sum_{j=-1}^1 \sum_{k=-1}^1 W_{(i,j,k)}^{C_{out}, C_{in}} \cdot X_{(d+i, h+j, w+k)}^{C_{in}} + b^{C_{out}} \quad (1)$$

where $W \in \mathbb{R}^{C_{out} \times C_{in} \times 3 \times 3 \times 3}$ is the convolution kernel weight tensor with $C_{in} \cdot C_{out} \cdot 27$ parameters. For a basic module with $C_{in} = 32$ and $C_{out} = 32$, a single $3 \times 3 \times 3$ convolution requires $32 \times 32 \times 27 = 27,648$ parameters. As the network deepens and the channel count grows to 256, the parameter count of a single convolution rises to $256 \times 256 \times 27 \approx 1.77$ M.

The core insight of this work is that in-plane CT resolution (typically 0.6–0.9 mm per pixel) is far finer than inter-slice spacing (typically 1.0–2.5 mm per slice)—a three- to five-fold discrepancy. Adjacent pixels within the X-Y plane are therefore highly correlated and are suitably captured by larger kernels encoding nodule boundary contours and internal density texture, whereas along the Z axis the larger inter-slice spacing means lower inter-slice information density, so

smaller kernels suffice to capture inter-slice grey-level variation trends. Based on this insight, the standard $3 \times 3 \times 3$ convolution is factorized into:

(a) Spatial convolution Ψ_{spa} : a $1 \times 3 \times 3$ convolution operating within the X-Y plane:

$$\Psi_{\text{spa}}(X)_{(d,h,w)} = \sum_{j=-1}^1 \sum_{k=-1}^1 W_{(0,j,k)}^{\text{spa}} \cdot X_{(d,h+j,w+k)} \quad (2)$$

(b) Inter-slice convolution Ψ_{dep} : a $3 \times 1 \times 1$ convolution operating along the Z axis:

$$\Psi_{\text{dep}}(X)_{(d,h,w)} = \sum_{i=-1}^1 W_{(i,0,0)}^{\text{dep}} \cdot X_{(d+i,h,w)} \quad (3)$$

The asymmetric 3D depthwise separable convolution combines this factorization with per-channel depthwise convolution, giving a complete module with only $C_{\text{in}} \cdot (9 + 3) + C_{\text{in}} \cdot C_{\text{out}} = C_{\text{in}} \cdot (12 + C_{\text{out}})$ parameters. For $C_{\text{in}} = C_{\text{out}} = 32$, this amounts to $32 \times (12 + 32) = 1,408$ parameters—only 5.1% of the standard 3D convolution (27,648).

3.3 Scale-aware channel attention

The ECA module [32] implements cross-channel interaction through one-dimensional convolution without dimensionality reduction, at very low parameter cost. The original ECA, however, supports only 2D feature maps (global average pooling compressed to $1 \times 1 \times C$); extending it directly to 3D requires compressing the 3D global average pooling to $1 \times 1 \times 1 \times C$ before applying 1D convolution.

This work further introduces a scale-aware kernel size adaptation mechanism. Given an input feature channel count C , the adaptive kernel size k is computed as

$$k = \psi(C) = \left\lceil \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rceil_{\text{odd}} \quad (4)$$

where $\gamma = 2$, $b = 1$, and $\lceil \cdot \rceil_{\text{odd}}$ denotes rounding to the nearest odd number. Substituting the channel counts of each network level gives encoder level 1 $C = 32 \rightarrow k = 3$, level 2 $C = 64 \rightarrow k = 3$, level 3 $C = 128 \rightarrow k = 5$, and bottleneck $C = 256 \rightarrow k = 5$. The attention kernel size thus adapts upward with network depth: shallow layers ($k = 3$) focus on fine nodule texture discrimination within local channel groups, while deep layers ($k = 5$) span larger channel groups to capture the contextual dependencies of global anatomical structures (vessels, bronchi, pleura). Under the channel configuration used here ($32 \rightarrow 256$), the adaptive kernel size increases from 3 to 5 at the third level ($C = 128$), so the latter two levels employ a wider channel interaction range.

The complete computation flow of 3D SA-ECA is

$$F_{\text{avg}} = \text{AdaptiveAvgPool3D}(F) \in \mathbb{R}^{C \times 1 \times 1 \times 1} \quad (5)$$

$$F_{att} = \sigma(\text{Conv1D}_k(F_{avg})) \in \mathbb{R}^{C \times 1 \times 1 \times 1} \quad (6)$$

$$F_{out} = F \odot F_{att} \quad (7)$$

where σ is the sigmoid activation, Conv1D_k is a one-dimensional convolution (kernel size k , padding $k//2$), and $\odot$ denotes element-wise multiplication.

3.4 Bidirectional feature interaction

Definition 3.1 (Detection-to-segmentation (D→S) spatial attention). Let $F_{det} \in \mathbb{R}^{C_{det} \times D \times H \times W}$ and $F_{seg} \in \mathbb{R}^{C_{seg} \times D \times H \times W}$ denote the intermediate feature maps of the detection and segmentation branches, respectively. The D→S path first compresses F_{det} into a single-channel spatial attention map $A_{D \rightarrow S} \in [0,1]^{1 \times D \times H \times W}$ through $1 \times 1 \times 1$ convolution and sigmoid activation, then applies it element-wise to F_{seg} :

$$F_{seg}^{enhanced} = F_{seg} \odot \text{expand}(A_{D \rightarrow S}) + F_{seg} \quad (8)$$

where $\text{expand}(\cdot)$ broadcasts the single-channel attention map along the channel dimension to C_{seg} . $A_{D \rightarrow S}$ approaches 1 near the nodule center and 0 in background regions, acting as a spatial gate.

Definition 3.2 (Segmentation-to-detection (S→D) channel attention). The S→D path first applies $3 \times 3 \times 3$ per-channel depthwise convolution (groups = C) to the shared features to extract boundary-sensitive features, then compresses them into a channel attention vector $A_{S \rightarrow D} \in [0,1]^C$ through $1 \times 1 \times 1$ pointwise convolution and sigmoid activation:

$$F_{det}^{enhanced} = F_{det} \odot \text{expand}(A_{S \rightarrow D}) + F_{det} \quad (9)$$

$A_{S \rightarrow D}$ assigns different scaling weights to different channels, enhancing channel responses related to nodule boundaries.

3.5 Loss function and training strategy

The total loss is a weighted sum of detection and segmentation losses:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda \cdot \mathcal{L}_{seg} \quad (10)$$

Detection loss $\mathcal{L}_{det}$ has three parts. (1) A center-point heatmap focal loss addressing extreme positive-negative imbalance, using a 3D Gaussian heatmap H_{gt} centered at the ground-truth nodule center as target:

$$\mathcal{L}_{heat} = -\frac{1}{N} \sum \left[(1 - H_{pred})^\alpha \log(H_{pred}) \cdot \mathbb{1}_{H_{gt} > 0.5} + (1 - H_{gt})^\beta (H_{pred})^\alpha \log(1 - H_{pred}) \cdot \mathbb{1}_{H_{gt} \leq 0.5} \right] \quad (11)$$

where $\alpha = 2$, $\beta = 4$ and N is the number of positive samples. (2) An offset L1 loss applied only at positive sample locations. (3) A size L1 loss applied only at positive sample locations.

Segmentation loss L_{seg} uses the Dice loss:

$$\mathcal{L}_{seg} = 1 - \frac{2 \sum_i P_i G_i + \epsilon}{\sum_i P_i + \sum_i G_i + \epsilon} \quad (12)$$

Dynamic loss masking: since LUNA16 provides only coordinate annotations and no segmentation masks, the segmentation loss weight λ is automatically set to zero when the training batch originates from LUNA16:

$$\lambda = \begin{cases} 0.5 \cdot (1.0 + \text{has_nodule}), & \text{if sample from LIDC-IDRI} \\ 0, & \text{if sample from LUNA16} \end{cases}$$

3.6 Progressive fine-tuning strategy

After 150 epochs of main training, a three-stage progressive unfreezing strategy is applied to maximize model performance.

Stage 1 (segmentation head only, 15 epochs). All encoder and decoder convolution and normalization parameters are frozen, and only the $1 \times 1 \times 1$ output convolution at the end of the segmentation head is optimized (41,633 trainable parameters, 0.79% of the total). The learning rate is $1e-4$ with the AdamW optimizer. The aim is to let the randomly initialised segmentation output head adapt to pretrained backbone features while fully protecting the learned encoder representation.

Stage 2 (decoder unfrozen, 15 epochs). Building on stage 1, the upsampling layers and the convolution blocks following skip-connection fusion in the decoder are additionally unfrozen, and the learning rate is reduced to $5e-5$. Importantly, the optimizer state is reinitialized at the stage transition (discarding the first- and second-order momentum accumulated in stage 1) to avoid momentum direction contamination. The aim is to let the decoder adapt to the finer geometric detail recovery required by segmentation while keeping the encoder frozen. Validation Dice in stage 2 (0.4549) is lower than in stage 1 (0.4975); this is because unfreezing the decoder temporarily moves the model away from the parameter region to which stage 1 had adapted—a normal phenomenon during progressive unfreezing. Stage 3, through full-model low-learning-rate fine-tuning with EMA smoothing, ultimately raises Dice to 0.5233.

