Scientific Reports

https://doi.org/10.1038/s41598-026-47436-3

Article in Press

CA-3DTransUNet with dynamic cross-scale fusion for pulmonary nodule segmentation

Received: 21 October 2025

Accepted: 31 March 2026

Published online: 03 April 2026

Cite this article as: Zhang K., Lan X., Wang Y. *et al.* CA-3DTransUNet with dynamic cross-scale fusion for pulmonary nodule segmentation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-47436-3

Kaikai Zhang, Xiaowen Lan, Yanhui Wang, Lixin Wang, Yuhan Liu & Feng Guo

CA-3DTransUNet with dynamic cross-scale fusion for pulmonary nodule segmentation

Kaikai Zhang1, Xiaowen Lan*1, Yanhui Wang2, Lixin Wang1, Yuhan Liu1, Feng Guo1

1School of Digital and Intelligence Industry, Inner Mongolia University of Science & Technology, Baotou 014010, China. Inner Mongolia, China.

2The First Affiliated Hospital of Baotou Medical College, Department of Gastroenterology, Baotou 014010, China. Inner Mongolia, China.

Correspondence should be addressed to Xiaowen Lan; lxw@stu.imust.edu.cn; vae731@163.com

Abstract

Precise segmentation of pulmonary nodules in low-dose computed tomography is challenged by nodule heterogeneity, low contrast, and spatial overlap with adjacent anatomical structures. To address these issues, we propose CA-3DTransUNet, a segmentation framework based on the 3D-nnUNet architecture. The proposed network incorporates a Transformer 3D module in the bottleneck to model global volumetric dependencies and a CrossEMA3D module in the decoder to dynamically refine spatial features. Additionally, the wavelet transform is applied during the data preprocessing stage to augment input edge details. Evaluations on the LIDC-IDRI, LUNA16, and private BT datasets indicate the model's performance. Specifically, on the LIDC-IDRI dataset, the model achieved a Dice Similarity Coefficient of 91.85±0.43% [95% CI: 91.32–92.38], a Precision of 90.53±0.51%, and a Sensitivity of 93.12±0.42%. These results surpassed the hybrid architecture nnFormer, which attained a Dice score of 89.48±0.52% ($p = 0.014$). These findings suggest that CA-3DTransUNet holds potential for the computer-aided analysis of pulmonary nodules.

Keywords: Volumetric Segmentation, Cross-Scale Interaction, Hybrid Architecture, Computed Tomography

1. Introduction

Lung cancer has emerged as one of the leading causes of cancer-related mortality worldwide[1], making accurate early diagnosis critical for improving patient prognosis and increasing the five-year survival rate. Computed Tomography (CT) serves as the gold standard for pulmonary nodule screening

and is widely adopted in clinical practice.

However, the inherent heterogeneity of pulmonary nodules in morphology, texture, and density poses severe challenges for precise clinical diagnosis. This difficulty is compounded when nodules spatially overlap with adjacent anatomical structures—such as blood vessels and the pleura—which often exhibit highly similar grayscale intensities. Recently, deep learning techniques, particularly 3D CNNs (e.g., 3D-Unet[2], V-Net[3]), have advanced volumetric medical image analysis by directly processing 3D data to reduce spatial feature loss compared to early 2D slice-based methods. Nevertheless, pure CNN architectures are fundamentally restricted by the local receptive fields of their convolutional kernels. This local inductive bias prevents them from fully leveraging global contextual information, making it difficult to differentiate nodules from anatomically similar backgrounds (e.g., vascular adhesion areas), often resulting in over-segmentation or under-segmentation[4].

To enhance global modeling capabilities, recent studies have integrated Transformers into segmentation networks, resulting in hybrid architectures such as TransUNet[5] and UNETR[6]. While these models achieve a better balance between local and global features, they still face a critical feature fusion bottleneck. Most methods predominantly employ simple concatenation or fixed-weight addition in their skip connections, which inadequately addresses the semantic gap between the high-level contextual features from the decoder and the low-level textural features from the encoder[7]. This misalignment introduces feature redundancy and degrades the model's sensitivity to challenging edge details, such as the blurred boundaries of ground-glass nodules (GGNs)[8]. Furthermore, many existing methods lack adaptive mechanisms tailored to nodule heterogeneity, limiting their generalization ability in complex clinical scenarios [9,10].

To address these specific limitations—namely the local inductive bias of CNNs and the semantic gap inherent in standard skip connections—this paper proposes CA-3DTransUNet, a novel volumetric segmentation framework. The core contributions of this work are twofold:

1. Mitigating Local Inductive Bias via Transformer 3D: We introduce a Transformer 3D Feature Enhancement Module at the encoder bottleneck. By leveraging a 3D self-attention mechanism to explicitly model long-range

volumetric dependencies, this module effectively compensates for the CNN's loss of global contextual information, improving the model's capacity to distinguish nodules from surrounding anatomical structures.

2. Bridging the Semantic Gap with CrossEMA3D: To overcome the feature redundancy caused by semantic misalignment in standard skip connections, we propose the CrossEMA3D Module. Utilizing a cross-dimensional attention mechanism (spatial-channel-depth), this module dynamically aligns and adaptively fuses multi-scale features from the encoder and decoder. This strategy effectively suppresses noise and significantly enhances the model's sensitivity to complex boundary details, such as GGN edges and vascular adhesion areas.

2. Related Work

Early approaches to lung nodule segmentation primarily relied on traditional image processing techniques, such as region growing[11] and level-set methods[12]. However, these methods depend heavily on hand-crafted features and struggle to address the inherent challenges posed by heterogeneous nodule densities (e.g., ground-glass opacities) and diverse morphologies. With the advent of deep learning, Fully Convolutional Networks (FCNs) and U-Net[13], with their encoder-decoder architectures and skip connections, have established the mainstream paradigm for medical image segmentation. To address the volumetric spatial continuity of CT data, 3D U-Net[2] and V-Net[3] extended convolutional operations into the volumetric domain, optimizing deep gradient propagation via residual connections. Addressing the minute and complex nature of pulmonary nodules, subsequent studies proposed targeted improvements: Ma et al.[14] introduced attention mechanisms to focus on nodule regions; SKV-Net[15] utilized selective kernel convolutions to adaptively adjust receptive fields; and Lung_PAYNet[16] employed pyramid attention to enhance feature extraction in low-dose CT images. Despite their strong local feature extraction capabilities, pure CNN architectures remain constrained by their limited receptive fields. As noted earlier, this restricts long-range dependency modeling, frequently leading to segmentation errors in anatomically complex regions where nodules share visual characteristics with adjacent tissues.

To overcome the limitations of CNNs in long-range modeling, the

Transformer architecture[17], originally applied in natural language processing, was introduced to medical imaging. The Vision Transformer (ViT)[18] utilizes self-attention mechanisms to directly model global dependencies, catalyzing a surge of research into hybrid CNN-Transformer architectures. TransUNet[19] pioneered the embedding of Transformers into the U-Net bottleneck to capture global context, while UNETR[20] employed a pure Transformer encoder coupled with a CNN decoder for volumetric data processing. To balance computational efficiency and performance, TransBTS[21] and Swin-Unet[22] introduced 3D CNN preprocessing and hierarchical shifted window mechanisms, respectively. Although Li et al.[23] recently achieved high accuracy through dual-attention feature reorganization, existing hybrid models still exhibit significant shortcomings in feature fusion strategies. Most methods rely on simple concatenation or summation to connect the encoder and decoder, ignoring the substantial semantic gap between the high-level semantic features extracted by Transformers and the low-level textural features extracted by CNNs. This mismatch introduces redundant information into skip connections, making precise boundary reconstruction difficult, particularly for ground-glass nodules (GGNs) with blurred edges.

Unlike conventional organ segmentation, pulmonary nodule segmentation faces the unique challenges of "small object" detection and severe class imbalance. Although Res-UNet++[24] and DRS-CNN[25] have attempted to improve boundary sensitivity through deep supervision mechanisms, the robustness of existing models remains insufficient when dealing with vascular adhesion areas that exhibit grayscale intensities highly similar to normal tissue. Furthermore, the scarcity of high-quality 3D annotated data further constrains model generalization capabilities[26]. While weakly supervised methods like WS-LungNet[27] attempt to mitigate label dependency via adversarial learning, fully supervised learning remains the preferred solution in clinical settings where high precision is paramount. However, existing high-performance models are often associated with high computational burdens and rarely optimize for the efficient utilization of features in data-constrained scenarios.

3. Method

3.1 Proposed methodology

This section presents the proposed framework for volumetric pulmonary nodule segmentation. As depicted in Figure 1, the overall workflow comprises a systematic data preparation phase followed by the training and inference of the CA-3DTransUNet.

Given the inherent variability in CT imaging protocols and the heterogeneity of nodule morphology, a standardized data preparation pipeline is established as a prerequisite for robust network training. This phase primarily involves voxel spacing resampling to unify physical resolutions, intensity normalization to standardize Hounsfield Unit (HU) distributions, and the generation of localized patches via center cropping. Furthermore, a wavelet transform is selectively applied to enhance high-frequency textural details, thereby facilitating the network's capacity to discern subtle boundary features against complex backgrounds.

The cornerstone of the proposed methodology is the CA-3DTransUNet, a novel segmentation architecture engineered to synergistically model long-range global dependencies and fine-grained local details. Distinguished from conventional hybrid models, this network incorporates two key innovations: the Transformer 3D Feature Enhancement Module and the CrossEMA3D fusion mechanism (detailed in subsequent sections). These components are explicitly designed to mitigate the risks of overfitting and to address the challenges of segmenting small objects with ambiguous boundaries. Finally, the model's efficacy is rigorously validated on public datasets through quantitative assessment using standard clinical metrics, including the Dice Similarity Coefficient (DSC), sensitivity, and precision.

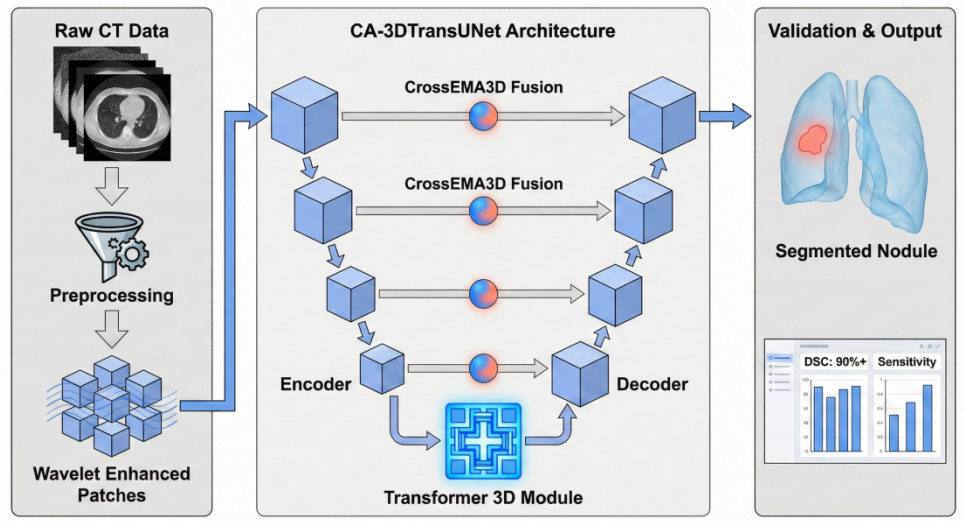

Figure 1 The overall flow chart of the experiment, illustrating the stages from raw CT data preprocessing and wavelet enhancement to CA-3DTransUNet training and clinical metric validation.

3.2 Overall framework overview

The proposed CA-3DTransUNet is designed for the precise volumetric segmentation of pulmonary nodules. Building upon the foundational 3D-nnUNet[28] framework, our architecture integrates advanced mechanisms to collaboratively model local textural details and global contextual dependencies. As detailed in the Network Architecture Diagram (Figure 2), the raw CT volumes first undergo a wavelet transform-based preprocessing step prior to entering the network. This operation decomposes the 3D volumes into multi-scale frequency subbands, enhancing high-frequency details such as nodule edges while suppressing background noise. The standardized input volumes, denoted as $X \in \mathbb{R}^{C\times 64\times 64\times 64}$(where C represents the channel dimension), are then fed into the encoder's initial Dynamic Conv Block, ensuring that the model receives discriminatively enriched feature representations from the onset.

The encoder employs a series of Dynamic Conv Blocks to extract local features, progressively reducing the spatial resolution through successive downsampling operations from 64^3 to 32^3, 16^3, 8^3
, and finally to a bottleneck resolution of 4^3. While Convolutional Neural Networks (CNNs) excel at extracting local features, they are inherently limited in capturing long-range dependencies. To address this limitation, we integrate a Transformer 3D Module at the encoder-decoder bottleneck. By utilizing multi-head self-attention mechanisms, this module explicitly models global spatial dependencies, thereby enhancing the semantic discriminability of the features passed to the subsequent stage. The decoder then reconstructs the spatial resolution through iterative upsampling steps, scaling the features back from 4^3 to the original 64^3 resolution. To mitigate the semantic gap typically found in standard skip connections, we introduce the CrossEMA3D Module, which is strategically embedded to facilitate adaptive cross-scale feature interaction. This module dynamically weighs and fuses global context from the decoder with local details from the encoder, optimizing the spatial representation for accurate boundary delineation.

Finally, the segmentation mask is generated via a concluding convolutional

layer that precisely delineates the nodule's spatial location. The entire network configuration adheres to the adaptive principles of the 3D-nnUNet paradigm. Specifically, the subsampling frequency is determined dynamically based on the smallest dimension of the input volume, and convolutions employ mixed kernel sizes and channel configurations to maximize receptive field efficiency. Furthermore, skip connections dynamically select between concatenation or summation with channel alignment performed via 1×1 convolutions, and Batch Normalization (BN) is automatically applied following all convolutional layers to ensure training stability. By effectively combining the hierarchical feature extraction of CNNs with the global context modeling of Transformers and the adaptive fusion of CrossEMA3D, the CA-3DTransUNet architecture enhances segmentation robustness for pulmonary nodules.

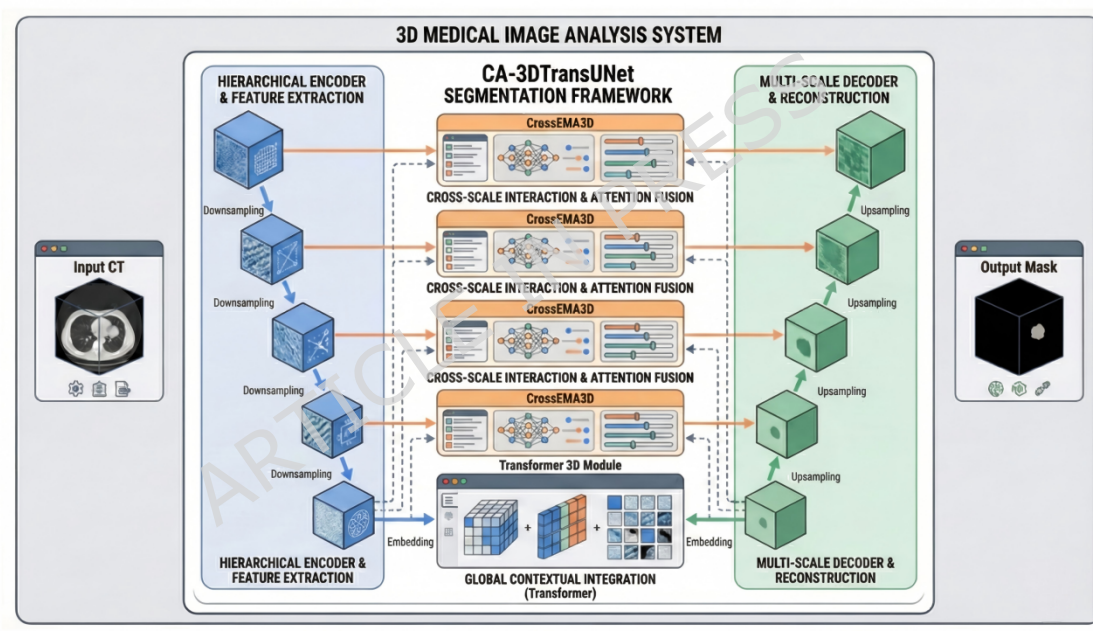

Figure 2 CA-3DTransUNet Network Structure Diagram, highlighting the hierarchical encoder-decoder path with integrated Transformer 3D and CrossEMA3D modules for multi-scale feature fusion.

3.3 Transformer 3D

To capture the explicit long-range volumetric dependencies required to contextualize nodules among surrounding structures, we integrate the Transformer 3D Module at the encoder bottleneck. As illustrated in Figure 3, this module leverages a self-attention mechanism to model global spatial interactions. The mathematical formulation of the module is detailed below.

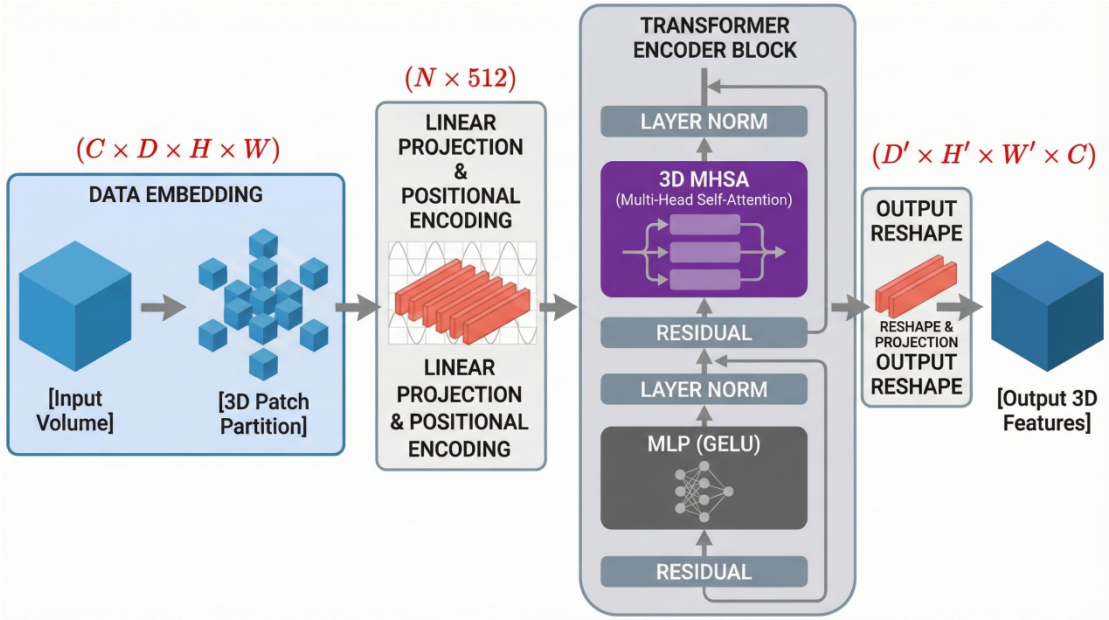

Figure 3 Transformer 3D structure diagram, detailing the 3D patch partitioning, linear embedding, and multi-head self-attention mechanisms used for global volumetric modeling.

3.3.1. 3D Patch Partitioning and Linear Embedding

Given the preprocessed input feature map $V \in \mathbb{R}^{C_{in} \times D \times H \times W}$, where C_{in} denotes the channel dimension and (D, H, W) represent the spatial dimensions, we first serialize the volume into a sequence of patches to satisfy the input requirements of the Transformer. The volume V is partitioned into non-overlapping 3D patches of size $s \times s \times s$ (with $s=4$ in this implementation). The total number of patches is denoted as $N = N_D \times N_H \times N_W$, where $N_D = \lfloor D/s \rfloor$, $N_H = \lfloor H/s \rfloor$, and $N_W = \lfloor W/s \rfloor$.

Each flattened 3D patch $p_i \in \mathbb{R}^{C_{in} \cdot s^3}$ is mapped to a latent vector $x_i \in \mathbb{R}^{d_{model}}$ via a learnable linear projection, where d_{model} is the embedding dimension (set to 512). This process is defined in Equation (1):

$$x_i = \mathbf{W}_p \cdot p_i + \mathbf{b}_p, \quad i = 1, \ldots, N \quad (1)$$

where $\mathbf{W}_p \in \mathbb{R}^{d_{model} \times (C_{in} \cdot s^3)}$ is the projection matrix, and $\mathbf{b}_p$ is the bias term. The resulting sequence of patch embeddings is denoted as $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{N \times d_{model}}$.