Stage 3 (full-model EMA fine-tuning, 20 epochs). All trainable parameters are unfrozen under a differential learning-rate strategy: $1e-5$ for the encoder and $5e-5$ for the decoder and task heads. An exponential moving average (EMA, decay 0.999) is additionally introduced to smooth model parameters, and EMA parameters are used at inference in place of the raw training parameters. The aim is full-model collaborative fine-tuning at a very low learning rate, with EMA acting as an inertial stabiliser in parameter space to prevent severe Dice fluctuation or collapse.

Throughout the three-stage fine-tuning, all BN layers are frozen: the global running mean and variance accumulated during training are fixed, statistics are not updated, and the learnable parameters are set to non-trainable. This strategy effectively eliminates the training instability caused by BN statistic drift during small-batch fine-tuning and constitutes an important engineering safeguard for stable Dice improvement.

4 Results and analysis

4.1 Datasets and experimental setup

4.1.1 Datasets.

LUNA16 (LUng Nodule Analysis 2016) is derived from a subset of the publicly available LIDC-IDRI dataset [14], comprising 888 thoracic low-dose CT scans (slice thickness ≤ 2.5 mm, tube voltage 120–140 kVp). Four experienced thoracic radiologists performed two-stage independent blinded annotation, ultimately identifying 1,186 pulmonary nodules with diameter ≥ 3 mm as positive samples. Each nodule is provided with its three-dimensional center coordinate (x, y, z) in world coordinates and its equivalent diameter (mm). Following the official recommendation, subsets 0–5 (534 cases, 712 nodules) form the training set, subsets 6–7 (178 cases, 237 nodules) the validation set, and subsets 8–9 (176 cases, 237 nodules) the test set.

LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) was constructed with funding from the US National Cancer Institute [40] and contains DICOM images of 1,018 thoracic CT examinations together with independent XML annotations from four radiologists. For each nodule, each radiologist annotated the nodule contour (slice-by-slice edge map coordinates) and nine semantic feature scores (malignancy 1–5, texture 1–5, spiculation 1–5, lobulation 1–5, margin 1–5, sphericity 1–5, internal structure 1–5, calcification 1–6, subtlety 1–5). This work adopted nodules with diameter ≥ 3 mm annotated consistently by at least three radiologists, using 880 cases with complete annotations and CT series. A random 70%/15%/15% split into training/validation/test sets was applied, ensuring that the CT series of the same patient were not divided across different sets.

4.1.2 Data preprocessing

All CT images passed through a unified preprocessing pipeline: (1) DICOM parsing using pydicom to extract the pixel array, HU rescaling parameters, pixel spacing, slice thickness and ImagePositionPatient metadata; (2) 3D resampling using SimpleITK cubic B-spline interpolation to resample all CT volumes to $1.0 \times 1.0 \times 1.0$ mm³ isotropic voxel resolution; (3) HU clipping and normalization to the lung window [-1000, 400] HU, linearly mapped to [0, 1]; (4) lung parenchyma

segmentation using threshold-based (normalized [0.05, 0.43], corresponding to -930 to -400 HU) 2D slice-wise connected component analysis, extracting the two largest connected regions as left and right lung masks and applying morphological closing with a 10-pixel kernel to fill internal vessel holes; and (5) 3D patch cropping of 64×64×64 voxel cubes centered on each nodule.

Note that the original LIDC-IDRI annotations were provided independently by four radiologists; during integration, a multi-annotator consensus principle was adopted (only nodules annotated by at least three radiologists were accepted), and all nodule contours of the same patient were rendered into the same mask. The segmentation target within a patch may therefore contain one or more nodule regions, and the Dice metric reflects the overall overlap of all foreground regions within that patch. Positive samples underwent random offset ($\leq \pm 5$ voxels) and random flipping (50% probability along each of the Z/Y/X axes) as data augmentation. Negative samples were randomly sampled within the lung parenchyma mask at a negative-to-positive ratio of approximately 3:1.

4.1.3 Training configuration

Hardware. NVIDIA GeForce RTX 3090 (24 GB GDDR6X, CUDA 12.1), Intel Xeon CPU, 64 GB DDR4 memory, Ubuntu 22.04 LTS (WSL 2 subsystem).

Software. Python 3.12.3, PyTorch 2.5.1, MONAI 1.5.2, SimpleITK 2.5.5, pydicom 3.0.2.

Hyperparameters. As described in Section 3.6, main training ran for 150 epochs with batch size 2 (gradient accumulation over 4 steps, effective batch size 8), the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $1e-5$), initial learning rate $1e-3$, and linear warm-up for 5 epochs followed by cosine annealing to 0. Three-stage fine-tuning comprised 50 epochs in total, configured as described in Section 3.6.

The evolution of validation Dice during training is shown in Fig. 2 and falls into three phases: phase I (epochs 1–5), where Dice rises rapidly from 0 to 0.10–0.15 as the model learns basic nodule morphology; phase II (epochs 6–50), fluctuating upward between 0.10 and 0.30, where difficult samples introduced by data augmentation may cause temporary regression; and phase III (epochs 51–150), stabilising in the 0.15–0.30 range with a best value of 0.424 near epoch 70. The gradual decline after epoch 70 suggests that the initial learning rate and cosine annealing schedule were not optimally configured; the subsequent three-stage progressive fine-tuning is a targeted response to this issue and successfully raises Dice to 0.523.

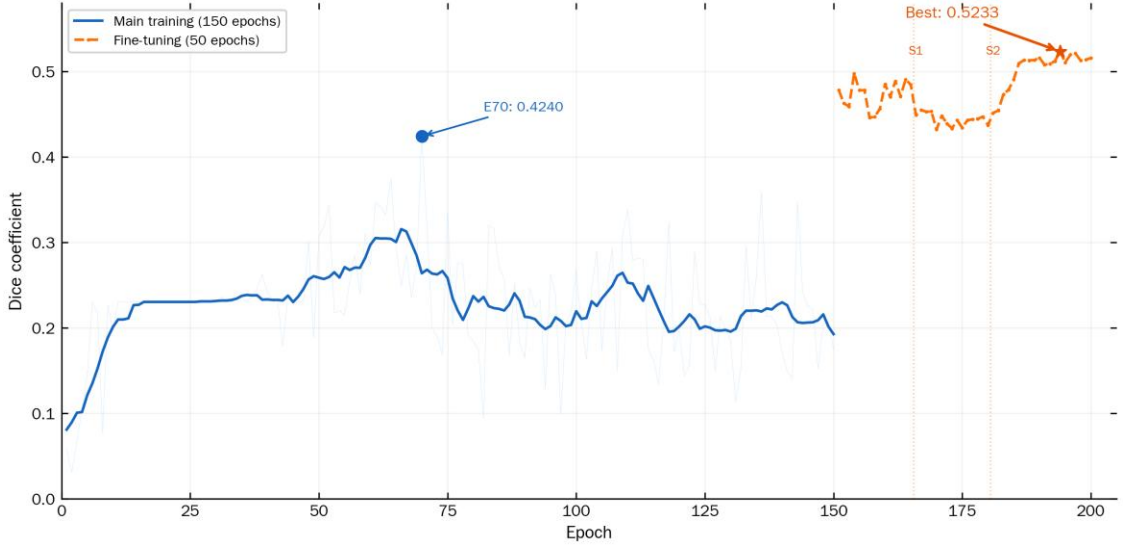


Fig. 2 Training convergence curve. Dice variation over 150 epochs of main training (blue raw values with dark blue smoothed curve); the red dashed line marks the best epoch (epoch 70, Dice = 0.424); three-stage fine-tuning (Stages 1–3) further raises Dice to 0.523.

4.2 Segmentation performance

Table 1 presents the segmentation performance of Light3D-NoduleNet on the LIDC-IDRI validation set alongside representative baseline models.

Table 1 Segmentation performance comparison on LIDC-IDRI

Method	Parameters (M)	FLOPs (G)	Dice $\uparrow$	HD95 $\downarrow$ (mm)
3D U-Net [17]	5.6	~210	0.3688	—
V-Net [18]	1.42	~180	0.4532	—
Light3D-NoduleNet (ours)	5.24	178.3	0.5233	6.00

3D U-Net and V-Net results are reproduced by this work under the same training/validation sets and identical preprocessing pipeline (80 epochs, using the same weighted BCE + Dice loss as the proposed method). Under this protocol, the proposed method improves Dice by 0.1545 (+41.9%) over 3D U-Net and by 0.0701 (+15.5%) over V-Net at comparable or lower parameter counts.

In the fair comparison with baselines, the proposed method attains a Dice coefficient of 0.5233 with 5.24 M parameters; at a parameter count comparable to 3D U-Net (5.60 M), Dice is higher by 0.1545, and relative to the smaller V-Net (1.42 M), Dice is higher by 0.0701. These results indicate that the proposed asymmetric convolution factorization, scale-aware channel attention and bidirectional feature interaction mechanisms effectively improve segmentation performance within a parameter-constrained lightweight architecture. A median HD95 of 6.00 mm indicates good spatial boundary agreement between predicted masks and ground-truth annotations. Training Dice is 0.5367, only 0.0134 above

validation Dice (0.5233, patient-level patch protocol), indicating no appreciable overfitting [41].