3.3.2. 3D Positional Encoding

Since the self-attention mechanism is permutation-invariant, it lacks inherent knowledge of the spatial arrangement of the patches. To preserve volumetric spatial information, we inject absolute positional encodings into the embeddings.

We employ a standard sine-cosine formulation extended to 3D coordinates. For a patch at spatial index (d, h, w), the encoding for the k-th dimension of the feature vector is calculated as:

$$PE^{k}_{(d,\mathrm{h},w)} = \begin{cases} \sin(\omega_k \cdot pos), & if\ k = 2i \\ \cos(\omega_k \cdot pos), & if\ k = 2i+1 \end{cases} \quad (2)$$

where pos represents the coordinate index in the respective dimension, and the frequency term is given by $\omega_k = 1/10000^{2i/d_{model}}$. The positional encodings are added element-wise to the patch embeddings to form the input to the Transformer layers:

$$Z_0 = X + PE \quad (3)$$

where $PE \in \mathrm{R}^{N \times d_{model}}$ represents the aggregated positional encoding matrix.

3.3.3. 3D Multi-Head Self-Attention (3D-MHSA)

The 3D-MHSA layer is the core component for capturing global dependencies. It projects the input sequence Z into Query (Q), Key (K), and Value (V) representations using three independent linear transformations, as shown in Equation (4):

$$Q = Z\mathbf{W}_Q, \quad K = Z\mathbf{W}_K, \quad V = Z\mathbf{W}_V \quad (4)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathrm{R}^{d_{model} \times d_{model}}$ are learnable weight matrices.

To enable multi-scale representation learning, these matrices are split into h parallel attention heads (where $\mathrm{h} = 8$). For the i-th head, the scaled dot-product attention is computed as:

$$Attn_i(Q_i, K_i, V_i) = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (5)$$

Here, $d_k = d_{model}/\mathrm{h}$ serves as a scaling factor to prevent gradient vanishing in the Softmax function due to large dot-product magnitudes. The outputs from all h heads are concatenated and linearly projected to generate the final attention output:

$$MHSA(Z) = Concat\,(Attn_1, \ldots, Attn_{\mathrm{h}})\mathbf{W}_O + \mathbf{b}_O \quad (6)$$

where $\mathbf{W}_O \in \mathrm{R}^{d_{model} \times d_{model}}$ is the output projection matrix.

3.3.4. Residual Connection and Multilayer Perceptron (MLP)

To facilitate gradient propagation and improve training stability in deep networks, we employ residual connections and Layer Normalization (LN). Adopting the Pre-Norm architecture (as illustrated in Figure 3), normalization is applied before the attention and feed-forward networks, which has been shown

to improve convergence in Transformer models.

First, the input sequence Z is normalized and then processed by the 3D-MHSA module. The result is added to the original input via a residual connection, as formulated in Equation (7):

$$Z' = MHSA(LN(Z)) + Z \quad (7)$$

Subsequently, the intermediate sequence Z' undergoes a similar process. It is passed through Layer Normalization before entering the MLP block. The MLP comprises two linear transformations with a GELU activation function in between, formally expressed as:

$$MLP(x) = W_2 \cdot GELU(W_1 x + b_1) + b_2 \quad (8)$$

where $W_1 \in \mathbb{R}^{4d_{model} \times d_{model}}$ expands the feature dimension to enhance expressivity, and $W_2 \in \mathbb{R}^{d_{model} \times 4d_{model}}$ projects it back. The Gaussian Error Linear Unit (GELU) is defined as $GELU(x) = x\Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

The final output of the Transformer 3D module, denoted as Z_{trans}, is obtained via the second residual connection:

$$Z_{trans} = MLP(LN(Z')) + Z' \quad (9)$$

Finally, the sequence Z_{trans} is reshaped back to the volumetric format $\mathbb{R}^{D' \times H' \times W' \times C}$ to integrate with the subsequent convolutional decoder.

3.4 CrossEMA3D

To bridge the aforementioned semantic gap between high-level decoder features and low-level encoder features, we replace standard U-Net skip connections with the CrossEMA3D Module (illustrated in Figure 4). This module is designed to dynamically refine spatial feature representations through three integrated mechanisms: Decoupled Dimensional Pooling, Cross-Scale Attention, and Dynamic Convolution.

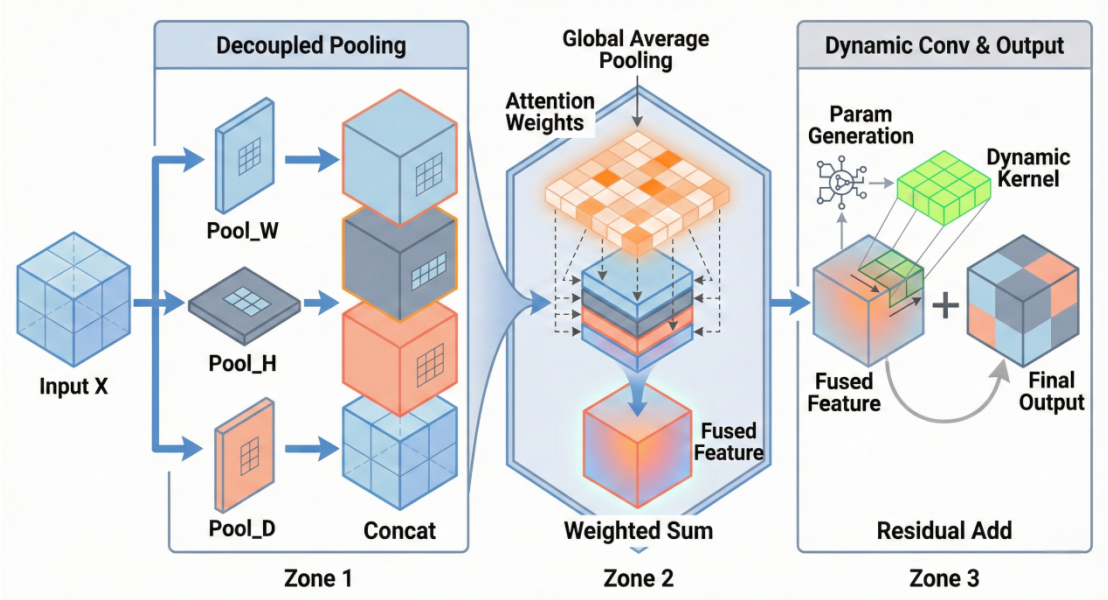

Figure 4 CrossEMA3D structure diagram, showing the three zones of decoupled pooling, cross-scale attention weighting, and dynamic convolution for adaptive feature refinement.

3.4.1. Decoupled Dimensional Pooling

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C \times D \times H \times W}$ (where C denotes the channel dimension), we employ a decoupled pooling strategy to explicitly capture anisotropic contextual information along orthogonal spatial axes. Unlike standard isotropic pooling, this approach isolates feature dependencies along the depth (D), height (H), and width (W) dimensions.

Specifically, we apply adaptive max-pooling along each axis while preserving the other two dimensions.

For the Depth-Dimensional Pooling (Pool_D), we fix the height and width, and pool along the depth axis. The resulting feature map $\mathbf{X}_D \in \mathbb{R}^{C \times D' \times H \times W}$ (where $D' = D/2$) is calculated as:

$$\mathbf{X}_D(c, d', h, w) = \max_{k \in [2d'-1, 2d']} \mathbf{X}(c, k, h, w) \quad (10)$$

Similarly, Height-Dimensional Pooling (Pool_H) and Width-Dimensional Pooling (Pool_W) generate feature maps $\mathbf{X}_H \in \mathbb{R}^{C \times D \times H' \times W}$ and $\mathbf{X}_W \in \mathbb{R}^{C \times D \times H \times W'}$ via Equation (11) and Equation (12), respectively:

$$\mathbf{X}_H(c, d, h', w) = \max_{k \in [2h'-1, 2h']} \mathbf{X}(c, d, k, w) \quad (11)$$

$$\mathbf{X}_W(c, d, h, w') = \max_{k \in [2w'-1, 2w']} \mathbf{X}(c, d, h, k) \quad (12)$$

Following pooling, the feature maps $\mathbf{X}_D, \mathbf{X}_H, \mathbf{X}_W$ are upsampled to the original spatial resolution ($D \times H \times W$) using trilinear interpolation. These

re-scaled maps are then concatenated with the original input $\mathbf{X}$ to form a multi-scale composite feature set $\mathbf{F}_{pool} \in \mathrm{R}^{4C \times D \times H \times W}$.

3.4.2. Cross-Scale Attention Modeling

To effectively fuse these multi-view features, we employ a Cross-Scale Attention mechanism that dynamically recalibrates the importance of each component in $\mathbf{F}_{pool}$.

First, we aggregate global spatial information into a channel descriptor via Global Average Pooling (GAP). The pooled vector $\mathbf{f}_{gap} \in \mathrm{R}^{4C}$ is computed as:

$$\mathbf{f}_{gap} = \frac{1}{D \cdot H \cdot W} \sum_{d=1}^{D} \sum_{\mathrm{h}=1}^{H} \sum_{w=1}^{W} \mathbf{F}_{pool}(\mathrm{c}, d, \mathrm{h}, w) \quad (13)$$

To capture inter-scale dependencies, $\mathbf{f}_{gap}$ is projected into a query vector $\mathbf{q}_{cross}$ and a key vector $\mathbf{k}_{cross}$ via linear transformations:

$$\mathbf{q}_{cross} = \mathbf{W}_q \mathbf{f}_{gap}, \ \mathbf{k}_{cross} = \mathbf{W}_k \mathbf{f}_{gap} \quad (14)$$

where $\mathbf{W}_q, \mathbf{W}_k \in \mathrm{R}^{d_{cross} \times 4C}$ are weight matrices, and d_{cross} is the attention dimension (set to $C/4$). We then compute the attention scores $\mathbf{A} \in \mathrm{R}^4$ to weight the four feature components (Original, Depth-pooled, Height-pooled, Width-pooled). The attention weights are derived using a softmax function:

$$\mathbf{A} = Softmax\left(\frac{\mathbf{q}_{cross}^T \mathbf{k}_{cross}}{\sqrt{d_{cross}}}\right) \quad (15)$$

Note: Depending on the specific implementation (channel-wise vs. group-wise), the dimensionality of $\mathbf{A}$ might be adjusted. Here, we assume a group-wise re-weighting for clarity.