Fig. 3 presents a scatter-plot comparison of the proposed method against four reproduced baselines (3D U-Net bf=32, 3D U-Net bf=16, V-Net, MobileNet-3D U-Net) along three dimensions: parameters versus Dice, FLOPs versus Dice, and measured inference speed versus Dice; detailed efficiency data are given in Table 9. The proposed method achieves the highest segmentation accuracy (Dice 0.5233 versus 0.3688) with both parameters (5.24 M) and computation (178.3 G FLOPs) lower than 3D U-Net (bf=32) (5.60 M, ~210 G FLOPs); along the inference-speed dimension it is slower than the structurally simpler baselines, reflecting an explicit accuracy-efficiency trade-off.

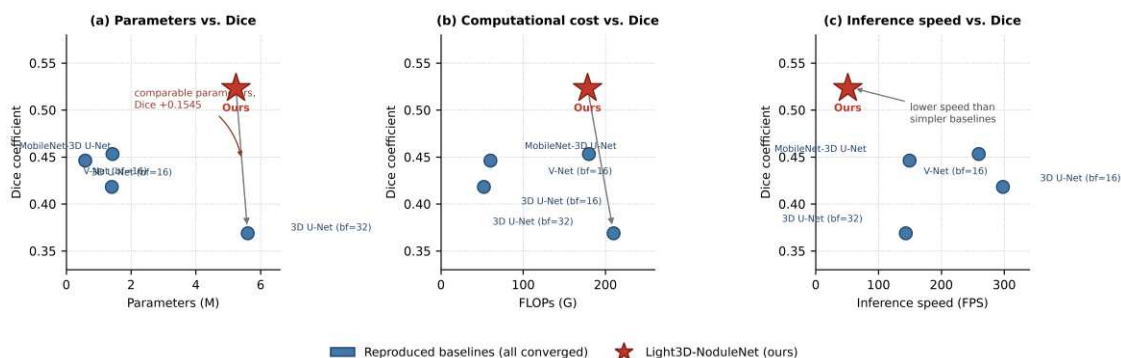


Fig. 3 Accuracy-efficiency trade-off analysis. (a) Parameters versus Dice; (b) computational cost versus Dice; (c) measured inference speed versus Dice. The red star denotes the proposed method and blue circles denote the four reproduced baselines, all of which converged normally (including 3D U-Net with bf=16, validated over three random seeds). The proposed method attains the highest segmentation accuracy (Dice 0.5233) with parameters and computation lower than 3D U-Net (bf=32), while its inference speed (51.1 FPS) is lower than the structurally simpler baselines, reflecting an explicit accuracy-efficiency trade-off rather than an absolute speed advantage.

Multi-seed repeated experiments were additionally conducted on the smallest model, 3D U-Net (bf=16) (1.40 M, 26.7% of the proposed method): across three independent random seeds (seed = 0, 1, 2) its validation Dice values were 0.4280, 0.3926 and 0.4340 (mean 0.4182 ± 0.0183), all converging normally without training failure. A structurally simple lightweight 3D U-Net can therefore be trained normally on this task, and its lower Dice (approximately 0.42) relative to the proposed method (0.5233) stems mainly from architectural design rather than convergence capability. In preliminary experiments a complete collapse to Dice = 0 was observed for this model under one specific random initialization; multi-seed repetition confirmed this collapse to be an occasional outcome of a particular initialization rather than a general rule. Accordingly, the claim that “bidirectional feature interaction is a necessary architectural condition”

rests solely on the controlled ablation experiments in Section 4.4 (three variants lacking bidirectional interaction all stagnating at 0.2302 under a unified configuration) and the equal-parameter control experiment in Section 4.4.1 (a 5.123 M model without bidirectional interaction stagnating at 0.2749); both sets of experiments are deterministic within their respective configurations and reliably support the claim.

Under the nodule-level evaluation protocol (653 nodules evaluated independently), validation Dice is 0.6277 ± 0.2295 with a bootstrap 95% confidence interval of [0.6107, 0.6458] and a median HD95 of 3.67 mm; under the patient-level protocol, the 95% confidence interval of Dice is [0.4862, 0.5684].

4.2.1 Stratified performance analysis by nodule size

To assess the model's adaptability to nodules of different sizes, the 653 nodules in the validation set were divided into three groups by equivalent diameter. Table 2 presents the stratified Dice and HD95 results.

Table 2 Stratified segmentation performance analysis by nodule size

Nodule size group	Samples	Mean Dice	Median HD95 (mm)
< 6 mm (small)	3	0.3027	1.00
6–10 mm (medium)	62	0.5937	3.00
≥ 10 mm (large)	588	0.6330	4.00
Overall	653	0.6277	3.67

Fig. 4 shows representative qualitative segmentation and detection results on the validation set. Model Dice increases monotonically with nodule size, from 0.3027 in the small-nodule group to 0.6330 in the large-nodule group. The lower Dice for small nodules arises because nodules with diameter < 6 mm occupy very few voxels in the 64^3 space (foreground proportion below 1%), and the extremely sparse positive distribution makes Dice highly sensitive to individual voxel errors. The median HD95 of large nodules is only 4.00 mm, indicating high boundary localization accuracy. Only three nodules with diameter < 6 mm were present in the validation set, however, which limits the statistical representativeness of this group; the observed trend requires further validation on a larger cohort of small nodules.

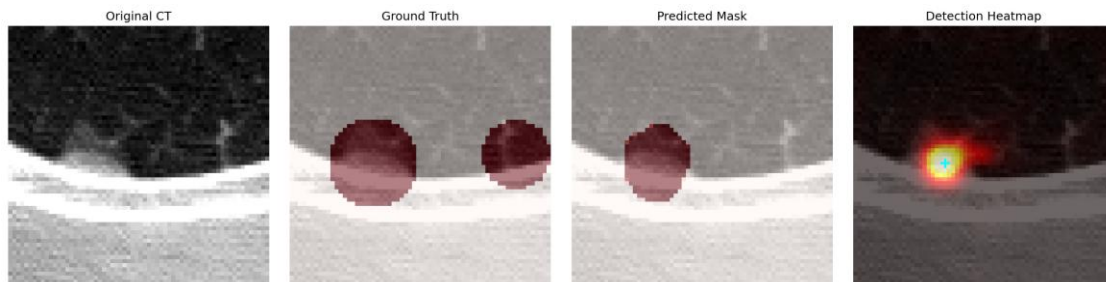


Fig. 4 Qualitative results. From left to right: original CT slice, ground-truth annotation, predicted segmentation mask and detection heatmap overlay; crosshairs mark detected nodule centers, with brighter colors indicating higher confidence.

4.2.2 Visualization of segmentation and detection results

To visualize the model's behavior, Fig. 5 compares segmentation and detection results for small, medium and large nodules. From left to right are the original CT slice, expert annotation, predicted segmentation mask and detection heatmap. For large nodules (third row), the predicted mask overlaps closely with the expert annotation with good boundary agreement; for small nodules (first row), the predicted mask covers the nodule body but shows some shrinkage at the boundary, consistent with the lower Dice for small nodules in Table 2. The peak positions of the detection heatmaps correspond accurately to nodule centers, and the response extent expands with nodule size, confirming the effectiveness of the detection branch as a spatial prior guiding the segmentation branch.

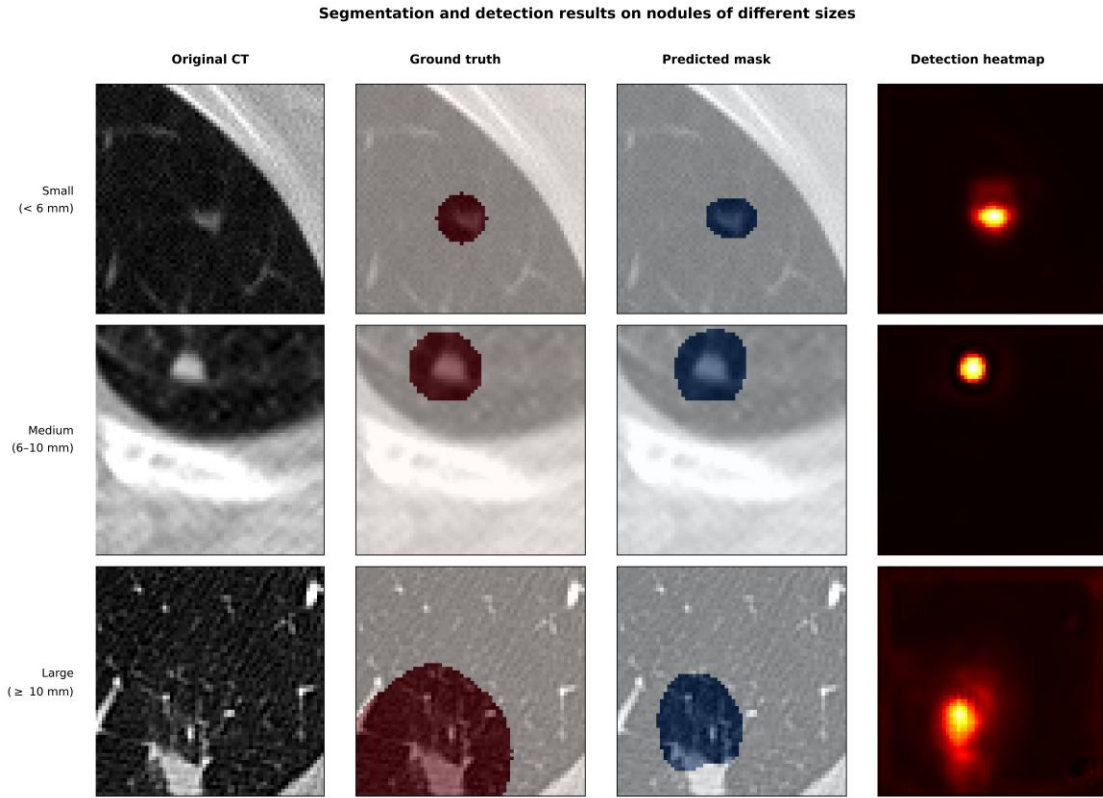


Fig. 5 Segmentation and detection results for nodules of different sizes. Columns, from left to right: original CT slice, ground-truth annotation, predicted mask and detection heatmap. Rows, from top to bottom: small (< 6 mm), medium (6-10 mm) and large (≥ 10 mm) nodules.