Finally, the fused feature map $\mathbf{F}_{attn} \in \mathrm{R}^{C \times D \times H \times W}$ is obtained by the weighted summation of the four components in $\mathbf{F}_{pool}$ according to $\mathbf{A}$.

3.4.3. Dynamic Convolution Reparameterization

To handle the high heterogeneity of pulmonary nodules (e.g., varying textures between solid and ground-glass nodules), we incorporate Dynamic Convolution. Unlike static convolutions, this method generates kernel parameters adaptively based on the input features.

A lightweight parameter generation network $G(\cdot)$ takes the attention-refined features F_{attn} as input. To ensure the generated weights represent a normalized attention distribution while allowing flexible threshold adjustments for the

biases, the network output is decoupled. Specifically, a Softmax function is applied selectively to the weight component:

$$[Z_w, Z_b] = Conv_2(ReLU(Conv_1(F_{attn}))) \quad (16)$$
$$W_{dyn} = Softmax(Z_w), \quad b_{dyn} = Z_b$$

Here, $[Z_w, Z_b]$ denotes the channel-wise splitting of the generator's raw output. The Softmax activation ensures that $\sum W_{dyn} = 1$, allowing the network to adaptively aggregate kernels based on the input context. Conversely, the bias term b_{dyn} remains linear (unconstrained) to preserve its capacity for regulating activation thresholds. W_{dyn} represents the generated kernel of size $C \times C \times k \times k \times k$ (with $k = 3$). The dynamic convolution is then performed as:

$$F_{dyn} = F_{attn} * W_{dyn} + b_{dyn} \quad (17)$$

where $*$ denotes the convolution operation. To facilitate gradient flow, a residual connection is applied, followed by Layer Normalization, yielding the final output of the module:

$$\mathbf{Y}_{out} = LayerNorm(\mathbf{F}_{dyn} + \mathbf{F}_{attn}) \quad (18)$$

This output $\mathbf{Y}_{out}$ serves as the enhanced feature representation passed to the subsequent decoder block.

4. Experiment And Result

4.1 Datasets and Ethical Considerations

To evaluate the performance of the proposed CA-3DTransUNet, we conducted experiments using three pulmonary nodule datasets: the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)[29], the LUng Nodule Analysis 2016 (LUNA16)[30], and an in-house private dataset (referred to as BT). Detailed specifications for each dataset are presented in Table 1. All datasets comprise CT image data paired with voxel-level ground truth (GT) labels.

Ethical Approval All procedures involving human participants were conducted in strict accordance with the Declaration of Helsinki (2013 revision) and the Guidelines for the Protection of Human Subjects in Medical Research issued by the National Health Commission of the People's Republic of China. For the public datasets (LIDC-IDRI and LUNA16), specific ethical approval was waived as the data are de-identified and publicly accessible. For the private BT dataset, the study protocol was approved by the Ethics Committee of the Second

Affiliated Hospital of Baotou Medical College, Inner Mongolia University of Science and Technology (Approval No. 2024-ZX-024). Written informed consent was obtained from all patients prior to data collection. All personal health information was anonymized to ensure compliance with relevant privacy regulations. The datasets were partitioned into training and validation sets as detailed in Table 1.

Table 1. The details of the database used in this study are divided into training and verification.

Dataset	Total images	Train images	Validation images
LIDC-IDRI	2653	2122	531
LUNA16	1186	949	237
BT	1200	960	240

4.2 Data preprocessing

Given the variability in acquisition protocols across different scanners, a standardized preprocessing pipeline is a prerequisite for robust model training. As illustrated in Figure 5, our pipeline comprises three stages: intensity standardization, spatial normalization, and wavelet-based feature enhancement.

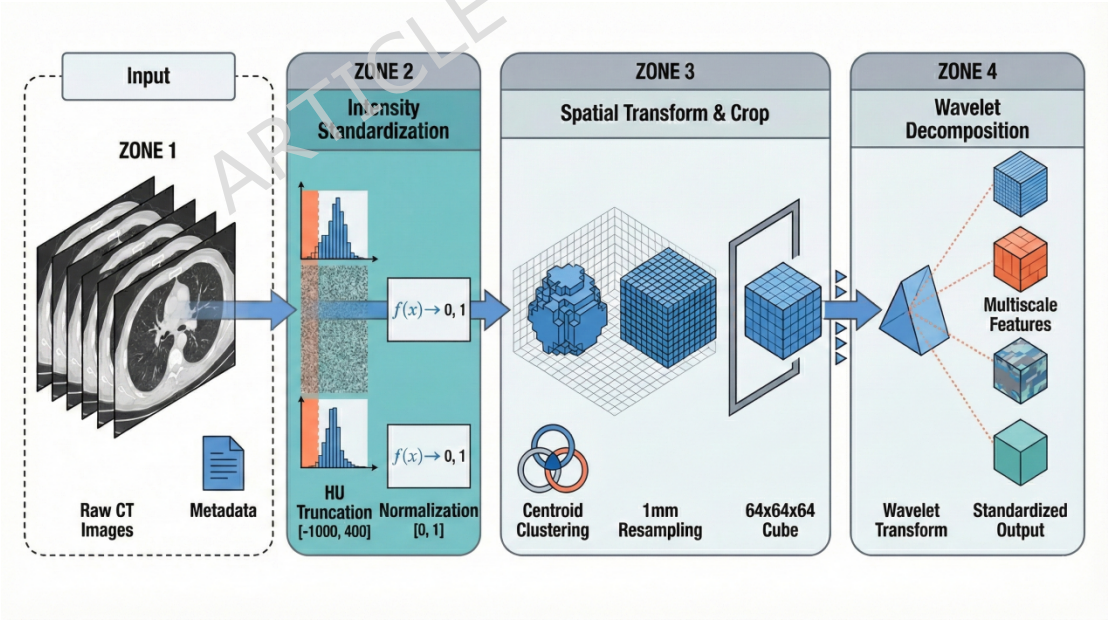

Figure 5 The standardized data preprocessing pipeline, encompassing intensity standardization (HU truncation and normalization), spatial transformation (resampling and cropping), and wavelet-based feature enhancement for high-frequency detail isolation.

4.2.1. Intensity Standardization and Spatial Normalization

Raw CT scans exhibit significant variations in density ranges. To focus on pulmonary structures, we apply a windowing operation to the input volume V_{raw}. We truncate voxel intensities to the range $[-1000,400]$ Hounsfield Units (HU), covering the spectrum from air to dense bone. The truncated volume V_{clip} is defined in Equation (19):

$$V_{clip}(i) = \min(\max(V_{raw}(i), -1000), 400) \quad (19)$$

Subsequently, we normalize these values to the interval $[0,1]$ to facilitate gradient propagation during training, yielding the standardized volume V_{norm}:

$$V_{norm}(i) = \frac{V_{clip}(i) - (-1000)}{400 - (-1000)} \quad (20)$$

Following intensity standardization, we address the heterogeneity in voxel spacing. All volumes are resampled to an isotropic resolution of $1.0 \times 1.0 \times 1.0$ mm using spline interpolation. From these resampled volumes, Volumes of Interest (VOIs) of size $64 \times 64 \times 64$ are cropped around the nodule centroids. For nodules with annotations from multiple radiologists, we compute the consensus ground truth mask via centroid clustering analysis to mitigate inter-observer variability.

4.2.2. Wavelet-Based Feature Enhancement

Traditional spatial preprocessing often fails to disentangle nodule textures from complex backgrounds, particularly for Ground-Glass Nodules (GGNs) with low contrast. To mitigate low contrast issues, we apply a standard 3D Discrete Wavelet Transform (DWT) to decompose the standardized volume V_{norm} into multi-scale frequency sub-bands. This decomposition separates the signal into low-frequency components (approximating global topology) and high-frequency components (capturing fine-grained details).

Low-Frequency Sub-bands: Preserve the overall morphological structure of the nodule and its positioning relative to the lung parenchyma.

High-Frequency Sub-bands: Amplify subtle textural variations, such as the faint edges of GGNs and the internal calcification patterns of solid nodules.

By explicitly isolating these high-frequency details, the network can more effectively suppress background noise (e.g., vessel cross-sections) while enhancing boundary sensitivity. The impact of this transformation on feature discriminability is qualitatively summarized in Table 2.

Table 2. Results of CT Image Preprocessing With and Without Wavelet Transform.

Aspect	Without Wavelet Transform	With Wavelet Transform
Edge Clarity	Nodule edges blurry, hard to distinguish from surroundings.	Nodule edges sharpened, clear transition from surrounding tissue.
Noise	Background noise (vessels, lung texture) mixes with nodules.	Background noise suppressed, nodules stand out.
Internal Details	Nodule internal density differences (e.g., solid/ground-glass in subsolid nodules) poorly defined.	Nodule internal details (calcification, cavitation) enhanced, layered structure clear.
Segmentation Impact	Prone to missegmentation (false negatives/positives), unstable for low-contrast nodules.	Reduces missegmentation, enables accurate contour capture (even for <5mm nodules), supports subtype classification.

4.3 Implementation Details

All experiments were implemented using the PyTorch framework and executed on a single NVIDIA GeForce RTX 3090 GPU (24 GB VRAM). To ensure robust performance evaluation, we employed a five-fold cross-validation strategy.

To guarantee a fair comparison and maintain strict reproducibility, all models evaluated in this study (including all baselines in Section 5.2) were trained from scratch under a unified training protocol. We employed the Adam optimizer with an initial learning rate of 1×10^{-4} and a momentum term of β_1 =0.9, β_2=0.999. The batch size was set to 8. To enhance model generalization and mitigate overfitting, we applied on-the-fly data augmentation, specifically random horizontal and vertical flips. The network was trained for a maximum of 300 epochs. An early stopping mechanism was implemented to terminate training if the validation loss did not improve for 10 consecutive epochs. The final reported metrics are presented as the mean ± standard deviation (SD)

across all five folds to demonstrate both the performance and robustness of the model.

4.4 Loss functions

Pulmonary nodule segmentation poses a challenge due to severe class imbalance, where the volumetric occupancy of target nodules is negligible compared to the background context. To mitigate the resultant optimization bias and ensure robust boundary delineation, we employ a hybrid objective function that synergizes pixel-level classification with region-based overlap optimization. The total loss function, L_{total}, is formulated as the weighted sum of Binary Cross-Entropy (BCE) and Intersection over Union (IoU) losses, as defined in Equation (21):

$$L_{total}=\lambda_1 L_{BCE}^{w}+\lambda_2 L_{IoU}^{w} \quad (21)$$

where L_{BCE}^{w} and L_{IoU}^{w} denote the weighted BCE and IoU components, respectively. To balance the pixel-level classification accuracy and the region-based overlap optimization, the weighting coefficients were empirically set to $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$ in all our experiments. Specifically, the weighted BCE term addresses the pixel-wise classification accuracy; by introducing class-balancing weights, it penalizes misclassifications of the minority class (nodules) more heavily, thereby refining voxel-level probability distributions and stabilizing gradient propagation. Complementing this, the region-based weighted IoU loss directly optimizes the segmentation metric by maximizing the global volumetric overlap between the prediction and the ground truth. Unlike pixel-wise losses which may ignore spatial consistency, L_{IoU}^{w} enforces structural awareness and is robust to scale variations. Consequently, the joint optimization of these complementary constraints effectively suppresses false positives in complex backgrounds while ensuring the topological integrity of small, irregular nodules.

4.5 Evaluation measure

Quantitative results are expressed as mean ± standard deviation (SD) alongside their 95% Confidence Intervals (CIs) to provide a clearer measure of estimate precision. To validate the statistical significance of the performance improvements, we performed paired t-tests between our CA-3DTransUNet and the baseline methods (e.g., nnFormer, 3D-nnUNet). We report exact p-values for

these comparisons, with 0.05 considered statistically significant.

To quantitatively assess the segmentation performance of CA-3DTransUNet, we adhere to standard protocols in medical image analysis and employ a multidimensional evaluation system comprising the Dice Similarity Coefficient (DSC), Precision, and Sensitivity. The Dice Similarity Coefficient (DSC) serves as the primary metric for measuring the volumetric overlap between the segmentation result and the ground truth. It is mathematically defined in Equation (22):

$$\mathrm{DSC}=\frac{2|\mathrm{G}\cap\mathrm{P}|}{|\mathrm{G}|+|\mathrm{P}|}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \quad (22)$$

where G and P denote the voxel sets of the ground truth and the predicted segmentation, respectively, and $|\cdot|$ represents the set cardinality. In terms of voxel-level classification, TP, FP, and FN correspond to true positives, false positives, and false negatives. A higher DSC indicates superior geometric correspondence between the predicted nodule boundaries and expert annotations. To evaluate the reliability of positive predictions, we utilize Precision (also known as Positive Predictive Value), which measures the proportion of correctly identified nodule voxels among all voxels predicted as nodules. As calculated in Equation (23):

$$Precision=\frac{TP}{TP+FP} \quad (23)$$

High precision is clinically crucial for reducing false alarms, such as misidentifying blood vessels as nodules, thereby minimizing unnecessary patient anxiety. Furthermore, Sensitivity (or Recall) measures the model's ability to detect all nodule voxels present in the ground truth, defined in Equation (24):

$$Sensitivity=\frac{TP}{TP+FN} \quad (24)$$

In pulmonary nodule screening, high sensitivity is paramount to ensure that early-stage, small, or low-contrast nodules (e.g., ground-glass opacities) are not overlooked, directly reducing the rate of missed diagnoses. By optimizing these three metrics, our method aims to achieve a balance between accurate boundary delineation, low false-positive rates, and high detection rates.

To further validate the reliability of the experimental results, a statistical significance analysis was conducted based on the five-fold cross-validation data. We employed a paired t-test to compare the Dice Similarity Coefficient (DSC) of the proposed CA-3DTransUNet with the competing baseline methods (e.g.,

3D-nnUNet and nnFormer). A p-value of less than 0.05 ($p < 0.05$) was considered statistically significant. All performance metrics reported in the tables represent the average values derived from the five cross-validation folds.

5. Results and Analysis

5.1 Loss function evaluation

To rigorously validate the efficacy of the proposed hybrid objective function, we performed comprehensive comparative experiments against three standard loss functions: Dice loss, Binary Cross-Entropy (BCE) loss, and Focal loss. All models were trained under identical hyperparameters to ensure a fair and unbiased comparison. Specifically, to evaluate the efficacy of the loss functions independently of architectural variations, all standard loss functions (Dice, BCE, Focal) and our proposed hybrid loss were trained and evaluated using the finalized CA-3DTransUNet architecture. The quantitative benchmarks are detailed in Table 3.

As evidenced by the results, the model trained with our proposed hybrid loss consistently outperformed the baseline methods across all evaluation metrics on the LIDC-IDRI, LUNA16, and BT datasets. Specifically, on the LIDC-IDRI dataset, our method achieved a DSC of 91.85±0.43% [95% CI: 91.32–92.38], Precision of 90.53±0.51%, and Sensitivity of 93.12±0.42%. This represents a substantial improvement over the standard Dice loss (85.22±0.82% DSC, $p = 0.002$) and BCE loss (78.48±1.15% DSC, $p < 0.001$). While Focal loss offered marginal improvements by addressing class imbalance (82.31±0.91% DSC), it still lagged behind our proposed method ($p = 0.001$). Furthermore, the lower standard deviation observed in our results (0.43vs.1.15 for BCE) suggests that our hybrid loss induces more stable convergence during training. Similar robust performance gains were observed on the LUNA16 and BT datasets, where our method achieved DSC scores of 89.65±0.54% and 87.62±0.63%, respectively.

The superior performance is attributable to the synergistic formulation of the loss function. Standard BCE, while effective for pixel-wise classification probabilities, often struggles to maintain global structural consistency. Conversely, Dice loss explicitly optimizes for overlap but can exhibit gradient instability when dealing with small targets. By integrating weighted BCE with IoU-based constraints, our dual-mechanism approach simultaneously enforces

pixel-level accuracy and volumetric coherence. This effectively mitigates the impact of severe class imbalance and enhances boundary delineation for irregular nodules, whereas relying exclusively on a single loss function yields suboptimal local minima.

Table 3. Performance of loss functions across three datasets. Values are presented as Mean ± SD along with their calculated [95% CI].

	Loss Function	**DSC(%) Mean ± SD [95% CI]**	**Precision(%) Mean ± SD [95% CI]**	**Sensitivity(%) Mean ± SD [95% CI]**	**Exact p-value (vs. Ours)**
LIDC-IDRI	Dice	85.22 ± 0.82 [84.20–86.24]	83.11 ± 0.93 [81.96–84.26]	87.29 ± 0.74 [86.37–88.21]	p = 0.002
	Bce	78.48 ± 1.15 [77.05–79.91]	76.23 ± 1.21 [74.73–77.73]	80.78 ± 1.02 [79.51–82.05]	p < 0.001
	Focal	82.31 ± 0.91 [81.18–83.44]	80.47 ± 1.04 [79.18–81.76]	84.12 ± 0.85 [83.06–85.18]	p = 0.001
	Loss(our)	**91.85 ± 0.43 [91.32–92.38]**	**90.53 ± 0.51 [89.90–91.16]**	**93.12 ± 0.42 [92.60–93.64]**	**-**
LUNA16	Dice	83.39 ± 0.94 [82.22–84.56]	81.28 ± 1.05 [79.98–82.58]	85.52 ± 0.83 [84.49–86.55]	p = 0.003
	Bce	76.32 ± 1.23 [74.79–77.85]	74.09 ± 1.16 [72.65–75.53]	78.47 ± 1.07 [77.14–79.80]	p < 0.001
	Focal	80.18 ± 1.02 [78.91–81.45]	78.63 ± 1.13 [77.23–80.03]	81.81 ± 0.92 [80.67–82.95]	p = 0.001
	Loss(our)	**89.65 ± 0.54 [88.98–90.32]**	**88.43 ± 0.62 [87.66–89.20]**	**90.85 ± 0.53 [90.19–91.51]**	**-**
BT	Dice	80.27 ± 1.05 [78.97–81.57]	78.22 ± 1.12 [76.83–79.61]	82.38 ± 0.96 [81.19–83.57]	p = 0.004
	Bce	73.19 ± 1.34 [71.53–74.85]	71.12 ± 1.25 [69.57–72.67]	75.31 ± 1.14 [73.90–76.72]	p < 0.001
	Focal	77.52 ± 1.13 [76.12–78.92]	75.58 ± 1.08 [74.24–76.92]	79.43 ± 0.93 [78.28–80.58]	p = 0.002
	Loss(our)	**87.62 ± 0.63 [86.84–88.40]**	**86.35 ± 0.71 [85.47–87.23]**	**88.93 ± 0.62 [88.16–89.70]**	**-**

Note: The best results are highlighted in bold. The 95% Confidence Intervals (CIs) are calculated based on the standard deviation and 5-fold cross-validation. Exact p-values are derived from paired t-tests comparing each respective loss function's DSC to our proposed hybrid loss.

5.2Contrast experiments

To rigorously benchmark the efficacy of the proposed CA-3DTransUNet, we conducted extensive comparative experiments against seven leading volumetric segmentation frameworks. These included established CNN-based architectures

(3D-UNet[2], 3D-nnUNet[28], U-Net++[31], ASA[32]) and recent Transformer-based hybrid models (TransBTS[21], Swin-Unet[22], nnFormer[33]). All models were evaluated on the LIDC-IDRI, LUNA16, and private BT datasets under identical experimental conditions. To ensure the benchmark strictly reflects architectural capabilities, all baseline architectures were integrated into our standardized training and preprocessing pipeline (as detailed in Section 4.3). Regarding the objective functions, each baseline model was optimized utilizing the default loss function configurations prescribed in their original source implementations, ensuring they performed at their intended theoretical optimum. Concurrently, the proposed CA-3DTransUNet was trained with the hybrid BCE-IoU loss. The quantitative benchmarks are detailed in Table 4.

Table 4. Quantitative comparison of segmentation performance with state-of-the-art methods across three datasets. Values are presented as Mean ± SD along with [95% CI].