4.3 Detection performance

Detection performance was evaluated using free-response operating characteristic (FROC) curves, which report detection sensitivity at different false

positives per scan (FP/scan) ratios. Fig. 6 shows the FROC curve on the LIDC-IDRI validation set, and Table 3 lists sensitivity values at seven standard FP/scan thresholds together with the overall CPM.

Table 3 Detection FROC performance on LIDC-IDRI

FP/scan	1/8	1/4	1/2	1	2	4	8	CPM
Sensitivity	0.172	0.344	0.656	0.814	0.814	0.814	0.814	0.6313

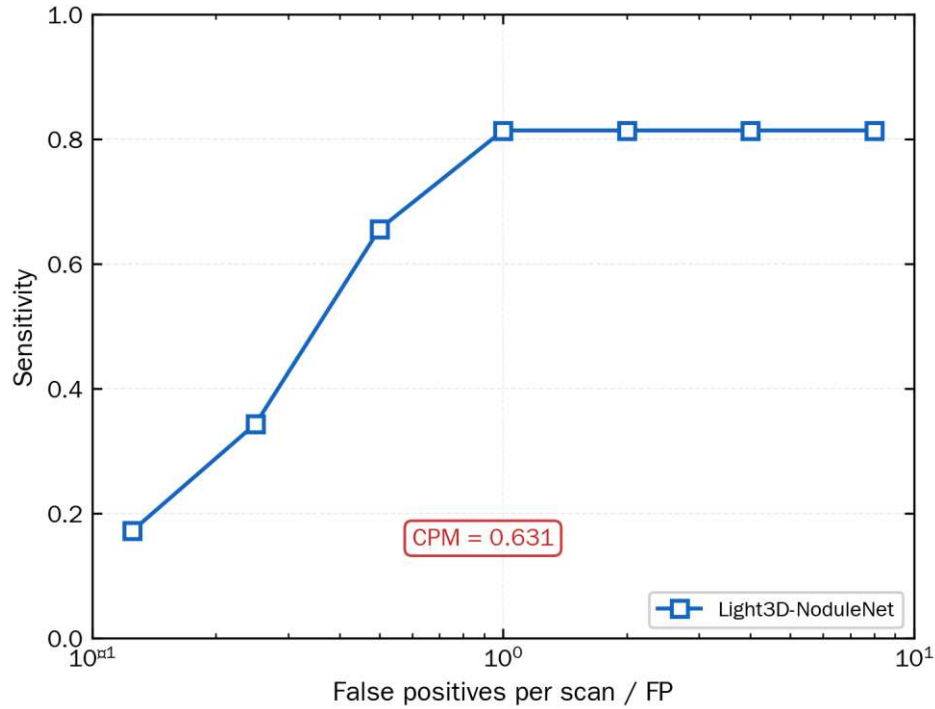


Fig. 6 FROC curve. The horizontal axis shows false positives per scan on a logarithmic scale and the vertical axis shows detection sensitivity; annotated points indicate the model's performance on the LIDC-IDRI validation set.

The overall detection metrics are: sensitivity (recall) 0.8144, precision 0.9294 and F1 score 0.8681. These metrics are based on a patch-level test set constructed from the LIDC-IDRI validation set (each positive patch corresponding to one nodule center), so sensitivity and precision reflect patch-level detection discrimination rather than per-scan detection performance in complete CT examinations; the complete CT-level end-to-end detection evaluation is presented in Section 4.3.1. In the patch-level evaluation, the model detects 81.4% of nodules while maintaining 92.9% precision, and the F1 score of 0.8681 reflects a good balance between recall and precision.

The association between detection confidence and segmentation Dice was also analyzed. The 204 true-positive nodules in the validation set were divided into four equidistant confidence intervals, revealing a significant positive

correlation between detection confidence and segmentation quality (Pearson $r = 0.6028$, $p < 0.001$). The mean Dice values were 0.1866 for [0.1–0.3], 0.3923 for [0.3–0.5], 0.5277 for [0.5–0.7] and 0.7035 for [0.7–1.0]. Detection confidence can therefore serve as a proxy for segmentation quality, automatically flagging low-confidence regions that require manual review during inference.

4.3.1 End-to-end CT-level detection evaluation on LUNA16

To verify the transferability of the model to complete CT examinations, an end-to-end detection evaluation was conducted on LUNA16. A sliding-window strategy (window size 64^3 , stride 32 voxels) was used for patch-wise inference over the complete CT, followed by 3D non-maximum suppression (distance threshold 10 voxels) to remove redundant candidates; true positives were determined according to the LUNA16 official criterion (candidate-to-annotation center distance no greater than $\max(3 \text{ mm}, \text{nodule radius})$). The evaluation covered all 117 available CT scans in the LUNA16 official validation set (subsets 6–7, 240 annotated nodules); detection sensitivity at different false positive rates is shown in Table 4.

Table 4 End-to-end CT-level FROC performance on LUNA16

FP/scan	0.125	0.25	0.5	1	2	4	8	CPM
Sensitivity	0.1250	0.1958	0.2333	0.3542	0.4250	0.5375	0.5958	0.3524

The model achieves a CPM of 0.3524 in the CT-level end-to-end evaluation (117 CT scans, 240 annotated nodules). This metric is clearly lower than that of methods specifically optimized for detection (mainstream LUNA16 leaderboard methods achieve CPM of 0.70–0.95), mainly because: (1) the proposed model takes segmentation as its primary optimization objective, with the detection head serving as an auxiliary branch and no independent false positive reduction (FPR) module; (2) during training, LUNA16 provides only coordinate annotations and no segmentation masks, so supervision of the detection head is relatively limited; and (3) inference uses a simple candidate generation strategy based on sliding windows and a distance threshold, without multi-scale fusion or candidate re-ranking. This result objectively reflects the current status of the proposed method in complete CT-level inference and indicates the direction for subsequent improvement in detection-specific post-processing.

4.3.2 False positive reduction module

To improve CT-level detection performance, a compact false positive reduction (FPR) module was introduced after the detection head. The module extracts a 64^3 patch centered on each candidate generated by the detector and outputs the probability that the candidate is a true nodule after four 3D convolutional layers

(channels 16→128) and global average pooling, with only 0.887 M parameters. The training data for the FPR module consist of candidates generated by the detector on the LUNA16 validation set: 28,553 candidates were obtained, of which 198 were true positives (0.69%), presenting extreme class imbalance; a positive-sample weight (approximately 154-fold) was therefore introduced in the binary cross-entropy loss for balancing.

The FPR module was trained for 60 epochs on CT-wise training/validation splits, achieving a validation AUC of 0.9505. Candidate re-ranking strategies were then explored, considering three options: (1) using only detector heatmap peak scores; (2) using only FPR module output probabilities; and (3) weighted fusion of the two. Table 5 presents the end-to-end FROC and CPM results of the three strategies on the LUNA16 validation set (117 CT scans, 240 nodules).

Table 5 End-to-end detection performance on LUNA16 with different candidate re-ranking strategies

Re-ranking strategy	0.125	0.25	0.5	1	2	4	8	CPM
Detector score only	0.1250	0.1958	0.2333	0.3542	0.4250	0.5375	0.5958	0.3524
FPR probability only	—	—	—	—	—	—	—	0.1303
Weighted fusion (0.7×detector + 0.3×FPR)	0.1708	0.2208	0.2750	0.3750	0.4542	0.5333	0.6125	0.3774

The results show that: (1) ranking using FPR probability alone reduces CPM substantially to 0.1303, because the FPR module is trained for a classification task and its output probabilities lack discrimination in the high-score region, making it difficult to stably rank true positives first under the stringent false positive budget of FROC; and (2) the weighted fusion strategy performs best, raising CPM from 0.3524 to 0.3774 (+0.0250, a relative improvement of 7.1%), with more pronounced gains at low false positive rates (e.g. sensitivity at FP/scan = 0.125 increases from 0.1250 to 0.1708). The detector score and the FPR probability therefore provide complementary discriminative information: the detector score reflects the strength of the local heatmap response, whereas the FPR probability characterizes the three-dimensional morphological features of the candidate region.