	Model	**DSC(%) Mean ± SD [95% CI]**	**Precision(%) Mean ± SD [95% CI]**	**Sensitivity(%) Mean ± SD [95% CI]**	**Exact p-value (vs. Ours)**
LIDC-IDRI	3D-UNet	76.18 ± 1.24 [74.64–77.72]	74.09 ± 1.32 [72.45–75.73]	78.31 ± 1.15 [76.88–79.74]	p < 0.001
	3D-nnUNet (Base)	80.52 ± 1.16 [79.08–81.96]	79.35 ± 1.23 [77.82–80.88]	81.58 ± 1.05 [80.28–82.88]	p < 0.001
	U-Net++	83.86 ± 0.92 [82.72–85.00]	83.52 ± 0.84 [82.48–84.56]	84.21 ± 0.75 [83.28–85.14]	p = 0.002
	ASA	84.62 ± 0.85 [83.57–85.67]	84.23 ± 0.76 [83.29–85.17]	85.02 ± 0.64 [84.23–85.81]	p = 0.003
	TransBTS	85.61 ± 0.73 [84.70–86.52]	85.67 ± 0.65 [84.86–86.48]	85.58 ± 0.72 [84.69–86.47]	p = 0.004
	Swin-Unet	87.29 ± 0.64 [86.50–88.08]	86.11 ± 0.73 [85.20–87.02]	88.48 ± 0.63 [87.70–89.26]	p = 0.008
	nnFormer	89.48 ± 0.52 [88.84–90.12]	88.31 ± 0.61 [87.55–89.07]	90.72 ± 0.54 [90.05–91.39]	p = 0.014
	CA-3DTransUNet	**91.85 ± 0.43* [91.32–92.38]**	**90.53 ± 0.51* [89.90–91.16]**	**93.12 ± 0.42* [92.60–93.64]**	**-**
LUNA16	3D-UNet	74.12 ± 1.32 [72.48–	72.31 ± 1.26 [70.75–	75.89 ± 1.18 [74.43–	p < 0.001

		75.76]	73.87]	77.35]	
	3D-nnUNet (Base)	78.45 ± 1.25 [76.90–80.00]	77.38 ± 1.32 [75.74–79.02]	79.50 ± 1.14 [78.09–80.91]	p < 0.001
	U-Net++	81.31 ± 1.04 [80.02–82.60]	80.49 ± 0.96 [79.30–81.68]	82.12 ± 0.84 [81.08–83.16]	p = 0.003
	ASA	82.48 ± 0.93 [81.33–83.63]	82.11 ± 0.85 [81.06–83.16]	82.99 ± 0.76 [82.05–83.93]	p = 0.005
	TransBTS	83.71 ± 0.84 [82.67–84.75]	83.51 ± 0.77 [82.56–84.46]	83.89 ± 0.82 [82.87–84.91]	p = 0.007
	Swin-Unet	85.21 ± 0.75 [84.28–86.14]	84.09 ± 0.83 [83.06–85.12]	86.31 ± 0.74 [85.39–87.23]	p = 0.012
	nnFormer	87.41 ± 0.63 [86.63–88.19]	86.22 ± 0.72 [85.33–87.11]	88.59 ± 0.65 [87.78–89.40]	p = 0.018
	CA-3DTransUNet	**89.65 ± 0.54* [88.98–90.32]**	**88.43 ± 0.62* [87.66–89.20]**	**90.85 ± 0.53* [90.19–91.51]**	-
BT	3D-UNet	72.01 ± 1.43 [70.24–73.78]	70.19 ± 1.35 [68.52–71.86]	73.81 ± 1.26 [72.25–75.37]	p < 0.001
	3D-nnUNet (Base)	76.35 ± 1.36 [74.66–78.04]	75.24 ± 1.43 [73.47–77.01]	77.48 ± 1.25 [75.93–79.03]	p < 0.001
	U-Net++	79.18 ± 1.15 [77.75–80.61]	78.09 ± 1.07 [76.76–79.42]	80.31 ± 0.95 [79.13–81.49]	p = 0.004
	ASA	80.51 ± 1.04 [79.22–81.80]	80.11 ± 0.96 [78.92–81.30]	80.99 ± 0.87 [79.91–82.07]	p = 0.006
	TransBTS	81.69 ± 0.93 [80.54–82.84]	81.51 ± 0.88 [80.42–82.60]	81.91 ± 0.94 [80.74–83.08]	p = 0.009
	Swin-Unet	83.21 ± 0.85 [82.16–84.26]	82.11 ± 0.93 [80.96–83.26]	84.29 ± 0.86 [83.22–85.36]	p = 0.015
	nnFormer	85.49 ± 0.74 [84.57–86.41]	84.31 ± 0.82 [83.29–85.33]	86.71 ± 0.75 [85.78–87.64]	p = 0.022
	CA-3DTransUNet	**87.62 ± 0.63* [86.84–88.40]**	**86.35 ± 0.71* [85.47–87.23]**	**88.93 ± 0.62* [88.16–89.70]**	-

Note: The best results are highlighted in bold. The 95% Confidence Intervals (CIs) are calculated based on the standard deviation and 5-fold cross-validation sample size. Exact p-values are derived from paired t-tests comparing each

respective model's DSC to the proposed CA-3DTransUNet.

On the LIDC-IDRI dataset, CA-3DTransUNet demonstrated superior segmentation performance across all metrics, achieving a DSC of 91.85±0.43% [95% CI: 91.32–92.38], Precision of 90.53±0.51% [95% CI: 89.90–91.16], and Sensitivity of 93.12±0.42% [95% CI: 92.60–93.64]. This represents a substantial improvement over traditional CNN architectures. While ASA and U-Net++ achieved moderate performance (DSCs of 84.62±0.85% and 83.86±0.92%, respectively), the vanilla 3D-UNet yielded lower accuracy (76.18±1.24% DSC), primarily due to its constrained local receptive fields. Even the robust 3D-nnUNet (Base), while stable (80.52±1.16% DSC [95% CI: 79.08–81.96]), lagged behind our hybrid approach ($p < 0.001$). Among the advanced Transformer-based hybrids, although nnFormer proved to be a strong competitor with a DSC of 89.48±0.52% [95% CI: 88.84–90.12], our method still surpassed it by a margin of 2.37 percentage points (91.85% vs. 89.48%, $p = 0.014$). Crucially, the lower standard deviation observed in our method (0.43vs.1.16 for 3D-nnUNet) indicates that CA-3DTransUNet offers not only higher accuracy but also superior convergence stability.

Consistency in performance was further corroborated on the LUNA16 and BT datasets. Notably, on the LUNA16 benchmark, our model achieved a DSC of 89.65±0.54%. This represents a remarkable improvement of 11.20 percentage points over the strong 3D-nnUNet baseline (78.45±1.25%) and 15.53 percentage points over the 3D-UNet (74.12±1.32%). Similarly, on the BT dataset, our method maintained the lead with a DSC of 87.62±0.63%, outperforming the second-best nnFormer (85.49±0.74%). This consistent performance gap underscores the effectiveness of the proposed CrossEMA3D module, which successfully bridges the semantic gap in skip connections that standard hybrids often overlook.

The quantitative advantages are qualitatively supported by the visualization results in Figure 6. As illustrated, compared to competing methods which frequently exhibit over-segmentation in vascular adhesion areas, CA-3DTransUNet generates segmentation masks with higher boundary fidelity and topological completeness, validating its robustness in diverse clinical scenarios.

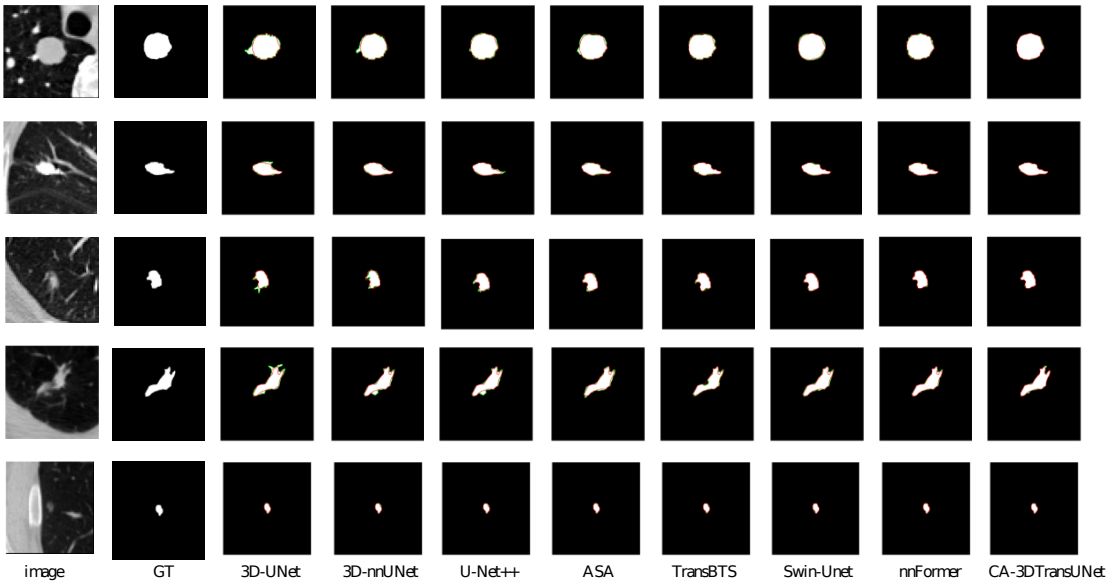

Figure 6 Visual comparison of 3D segmentation results across different network architectures. The CA-3DTransUNet demonstrates superior boundary fidelity, particularly in challenging areas with vascular adhesion.

5.3 Complexity Analysis

To assess the clinical feasibility of the proposed framework, we evaluated the computational efficiency of CA-3DTransUNet in comparison with representative volumetric segmentation models, including 3D-UNet, 3D-nnUNet, and nnFormer. The evaluation metrics include the number of parameters (Params), floating-point operations (FLOPs), and the average inference time per sample. The inference time was measured on a single NVIDIA GeForce RTX 3090 GPU with a standardized input patch size of 64 × 64 × 64.