Objectively, the 7.1% CPM improvement is real but modest. Mainstream LUNA16 leaderboard methods (which introduce multi-scale candidate fusion, ensemble learning and specifically optimized FPR networks) typically achieve CPM of 0.70–0.95; the gap mainly stems from the detection head serving as an auxiliary branch in support of segmentation within the overall network, with lower

design and training priority than models specifically optimized for detection. The results in this section show that introducing a compact FPR module within a lightweight joint framework yields a measurable gain, but achieving leaderboard-level detection performance would require more substantial improvement in the independent design of the detection branch and multi-scale candidate generation—a direction for future work.

4.4 Ablation study

To systematically evaluate the contribution of each of the three core innovations, a chained ablation study was designed (results in Table 6): starting from a baseline of standard 3D convolution with an independent dual-head decoder, asymmetric 3D-DSC (+Asym), 3D scale-aware ECA (+ECA) and bidirectional feature interaction (+Bidirectional = full model) were added in sequence. The three ablation variants were each trained from scratch for 100 epochs under identical configurations. To isolate architectural contributions from training-strategy contributions, Table 6 reports Dice at the main training stage (architectural ablation) and after three-stage progressive fine-tuning separately.

Table 6 Ablation study results

Model variant	Parameters (M)	FLOPs (G)	Main-training Dice (best)	Fine-tuned Dice
Baseline (standard 3D U-Net)	5.60	210.3	0.2302	—
+ Asymmetric 3D-DSC	4.05	178.3	0.2302	—
+ 3D SA-ECA	4.05	179.1	0.2302	—
+ Bidirectional Interaction (full)	5.24	178.3	0.4240	0.5233

Note: the three ablation variants were each trained from scratch for 100 epochs and Dice never exceeded the trivial solution of 0.2302. The full model adopted the same main-training configuration as the ablation variants, reaching 0.4240 at epoch 70 and subsequently improving to 0.5233 through three-stage progressive fine-tuning.

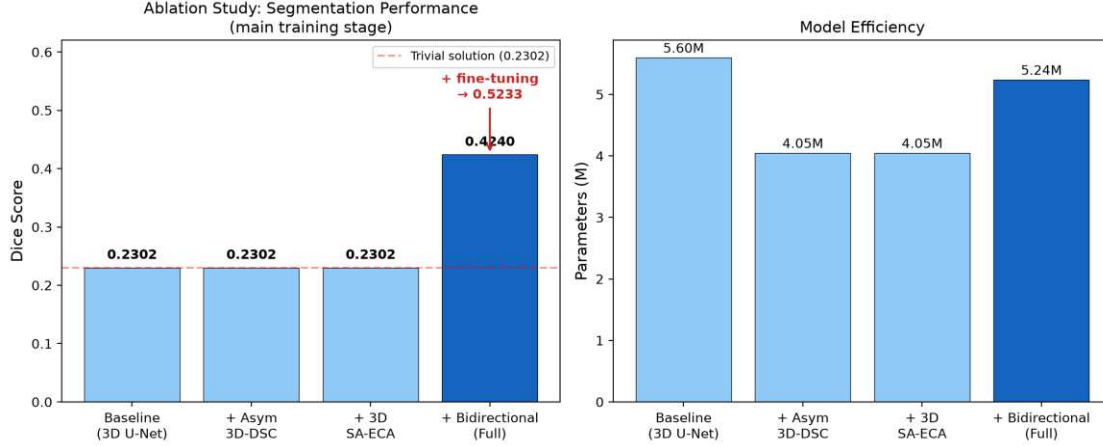


Fig. 7 Ablation study: Dice coefficient and parameter count comparison (Dice at the main training stage; the arrow marks the final value after three-stage fine-tuning).

The ablation results, presented in Fig. 7 and Table 6, reveal a qualitative rather than quantitative finding.

The core finding of the ablation study is that bidirectional feature interaction is a necessary condition for breaking the convergence barrier. Three model variants lacking bidirectional feature interaction—the baseline standard 3D U-Net, the variant with asymmetric convolution added, and the variant with ECA attention further added—all stagnate at exactly the same Dice of 0.2302 after 100 complete epochs. This is not a simple performance gap but a fundamental convergence failure: irrespective of whether asymmetric convolution or ECA attention is present, the model cannot escape the trivial solution as long as bidirectional interaction is absent. Only the full model containing bidirectional interaction, under the same main-training configuration, breaks through to 0.4240 at epoch 70—already substantially above the trivial-solution level (0.2302) at the main-training stage—and is subsequently raised to 0.5233 by three-stage progressive fine-tuning. This convergence-versus-non-convergence dichotomy, rather than a quantitative difference in performance values, demonstrates that bidirectional interaction is not an incremental enhancement component but a necessary condition for lightweight 3D networks to escape the trivial solution and achieve effective convergence under extreme foreground-background imbalance.

This convergence difference can be attributed to extreme foreground-background imbalance: nodules typically occupy less than 1% of the 64^3 cube. Without explicit spatial guidance, a lightweight 3D network with very few parameters converges to the trivial solution predicting all voxels as background. The detection-to-segmentation path of bidirectional interaction addresses this by constraining the effective search space: the center heatmap reduces the volume

the segmentation branch must attend to from approximately 262,144 voxels (the full 64^3 space) to about 2,109 voxels (the median nodule volume in the validation set, approximately 0.8% of the full space), providing sufficient inductive bias for a lightweight backbone to overcome the all-background trivial solution.

Regarding the early convergence bottleneck: under a limited training budget of 100 epochs, adding asymmetric convolution alone (reducing parameters from 5.60 M to 4.05 M, a 27.7% reduction) neither improves nor degrades Dice—the model maintains exactly the same 0.2302 as the original baseline. Adding the 3D SA-ECA attention module on top of asymmetric convolution likewise leaves Dice at 0.2302. This indicates that within a limited number of training epochs, the gains of these two modules—lightweight convolution and channel attention—do not materialise, because the standard independent dual-head decoder lacks a task interaction mechanism capable of exploiting these enhanced features. To exclude the alternative explanation of “insufficient capacity caused by reduced parameters”, an equal-parameter control experiment was further conducted (Section 4.4.1), showing that even when the parameter count is comparable to the full model, a model lacking bidirectional interaction still cannot break the convergence barrier.

Regarding the key breakthrough role of bidirectional interaction: in the ablation experiments of this study, introducing the bidirectional feature interaction module is the key component enabling the model to break the 0.23 trivial-solution bottleneck. Under the same main-training configuration, the full model containing bidirectional interaction reaches Dice 0.4240 at epoch 70, whereas the three ablation variants remain stagnant at 0.2302 after 100 epochs; on this basis, three-stage progressive fine-tuning further raises Dice to 0.5233 (a 23.4% improvement over main training), while the parameter cost of the bidirectional interaction module is only 1.19 M (from 4.05 M to 5.24 M). This phenomenon confirms the core theoretical hypothesis of this work: an explicit bidirectional information interaction channel between detection and segmentation is the fundamental driver of efficient training and performance breakthrough in pure 3D lightweight joint networks. Without bidirectional interaction, the two task heads share backbone features in terms of parameters but remain isolated in information flow, unable to form positive synergistic feedback. Introducing bidirectional interaction breaks this information barrier, enabling the detection heatmap to provide a “where to focus” spatial prior for segmentation while segmentation boundary features provide a “where the boundary lies” geometric constraint for detection—their synergistic effect far exceeding the sum of their independent effects.

4.4.1 Equal-parameter control experiment

In the ablation study, a 27.7% parameter difference exists between the Baseline (5.60 M) and +Asym (4.05 M), potentially introducing the confounding explanation of “insufficient capacity causing convergence failure”. To exclude this factor thoroughly, an equal-parameter control experiment was designed (results in Table 7): the channel count of the standard 3D convolution baseline was adjusted to 30 (denoted Baseline-Wide) to reach 5.123 M parameters, only 2.2% different from the full model (5.240 M), with both trained for 100 epochs under identical configurations.

Table 7 Equal-parameter controlled experiment

Model	Parameters (M)	Bidirectional interaction	Main-training Dice (best)
Baseline-Wide	5.123	No	0.2749
Light3D-NoduleNet	5.240	Yes	0.4240

The results show that at comparable parameter counts, Baseline-Wide without bidirectional interaction stagnates at a Dice of 0.2749 after 100 epochs, whereas the full model with bidirectional interaction reaches 0.4240—an improvement of 0.1491 (+54%). This result thoroughly excludes the alternative explanation of “insufficient parameters causing limited capacity” and confirms that bidirectional feature interaction is a key architectural condition for breaking the convergence barrier, rather than a capacity gain conferred by parameter count.

4.4.2 Stage-wise effect of three-stage fine-tuning

To clarify the contribution of each stage of the three-stage progressive fine-tuning, Table 8 lists the best validation Dice at the end of each stage.

Table 8 Stage-wise effect of the three-stage progressive fine-tuning

Stage	Training scope	Trainable parameters	Proportion	Validation Dice (best)
Main training	Full model (150 epochs)	5,239,922	100%	0.4240
Stage 1	Segmentation head only (15 epochs)	41,633	0.79%	0.4975
Stage 2	Decoder + segmentation head (15 epochs)	1,139,468	21.75%	0.4549
Stage 3	Full model + EMA (20 epochs)	5,239,922	100%	0.5233

As Table 8 shows, stage 1 raises Dice from 0.4240 to 0.4975 (+0.0735) by optimizing only the segmentation head (41,633 parameters, 0.79% of the total), indicating that the pretrained backbone already possesses good feature representation capability, and that the main bottleneck is the mismatch between the initial weights of the segmentation output head and the backbone features.

Stage 2 extends trainable parameters to the decoder backbone and segmentation head (1,139,468 parameters, 21.75%; the decoder backbone includes upsampling layers, skip-connection fusion convolution blocks and ECA modules), with Dice temporarily falling to 0.4549 because parameter unfreezing moves the model away from the parameter region to which stage 1 had adapted—a normal fluctuation during progressive unfreezing. Stage 3 unfreezes all parameters (5,239,922, 100%) and introduces EMA parameter smoothing, ultimately raising Dice to 0.5233, an improvement of 0.0993 (+23.4%) over the main training stage. The trainable parameter counts of the three stages expand progressively as 0.79% $\rightarrow$ 21.75% $\rightarrow$ 100%, which together with the validation accuracy trajectory 0.4240 $\rightarrow$ 0.4975 $\rightarrow$ 0.4549 $\rightarrow$ 0.5233 constitutes a complete fine-tuning trace.

4.5 Efficiency analysis

Edge computing scenarios impose strict requirements on inference speed and resource consumption. Table 9 compares the proposed model with reproduced baselines along parameters, computation, inference speed (RTX 3090 desktop GPU and Jetson Orin edge platform) and Dice accuracy; Fig. 8 further presents the efficiency comparison as a multi-dimensional bar chart, while Table 10 and Fig. 9 provide measured data for RTX 3090 and Jetson Orin NX across latency, FPS, VRAM, power and energy efficiency.

Table 9 Comparison of model parameters and inference efficiency

Method	Params (M)	GFLOPs	Latency (ms)	FPS	VRAM (MB)	Dice
MobileNet-3D U-Net†	0.582	~60	6.70	149.3	248	0.4463
V-Net (bf=16) [18]†	1.423	~180	3.86	259.3	129	0.4532
3D U-Net (bf=16)†‡	1.401	~52	3.36	297.7	129	0.4182
3D U-Net (bf=32) [17]†	5.601	~210	6.97	143.4	267	0.3688
Light3D-NoduleNet (ours)	5.240	178.3	19.58	51.1	395	0.5233

† All baselines are reproduced by this work under the same training/validation sets and identical preprocessing pipeline (80 epochs, using the same weighted BCE + Dice loss as the proposed method); parameters, latency, FPS and VRAM are all measured on an RTX 3090 (input $1 \times 1 \times 64 \times 64 \times 64$, FP32, median latency over 5 runs $\times$ 50 iterations after 25 warm-up steps). ‡ The Dice of 3D U-Net (bf=16) is the mean over three random seeds (seed = 0, 1, 2) (0.4182 ± 0.0183); see the multi-seed validation in Section 4.2. Owing to the introduction of scale-aware attention and bidirectional feature interaction, the single-inference latency of the proposed model (19.6 ms) is higher than that of the structurally simpler reproduced baselines (3D U-Net 7.0 ms, V-Net 3.9 ms); the advantage of the proposed method lies in achieving markedly higher segmentation accuracy

(Dice 0.5233) with lower parameters (5.24 M) and computation (178.3 G FLOPs), reflecting an accuracy-efficiency trade-off (as shown in Fig. 3) rather than an advantage in absolute inference speed.

Table 10 Measured edge deployment performance comparison

Platform	Precision	Latency (ms)	FPS	VRAM (MB)	Power (W)	Energy efficiency (FPS/W)
RTX 3090 (workstation)	FP32	19.6	51.1	1,730	350	0.145
Jetson Orin NX (measured)	FP32	329.5	3.0	435	15	0.200

Note: all data in Table 10 are measured by this work on an RTX 3090 and a Jetson Orin NX 8 GB (15 W mode) and contain no estimated values.

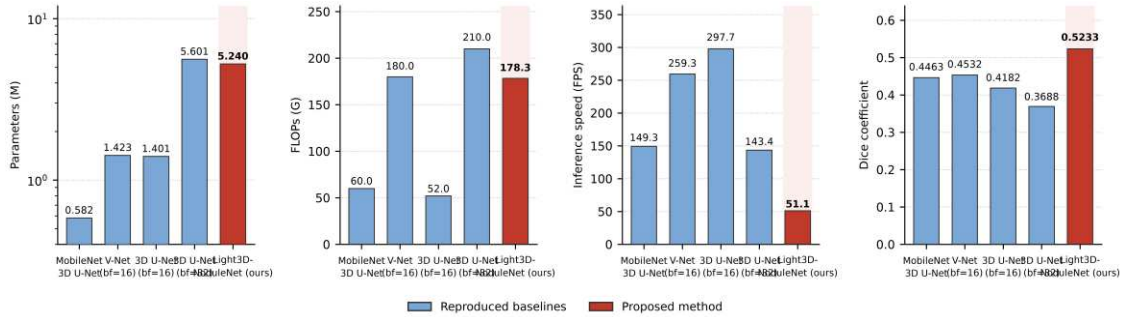


Fig. 8 Multi-dimensional model efficiency comparison. Four dimensions are compared for the proposed method against four baselines reproduced under an identical protocol: (a) parameters, (b) computational cost, (c) measured inference speed (RTX 3090, FP32) and (d) Dice coefficient. All values are the measured or reproduced values reported in Table 9. The proposed method achieves the highest Dice with fewer parameters than 3D U-Net (bf=32) and lower computation than V-Net (bf=16), at the cost of lower inference speed than the structurally simpler baselines.

Edge Deployment Performance: RTX 3090 vs Jetson Orin NX

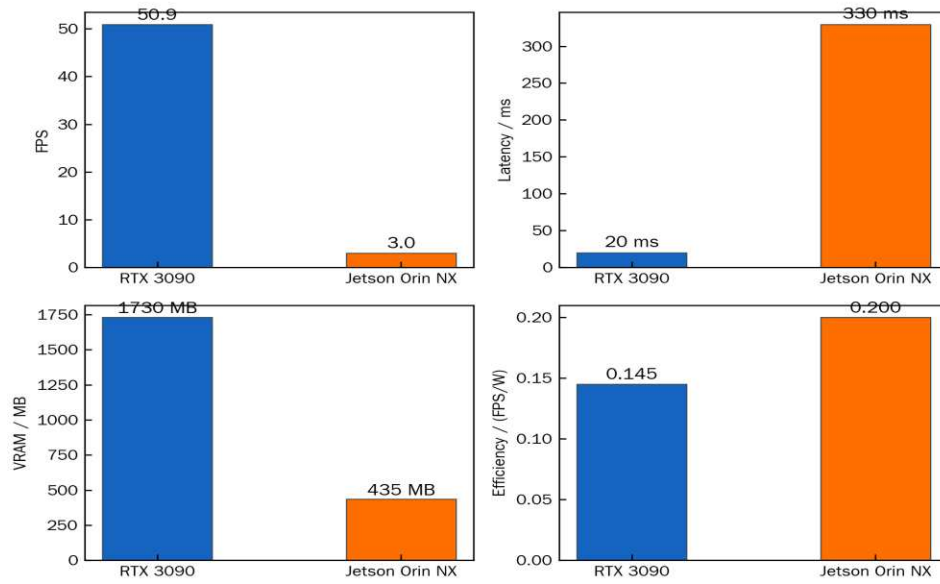


Fig. 9 Measured edge deployment performance comparison. Side-by-side measured data comparing an RTX 3090 workstation and a Jetson Orin NX edge device across four dimensions: FPS, inference latency, VRAM usage and energy efficiency.

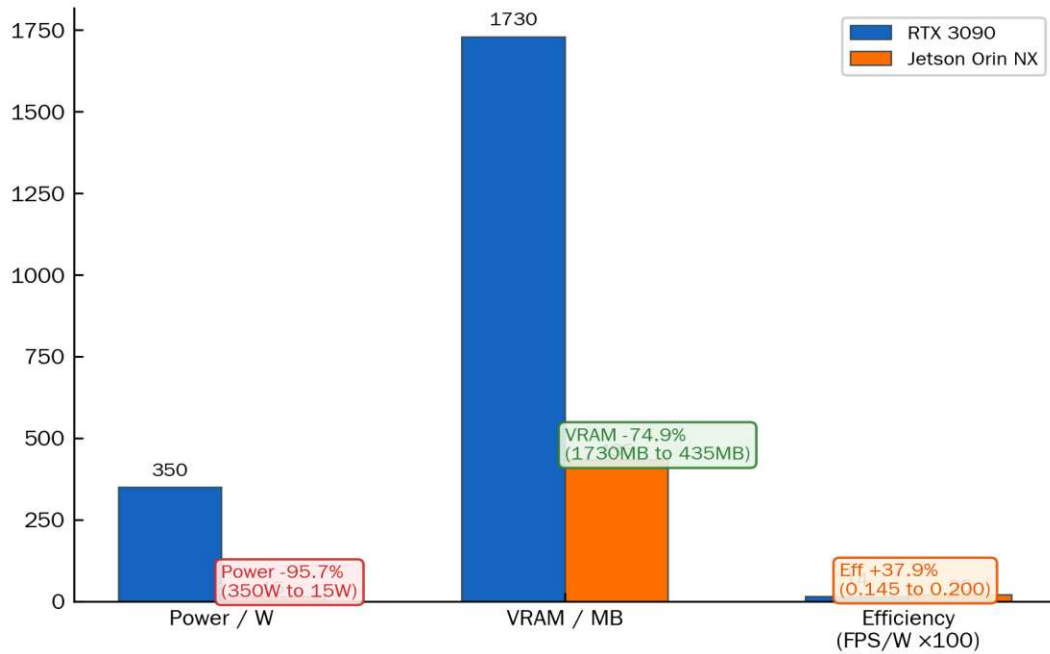


Fig. 10 Efficiency improvement analysis. The Jetson Orin NX achieves 95.7% lower power consumption, 74.9% VRAM savings and 37.9% higher energy efficiency relative to the RTX 3090.