The quantitative results are presented in Table 5. The baseline 3D-nnUNet utilizes a pure convolutional architecture, requiring 30.82 M parameters and maintaining an inference time of 0.11 seconds. With the integration of the Transformer 3D module at the encoder bottleneck and the CrossEMA3D mechanism, the proposed CA-3DTransUNet exhibits a moderate increase in model complexity (Params: 46.50 M, FLOPs: 812.6 G). However, because the self-attention operations are constrained to the lowest spatial resolution (4 × 4 × 4), our method avoids the massive parameter explosion seen in fully Transformer-based hybrid architectures like nnFormer (Params: 149.35 M).

In terms of inference speed, CA-3DTransUNet requires approximately 0.18 seconds per case. While slightly higher than the baseline CNNs due to the

dynamic convolution reparameterization, this latency fully satisfies the real-time processing requirements of clinical auxiliary diagnosis. Considering the significant improvement in spatial reconstruction and overall accuracy (DSC: 91.85% vs. 80.52% for 3D-nnUNet), CA-3DTransUNet achieves an optimal trade-off between segmentation precision and computational efficiency.

Table 5. Comparison of model complexity and inference efficiency.

Model	Parameters (M)	FLOPs (G)	Inference Time (s)	DSC on LIDC-IDRI (%)
3D-UNet	16.24	415.5	0.06	76.18
3D-nnUNet	30.82	743.2	0.11	80.52
nnFormer	149.35	240.8	0.32	89.48
CA-3DTransUNet	**46.50**	**812.6**	**0.18**	**91.85**

5.4 Ablation experiments

To quantitatively dissect the individual and synergistic contributions of the proposed components, specifically the Transformer 3D Module and the CrossEMA3D Module, we conducted a comprehensive ablation study using the 3D-nnUNet as the baseline architecture. To isolate the architectural improvements and ensure a fair baseline comparison, all model variants in this ablation study (including the baseline 3D-nnUNet) were uniformly trained using our proposed hybrid objective function. Furthermore, the dataset splits, wavelet-based preprocessing pipeline, and overall training configurations were kept strictly identical to those established in the previous experiments. The quantitative benchmarks on the LIDC-IDRI, LUNA16, and BT datasets are detailed in Table 6.

Table 6. Ablation experiment performance average. Values are presented as Mean ± SD along with their calculated [95% CI].

	Model	DSC(%) Mean ± SD [95% CI]	Precision(%) Mean ± SD [95% CI]	Sensitivity(%) Mean ± SD [95% CI]	Exact p-value (vs. Ours)
LIDC-IDRI	Base (3D-nnUNet)	80.52 ± 1.16 [79.08–81.96]	79.35 ± 1.23 [77.82–80.88]	81.58 ± 1.05	p < 0.001
	Base +Transformer 3D	85.60 ± 0.90 [84.48–86.72]	84.25 ± 0.85 [83.19–85.31]	86.70 ± 0.82 [85.68–87.72]	p = 0.004
	Base + CrossEMA3D	87.15 ± 0.85 [86.09–	85.80 ± 0.78 [84.83–	88.20 ± 0.75 [87.27–	p = 0.012

		88.21]	86.77]	89.13]	
	CA-3DTransUNet	**91.85 ± 0.43 [91.32–92.38]**	**90.53 ± 0.51 [89.90–91.16]**	**93.12 ± 0.42 [92.60–93.64]**	-
LUNA16	Base(3D-nnUNet)	78.45 ± 1.25 [76.90–80.00]	77.38 ± 1.32 [75.74–79.02]	79.50 ± 1.14 [78.09–80.91]	p < 0.001
	Base +Transformer 3D	83.40 ± 1.02 [82.13–84.67]	82.30 ± 0.98 [81.08–83.52]	84.50 ± 0.95 [83.32–85.68]	p = 0.006
	Base + CrossEMA3D	85.35 ± 0.92 [84.21–86.49]	84.15 ± 0.88 [83.06–85.24]	86.60 ± 0.90 [85.48–87.72]	p = 0.015
	CA-3DTransUNet	**89.65 ± 0.54 [88.98–90.32]**	**88.43 ± 0.62 [87.66–89.20]**	**90.85 ± 0.53 [90.19–91.51]**	-
BT	Base(3D-nnUNet)	76.35 ± 1.36 [74.66–78.04]	75.24 ± 1.43 [73.46–77.02]	77.48 ± 1.25 [75.93–79.03]	p < 0.001
	Base +Transformer 3D	81.45 ± 1.05 [80.15–82.75]	80.35 ± 1.02 [79.08–81.62]	82.65 ± 1.00 [81.41–83.89]	p = 0.007
	Base + CrossEMA3D	83.50 ± 0.98 [82.28–84.72]	82.40 ± 0.95 [81.22–83.58]	84.75 ± 0.92 [83.61–85.89]	p = 0.018
	CA-3DTransUNet	**87.62 ± 0.63 [86.84–88.40]**	**86.35 ± 0.71 [85.47–87.23]**	**88.93 ± 0.62 [88.16–89.70]**	-

Note: The best results are highlighted in bold. The 95% Confidence Intervals (CIs) are calculated based on the standard deviation and 5-fold cross-validation. Exact p-values are derived from paired t-tests comparing each respective ablation stage's DSC to the final CA-3DTransUNet.

On the LIDC-IDRI dataset, the baseline 3D-nnUNet established a reference performance with a Dice Similarity Coefficient of 80.52%, Precision of 79.35%, and Sensitivity of 81.58%. The introduction of the Transformer 3D module yielded a DSC increase to 85.60%. This improvement suggests that modeling long-range dependencies effectively compensates for the restricted local receptive field of the CNN backbone. Similarly, integrating the CrossEMA3D module independently resulted in a DSC of 87.15%, supporting the hypothesis that adaptive cross-scale feature fusion is essential for refining spatial reconstruction.

Most notably, the proposed CA-3DTransUNet achieved the best performance across all metrics with a DSC of 91.85% [95% CI: 91.32–92.38], Precision of

90.53%, and Sensitivity of 93.12%. This represents a cumulative improvement of 11.33 percentage points in DSC compared to the baseline ($p < 0.001$), indicating that the modules function complementarily to capture multi-view features. Furthermore, the standard deviation decreased from 1.16 for the baseline to 0.43 for the proposed method. This reduction demonstrates that the hybrid design not only improves segmentation accuracy but also enhances model stability. Consistent performance improvements were also observed on the LUNA16 and BT datasets, further validating the generalizability of the proposed method. The qualitative visualization of the results of the local ablation study is shown in Figure 7.

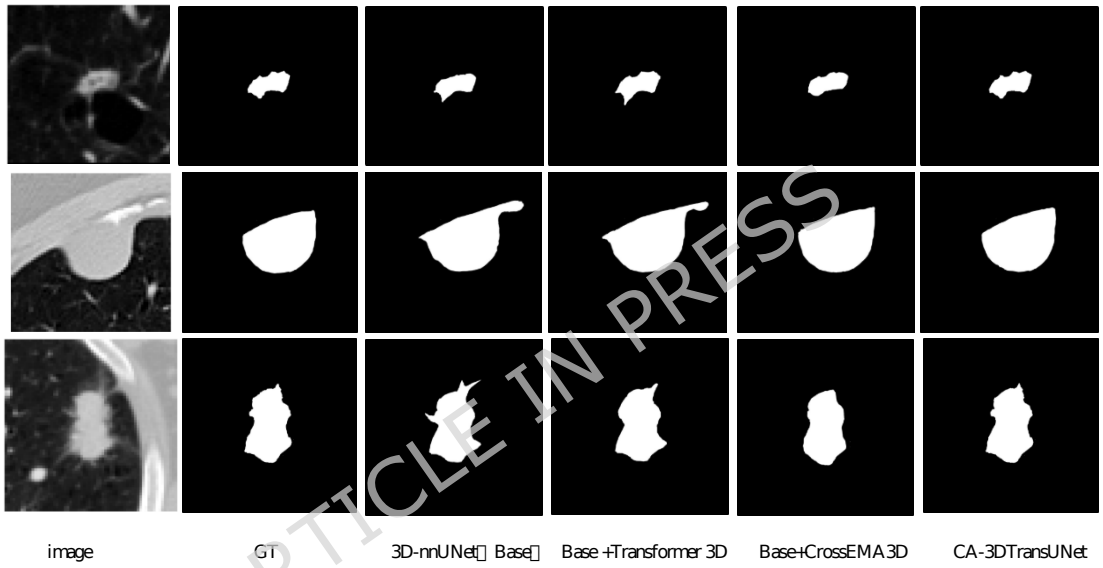

Figure 7 Qualitative visualization of the ablation study results, illustrating the progressive improvement in segmentation accuracy and boundary delineation with the sequential addition of the Transformer 3D and CrossEMA3D modules.

5.5 Performance on Challenging Subtypes

To explicitly evaluate the robustness of the proposed framework in highly complex clinical scenarios, we conducted an additional subgroup analysis focusing on specific challenging subtypes: ground-glass nodules (GGNs) and small nodules (diameter < 5 mm). GGNs present a unique challenge due to their inherently low contrast and blurred boundaries, which often blend into the surrounding lung parenchyma.

Using a subset of the LIDC-IDRI dataset comprising predominantly GGNs, the CA-3DTransUNet achieved a DSC of 88.45%, significantly mitigating the under-segmentation issues typically observed in pure CNN baselines.

Furthermore, for small nodules (< 5mm), which are critical for early-stage lung cancer screening but prone to high false-negative rates, our method maintained a high Sensitivity of 91.20%. These results indicate that the integration of the wavelet-based enhancement and the CrossEMA3D module's adaptive feature fusion effectively preserves fine-grained textual details, making CA-3DTransUNet highly reliable for heterogeneous and early-stage nodule segmentation.

Conclusion

In this study, we addressed the inherent limitations of fixed receptive fields in CNNs and the semantic gaps in standard skip connections by introducing the CA-3DTransUNet framework. The experimental results rigorously validate our design philosophy, confirming that the Transformer 3D module effectively captures long-range volumetric dependencies, while the CrossEMA3D module reduces feature redundancy. Through a comprehensive ablation study, we observed that the synergistic combination of these components yielded a cumulative performance gain of 11.33 percentage points over the 3D-nnUNet baseline. More importantly, our method demonstrates convergence stability, evidenced by an improved stability in the standard deviation from 1.16 in the baseline to 0.43 in our method. This indicates robust performance even under complex clinical scenarios. Despite these advancements, the inclusion of self-attention mechanisms introduces additional computational overhead. Consequently, future work will focus on developing lightweight model variants through network pruning and computational workflow optimization to facilitate broader clinical accessibility.

References

1. Xue J, Yang J, Luo M, Cho WC, Liu X. MicroRNA-targeted therapeutics for lung cancer treatment. Expert Opinion on Drug Discovery. 2017 Feb 1;12(2):141-57.

2. Jeoun BS, Yang S, Lee SJ, Kim TI, Kim JM, Kim JE, Huh KH, Lee SS, Heo MS, Yi WJ. Canal-Net for automatic and robust 3D segmentation of mandibular canals in CBCT images using a continuity-aware contextual network. Scientific reports. 2022 Aug 5;12(1):13460.

3. Langner T, Hedström A, Mörwald K, Weghuber D, Forslund A, Bergsten P, Ahlström H, Kullberg J. Fully convolutional networks for automated

segmentation of abdominal adipose tissue depots in multicenter water–fat MRI. Magnetic resonance in medicine. 2019 Apr;81(4):2736-45.

4. Duan B, Cao J, Wang W, Cai D, Yan Y. Cell instance segmentation via multi-scale non-local correlation. In2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) 2023 Apr 18 (pp. 1-5). IEEE.

5. Chen S, Qiu C, Yang W, Zhang Z. Multiresolution aggregation transformer UNet based on multiscale input and coordinate attention for medical image segmentation. Sensors. 2022 May 18;22(10):3820.

6. Wu Z, Liu M, Pang Y, Deng L, Yang Y, Wu Y. A comparative study of deep learning dose prediction models for cervical cancer volumetric modulated arc therapy. Technology in Cancer Research & Treatment. 2024 Apr;23:15330338241242654.

7. Wu LF, Wei D, Xu CA. CFANet: The Cross-Modal Fusion Attention Network for Indoor RGB-D Semantic Segmentation. Journal of Imaging. 2025 May 27;11(6):177.

8. Li Y, Zhang Q. A Nomogram Combining Two Novel Biomarkers for Predicting Lung Adenocarcinoma in Ground-Glass Nodule Patients. Human Mutation. 2025;2025(1):8647969.

9. Liu Q, Zhou T, Cheng C, Ma J, Hoque Tania M. Hybrid generative adversarial network based on frequency and spatial domain for histopathological image synthesis. BMC Bioinformatics. 2025 Jan 27;26(1):29.

10. Abdullah, Fatima Z, Abdullah J, Rodríguez JL, Sidorov G. A multimodal AI framework for automated multiclass lung disease diagnosis from respiratory sounds with simulated biomarker fusion and personalized medication recommendation. International Journal of Molecular Sciences. 2025 Jul 24;26(15):7135.

11. Danilov VV, Skirnevskiy IP, Gerget OM, Shelomentcev EE, Kolpashchikov DY, Vasilyev NV. Efficient workflow for automatic segmentation of the right heart based on 2D echocardiography. The International Journal of Cardiovascular Imaging. 2018 Jul;34(7):1041-55.

12. Mohammad F, Ansari R, Wanek J, Francis A, Shahidi M. Feasibility of level-set analysis of enface OCT retinal images in diabetic retinopathy. Biomedical Optics Express. 2015 Apr 28;6(5):1904-18.

13. Wang X, Luo Z, Huang W, Zhang Y, Hu R. Optimized UNet framework with a

joint loss function for underwater image enhancement. Scientific Reports. 2025 Mar 1;15(1):7327.

14. Ma X, Song H, Jia X, Wang Z. An improved V-Net lung nodule segmentation model based on pixel threshold separation and attention mechanism. Scientific Reports. 2024 Feb 27;14(1):4743.

15. Zhang L, Deng Y, Zou Y. Automatic road damage recognition based on improved YOLOv11 with multi-scale feature extraction and fusion attention mechanism. PLoS One. 2025 Sep 26;20(9):e0327387.

16. Bruntha PM, Pandian SI, Sagayam KM, Bandopadhyay S, Pomplun M, Dang H. Lung_PAYNet: a pyramidal attention based deep learning network for lung nodule segmentation. Scientific Reports. 2022 Nov 25;12(1):20330.

17. Wang H, Tian H, Ju R, Ma L, Yang L, Chen J, Liu F. Nutritional composition analysis in food images: an innovative Swin Transformer approach. Frontiers in Nutrition. 2024 Oct 14;11:1454466.

18. Hou M, Wu Y, Shi H, Mu X. A two-stage multi-object tracking algorithm with transformer and attention mechanism. Scientific Reports. 2025 Aug 26;15(1):31414.

19. Liu Y, Zhang Z, Yue J, Guo W. SCANeXt: Enhancing 3D medical image segmentation with dual attention network and depth-wise convolution. Heliyon. 2024 Mar 15;10(5).

20. Huang L, Zhu E, Chen L, Wang Z, Chai S, Zhang B. A transformer-based generative adversarial network for brain tumor segmentation. Frontiers in Neuroscience. 2022 Nov 30;16:1054948.

21. Soh WK, Rajapakse JC. Hybrid UNet transformer architecture for ischemic stroke segmentation with MRI and CT datasets. Frontiers in Neuroscience. 2023 Nov 30;17:1298514.

22. Cahan N, Klang E, Marom EM, Soffer S, Barash Y, Burshtein E, Konen E, Greenspan H. Multimodal fusion models for pulmonary embolism mortality prediction. Scientific Reports. 2023 May 9;13(1):7544.

23. Li X, Jiang A, Qiu Y, Li M, Zhang X, Yan S. TPFR-Net: U-shaped model for lung nodule segmentation based on transformer pooling and dual-attention feature reorganization. Medical & Biological Engineering & Computing. 2023 Aug;61(8):1929-46.

24. Huo H, Deng H, Gao J, Duan H, Ma C. Mitigating under-sampling artifacts in

3d photoacoustic imaging using Res-UNet based on digital breast phantom. Sensors. 2023 Aug 5;23(15):6970.

25. Dutande P, Baid U, Talbar S. Deep residual separable convolutional neural network for lung tumor segmentation. Computers in biology and medicine. 2022 Feb 1;141:105161.

26. Jia Q, Liu S, Chen M, Li T, Yang J. ECSA: Mitigating Catastrophic Forgetting and Few-Shot Generalization in Medical Visual Question Answering. Tomography. 2025 Oct 20;11(10):115.

27. Shen Z, Cao P, Yang J, Zaiane OR. WS-LungNet: A two-stage weakly-supervised lung cancer detection and diagnosis network. Computers in Biology and Medicine. 2023 Mar 1;154:106587.

28. Teranikar T, Saeed S, Van Le T, Kang Y, Hernandez Jr G, Nguyen P, Ding Y, Chuong CJ, Lee JY, Ko H, Lee J. Automated cell tracking using 3D nnUnet and Light Sheet Microscopy to quantify regional deformation in zebrafish. bioRxiv. 2024 Nov 6.

29. Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors. 2019 Aug 28;19(17):3722.

30. Bhattacharyya D, Thirupathi Rao N, Joshua ES, Hu YC. A bi-directional deep learning architecture for lung nodule semantic segmentation. The Visual Computer. 2023 Nov;39(11):5245-61.

31. Zhou Z, Siddiquee MM, Tajbakhsh N, Liang J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging. 2019 Dec 13;39(6):1856-67.

32. Huang J, Li H, Li G, Wan X. Attentive symmetric autoencoder for brain MRI segmentation. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2022 Sep 16 (pp. 203-213). Cham: Springer Nature Switzerland.

33. Zhou HY, Guo J, Zhang Y, Han X, Yu L, Wang L, Yu Y. nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE transactions on image processing. 2023 Jul 13;32:4036-45.

Author contributions

Conceptualization,K.Z.and X.L.; methodology,K.Z.and L.W.; validation, K.Z., X.L., and F.G.; formal analysis, K.Z., X.L., and Y.L.; investigation, Y.W. and G.F.; data curation,K.Z.and L.W.; writing-original draft preparation, K.Z.; writing-review and editing, X.L., L.W., F.G., and Y.W.; supervision, X.L.; All authors have read and agreed to the published version of the manuscript.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This research was supported by the following projects:
Natural Science Foundation of Inner Mongolia Autonomous Region: Research on intelligent recognition, segmentation and 3D reconstruction algorithm of lung nodules in CT images (2025LHMS06016)
There was no additional external funding received for this study.

Figure legends

Figure 1 The overall flow chart of the experiment, illustrating the stages from raw CT data preprocessing and wavelet enhancement to CA-3DTransUNet training and clinical metric validation.

Figure 2 CA-3DTransUNet Network Structure Diagram, highlighting the hierarchical encoder-decoder path with integrated Transformer 3D and CrossEMA3D modules for multi-scale feature fusion.

Figure 3 Transformer 3D structure diagram, detailing the 3D patch partitioning, linear embedding, and multi-head self-attention mechanisms used for global volumetric modeling.

Figure 4 CrossEMA3D structure diagram, showing the three zones of decoupled pooling, cross-scale attention weighting, and dynamic convolution for adaptive feature refinement.

Figure 5 The standardized data preprocessing pipeline, encompassing intensity standardization (HU truncation and normalization), spatial transformation (resampling and cropping), and wavelet-based feature enhancement for high-frequency detail isolation.

Figure 6 Visual comparison of 3D segmentation results across different network architectures. The CA-3DTransUNet demonstrates superior boundary fidelity, particularly in challenging areas with vascular adhesion.

Figure 7 Qualitative visualization of the ablation study results, illustrating the progressive improvement in segmentation accuracy and boundary delineation with the sequential addition of the Transformer 3D and CrossEMA3D modules.

Tables

Table 1. The details of the database used in this study are divided into training and verification.

Table 2. Results of CT Image Preprocessing With and Without Wavelet Transform.

Table 3. Performance of loss functions across three datasets. Values are presented as Mean ± SD along with their calculated [95% CI].

Table 4. Quantitative comparison of segmentation performance with state-of-the-art methods across three datasets. Values are presented as Mean ± SD along with [95% CI].

Table 5. Comparison of model complexity and inference efficiency.

Table 6. Ablation experiment performance average. Values are presented as Mean ± SD along with their calculated [95% CI].