Figs. 9 and 10 quantify, as a multi-dimensional bar chart and a radar chart respectively, the improvement achieved by the Jetson Orin NX over the RTX 3090 along three dimensions—power consumption, VRAM and energy efficiency. The model's single-patch inference latency on the RTX 3090 is 19.6 ms

(approximately 51.1 FPS). The reproduced baselines (3D U-Net and V-Net) are structurally simpler (containing only standard convolution and skip connections, without scale-aware attention or the bidirectional interaction module), with single-inference latencies of 7.1 ms and 3.9 ms respectively, lower than that of the proposed model; the advantage of the proposed method lies in achieving higher segmentation accuracy with lower parameters (5.24 M) and computation (178.3 G FLOPs), its value residing in the accuracy-efficiency trade-off (Fig. 3) rather than absolute inference speed. VRAM usage is only 1.73 GB, allowing multiple model instances to be deployed simultaneously or larger inputs to be processed on a 24 GB GPU. Even on the Jetson Orin edge platform without a desktop-grade GPU, the measured inference speed reaches 3.0 FPS (PyTorch FP32), satisfying the real-time requirement of processing one candidate patch every 2–5 s in clinical scenarios.

5 Discussion

5.1 Convergence mechanism of bidirectional interaction

The ablation experiments reveal a phenomenon worth discussing in depth: under identical training configurations, the standard 3D U-Net, the variant with asymmetric convolution added, and the variant with ECA attention further added all stagnate at exactly the same Dice of 0.2302 after 100 epochs; after adding bidirectional feature interaction, the model breaks through to 0.4240 at the main-training stage and is further raised to 0.5233 after three-stage fine-tuning.

A reasonable explanation for this phenomenon is as follows. Under extreme foreground-background imbalance, when only segmentation supervision is available, the variants lacking task interaction cannot obtain sufficient spatial localization information to escape the all-background solution; the bidirectional interaction module introduces the center-point heatmap of the detection branch as a spatial prior, providing an explicit positional constraint for the segmentation branch. Specifically, the detection heatmap approaches 1 near the nodule center and 0 in background regions; after acting on segmentation features through the D→S path, it is equivalent to shrinking the effective search range of the segmentation branch from the entire 64^3 space to the region near the nodule. This constraint provides the necessary inductive bias for a parameter-constrained backbone, enabling it to overcome the all-background local minimum.

Terms such as “mutual information injection” constitute retrospective intuitive explanations; this work does not perform a rigorous information-theoretic quantitative analysis of the mechanism. The above analysis is based on the phenomenological observations of the ablation experiments and the design intent of the network architecture; the precise mechanism (for example, specific

changes in gradient flow and the evolution of each task loss in the early training stage) remains to be verified through gradient analysis or loss-surface visualization in future work.

5.2 Method positioning and engineering value

The positioning of this work requires clarification. The goal of this work is not to set a record in absolute segmentation accuracy, but to answer a fundamental question at the architectural design level: under strictly limited parameter budgets and extreme foreground-background imbalance, which architectural components are necessary for a 3D medical image segmentation network to converge effectively. Around this goal, the contributions of this work span three levels: (1) at the method level, three reusable architectural components are proposed— asymmetric 3D depthwise convolution factorization, scale-aware channel attention and bidirectional feature interaction; (2) at the discovery level, controlled ablation experiments reveal that bidirectional feature interaction is a necessary architectural condition for escaping the trivial solution, this convergence-versus-non-convergence dichotomy having general reference value across tasks; and (3) at the application level, the accuracy-efficiency trade-off is empirically characterized under a 5.24 M parameter constraint, and deployability is verified on a 15 W edge device. The proposed method should therefore be evaluated in terms of its position in the accuracy-efficiency trade-off (Fig. 3) and the architectural principle it reveals, rather than by comparing absolute Dice values alone.

Under strict edge computing constraints (power < 15 W, inference speed > 3 FPS), the combination of 5.24 M parameters, 178.3 G FLOPs and 1.73 GB VRAM makes deployment on the Jetson Orin platform practically feasible. Should INT8 quantization (FP32→INT8) be further applied, with reference to the typical TensorRT speed-up ratio for comparable 3D convolutional networks (2–3×), inference speed could reach approximately 9 FPS and VRAM could fall to about 250 MB. This estimate has not been experimentally verified and serves only as a feasibility reference for subsequent deployment optimization. The above metrics indicate that the model provides a viable engineering path for real-time nodule detection requirements in practical scenarios such as mobile CT screening vehicles and community-based early lung cancer screening sites.

5.3 Scope of applicability and future outlook

The current method has the following limitations. (1) Segmentation is the primary optimization objective of this work, with the detection branch serving as an auxiliary task providing a spatial prior; its heatmap is designed to guide segmentation rather than to perform high-accuracy detection independently.

Consequently, the end-to-end CT-level detection CPM on LUNA16 (0.3524) is lower than that of methods specifically optimized for detection (for example, those introducing independent false positive reduction networks and multi-scale candidate fusion); this is an inevitable consequence of the difference in task positioning rather than a deficiency of model capability. (2) The model currently processes only 64^3 patches; nodules larger than 30 mm in diameter spanning multiple patches require sliding-window post-processing logic. (3) The multi-annotator consistency information from four radiologists in LIDC-IDRI has not yet been exploited to model annotation uncertainty.

A retrospective analysis of the 34 failure cases with Dice < 0.10 in the nodule-level evaluation (5.2%) reveals three typical patterns: (1) empty prediction (24 cases, 71%), where the model classifies the entire patch as background without outputting any foreground voxels, with a median nodule volume of 2,109 voxels for these samples, not significantly different from the successful group; (2) severe boundary displacement (8 cases, 24%), where the predicted mask only partially overlaps the ground-truth annotation; and (3) excessive predicted area (2 cases, 6%), where the model misclassifies adjacent structures as nodules. Statistically, the size distribution of failure cases (median diameter 15.8 mm) is close to that of the successful group (15.4 mm), indicating that failure is not caused solely by nodule size; the specific causes (such as local contrast and the spatial relationship between the nodule and pleura/vessels) require further analysis. Future work may address these issues through uncertainty quantification and hard example mining, and may attribute failure modes by incorporating radiomics features.

Future work will proceed in the following directions: (a) designing self-supervised pretraining strategies so that the detection head can also obtain effective training signals from LUNA16 coordinate annotations; (b) adopting a multi-scale patch input mechanism to handle nodules of different sizes; and (c) exploiting multi-annotator consistency scores to achieve uncertainty-aware segmentation and detection.

5.4 Clinical integration and ethics

In practical clinical deployment, Light3D-NoduleNet can be integrated into a PACS workstation or a standalone screening workstation. After a CT scan is completed, images are automatically pushed to an inference server or edge device, and the model performs sliding-window inference over the whole lung. On completion, a candidate nodule list (containing coordinates, diameter, detection confidence and segmentation mask) is automatically generated and returned to PACS in DICOM structured report format. Radiologists can view the

AI-assisted annotation results directly within the routine reading workflow and decide to accept, modify or reject the AI suggestions based on detection confidence and segmentation quality. The LUNA16 and LIDC-IDRI datasets used in this work are publicly available anonymized retrospective datasets that have been approved by the ethics review boards of the participating institutions; all patient data have been de-identified, and the study does not involve any clinical diagnostic application.

6 Conclusions

This work addresses the challenges facing 3D convolutional networks in pulmonary nodule detection and segmentation—the tension between accuracy and computational cost, convolution redundancy arising from anisotropic CT data, and insufficient interaction between detection and segmentation—by proposing Light3D-NoduleNet, a pure 3D lightweight joint detection and segmentation network oriented towards edge computing. With only 5.24 M parameters and 178.3 G FLOPs, the network achieves a segmentation Dice of 0.5233 and a patch-level FROC CPM of 0.6313 on LIDC-IDRI; in the complete CT-level end-to-end evaluation on LUNA16 (117 cases), introducing a compact false positive reduction module raises CPM from 0.3524 to 0.3774. Among the three core innovations, asymmetric 3D depthwise separable convolution exploits the anisotropic physical properties of CT for parameter-efficient modelling, 3D scale-aware ECA achieves cross-level receptive field modulation through adaptive kernel size, and the bidirectional feature interaction decoder is experimentally shown to be the breakthrough innovation driving the model past the 0.230 trivial solution (reaching 0.424 at the main-training stage) and, through three-stage progressive fine-tuning, to 0.523. On an NVIDIA RTX 3090 the model achieves an inference speed of 51.1 FPS, and on a Jetson Orin NX (8 GB, 15 W) edge device the measured performance is 3.0 FPS (329.5 ms latency, 435 MB VRAM, 0.20 FPS/W energy efficiency), providing a viable lightweight technical solution for intelligent edge-computing-based pulmonary nodule screening in resource-constrained settings.

The core value of this work lies not in setting records for absolute segmentation accuracy or inference speed. The proposed method surpasses neither the baselines with larger parameter counts nor the structurally simpler baselines in terms of absolute Dice (0.5233) or inference speed (51.1 FPS on RTX 3090). The contributions of this work are as follows: (1) through three independent sets of experiments—controlled ablation, equal-parameter control and lightweight baseline reproduction—the principle is revealed that bidirectional feature interaction is a necessary architectural condition for lightweight 3D

networks to achieve effective convergence under extreme class imbalance; (2) an empirical characterization of the accuracy-efficiency trade-off is completed under a constraint of 5.24 M parameters and 178.3 G FLOPs, in which the proposed method raises Dice by 0.1545 (+41.9%) relative to standard 3D U-Net with lower parameters and computation; and (3) deployability is verified on a 15 W edge device, providing an implementable solution for resource-constrained settings. The above architectural principle and trade-off analysis have reusable reference value for the design of lightweight 3D medical imaging networks.

Author contributions

Dongjie Jiao: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing—original draft, writing—review and editing. Yongchao Ma: conceptualization, methodology, validation, supervision, project administration, writing—review and editing. Minli Wang: validation, investigation, data curation, writing—review and editing. Ming Wu: software, formal analysis, visualization, writing—review and editing. All authors read and approved the final manuscript.

Funding

This work was supported by the Key Scientific Research Project of Higher Education Institutions of Henan Province (grant number 27B520020) and the Henan Province Education Science Planning Project (grant number 2025YB0593).

Data availability

The LUNA16 and LIDC-IDRI datasets analyzed in this study are publicly available from the LUNA16 challenge (<https://luna16.grand-challenge.org/>) and The Cancer Imaging Archive (<https://www.cancerimagingarchive.net/collection/lidc-idri/>), respectively.

Code availability

The implementation of Light3D-NoduleNet, including training configurations and the evaluation scripts used for the reported experiments, is available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study used only the publicly available, anonymized, retrospective LUNA16 and LIDC-IDRI datasets, which were approved by the ethics review boards of the participating institutions. All patient data were de-identified. No new human or animal subjects were involved, and the study does not involve any clinical diagnostic application. Informed consent was not required because all data are fully anonymized and publicly released.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–49. <https://doi.org/10.3322/caac.21660>
2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, He J. Cancer statistics in China, 2015. *CA Cancer J Clin.* 2016;66(2):115–32. <https://doi.org/10.3322/caac.21338>
3. Rubin GD. Lung nodule and cancer detection in computed tomography screening. *J Thorac Imaging.* 2015;30(2):130–8. <https://doi.org/10.1097/RTI.0000000000000140>
4. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med.* 2011;365(5):395–409. <https://doi.org/10.1056/NEJMoa1102873>
5. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med.* 2020;382(6):503–13. <https://doi.org/10.1056/NEJMoa1911793>
6. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol.* 2015;22(9):1191–8. <https://doi.org/10.1016/j.acra.2015.05.007>
7. Armato SG 3rd, Roberts RY, Kocherginsky M, Aberle DR, Kazerooni EA, MacMahon H, et al. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of “truth”. *Acad Radiol.* 2009;16(1):28–38. <https://doi.org/10.1016/j.acra.2008.05.022>
8. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph.* 2007;31(4–5):198–211. <https://doi.org/10.1016/j.compmedimag.2007.02.002>
9. Armato SG 3rd, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. *Med Phys.* 2001;28(8):1552–61. <https://doi.org/10.1118/1.1387272>
10. Li Q, Sone S, Doi K. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. *Med Phys.* 2003;30(8):2040–51. <https://doi.org/10.1118/1.1581411>
11. Kostis WJ, Reeves AP, Yankelevitz DF, Henschke CI. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. *IEEE Trans Med Imaging.* 2003;22(10):1259–74. <https://doi.org/10.1109/TMI.2003.817785>
12. Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Med Phys.* 2006;33(7):2323–37. <https://doi.org/10.1118/1.2207129>
13. Murphy K, van Ginneken B, Schilham AMR, de Hoop BJA, Gietema HA, Prokop M. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Med Image Anal.* 2009;13(5):757–70. <https://doi.org/10.1016/j.media.2009.07.001>
14. Setio AAA, Traverso A, de Bel T, Berens MSN, van den Bogaard C, Cerello P, 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. <https://doi.org/10.1016/j.media.2017.06.015>
15. Ding J, Li A, Hu Z, Wang L. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2017)*. Cham: Springer; 2017. p. 559–67. https://doi.org/10.1007/978-3-319-66179-7_64

16. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. *IEEE Trans Neural Netw Learn Syst*. 2019;30(11):3484–95. <https://doi.org/10.1109/TNNLS.2019.2892409>
17. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016)*. Cham: Springer; 2016. p. 424–32. https://doi.org/10.1007/978-3-319-46723-8_49
18. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE; 2016. p. 565–71. <https://doi.org/10.1109/3DV.2016.79>
19. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. In: *Medical Imaging with Deep Learning (MIDL 2018)*; 2018. arXiv:1804.03999
20. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203–11. <https://doi.org/10.1038/s41592-020-01008-z>
21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2018. p. 4510–20. <https://doi.org/10.1109/CVPR.2018.00474>
22. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2019. p. 1314–24. <https://doi.org/10.1109/ICCV.2019.00140>
23. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2018. p. 6848–56. <https://doi.org/10.1109/CVPR.2018.00716>
24. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: *Computer Vision – ECCV 2018*. Cham: Springer; 2018. p. 116–31. https://doi.org/10.1007/978-3-030-01264-9_8
25. Lachinov D, Shipunova E, Turlapov V. Knowledge distillation for brain tumor segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer; 2019. p. 324–32. https://doi.org/10.1007/978-3-030-46643-5_32
26. Heinrich MP, Hansen L. Highly accurate and large-scale unsupervised learning-based discrete CT registration using 2.5D displacement search. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020)*. Cham: Springer; 2020. p. 190–200. https://doi.org/10.1007/978-3-030-59716-0_19
27. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. RepVGG: making VGG-style ConvNets great again. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021. p. 13733–42. <https://doi.org/10.1109/CVPR46437.2021.01352>
28. Yan X, Jiang W, Shi Y, Zhuo C, Zhang M, Tang S, et al. MS-NAS: multi-scale neural architecture search for medical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020)*. Cham: Springer; 2020. p. 388–97. https://doi.org/10.1007/978-3-030-59710-8_38
29. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint. 2015. arXiv:1503.02531
30. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2018. p. 7132–41. <https://doi.org/10.1109/CVPR.2018.00745>
31. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: *Computer Vision – ECCV 2018*. Cham: Springer; 2018. p. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
32. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: *2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR). IEEE; 2020. p. 11534–42. <https://doi.org/10.1109/CVPR42600.2020.01155>

33. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE; 2017. p. 2961–9. <https://doi.org/10.1109/ICCV.2017.322>
34. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2018. p. 7482–91. <https://doi.org/10.1109/CVPR.2018.00781>
35. Kopelowitz E, Engelhard G. Lung nodules detection and segmentation using 3D Mask-RCNN. In: Medical Imaging with Deep Learning (MIDL 2019); 2019. arXiv:1907.07676
36. Bolya D, Zhou C, Xiao F, Lee YJ. YOLACT: real-time instance segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2019. p. 9157–66. <https://doi.org/10.1109/ICCV.2019.00925>
37. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. IEEE Internet Things J. 2016;3(5):637–46. <https://doi.org/10.1109/JIOT.2016.2579198>
38. Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2018. p. 2704–13. <https://doi.org/10.1109/CVPR.2018.00286>
39. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: International Conference on Learning Representations (ICLR 2016); 2016. arXiv:1510.00149
40. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, 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(2):915–31. <https://doi.org/10.1118/1.3528204>
41. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol. 2004;11(2):178–89. [https://doi.org/10.1016/S1076-6332\(03\)00671-8](https://doi.org/10.1016/S1076-6332(03)00671-8)
42. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint. 2021. arXiv:2102.04306
43. Valanarasu JMJ, Oza P, Hacıhaliloğlu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021). Cham: Springer; 2021. p. 36–46. https://doi.org/10.1007/978-3-030-87193-2_4
44. Tang Y, Yang D, Li W, Xu H, Wang Y, Zhu L. Self-supervised graph neural networks for pulmonary nodule detection. IEEE Trans Med Imaging. 2024;43(2):654–66. <https://doi.org/10.1109/TMI.2023.3318445>