Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-025-03425-8

ORIGINAL ARTICLE

S^{3}TU-Net: Structured convolution and superpixel transformer for lung nodule segmentation

Yuke Wu[1] · Xiang Liu[1] · Yunyu Shi[1] · Xinyi Chen[1] · Zhenglei Wang[2] · YuQing Xu[3] · ShuoHong Wang[4]

Received: 18 February 2025 / Accepted: 16 July 2025

Abstract
Accurate segmentation of lung adenocarcinoma nodules in computed tomography (CT) images is critical for clinical staging and diagnosis. However, irregular nodule shapes and ambiguous boundaries pose significant challenges for existing methods. This study introduces S^{3}TU-Net, a hybrid CNN-Transformer architecture designed to enhance feature extraction, fusion, and global context modeling. The model integrates three key innovations: (1) structured convolution blocks (DWF-Conv/D^{2}BR-Conv) for multi-scale feature extraction and overfitting mitigation; (2) S^2-MLP Link, a spatial-shift-enhanced skip-connection module to improve multi-level feature fusion; and 3) residual-based superpixel vision transformer (RM-SViT) to capture long-range dependencies efficiently. Evaluated on the LIDC-IDRI dataset, S^{3}TU-Net achieves a Dice score of 89.04%, precision of 90.73%, and IoU of 90.70%, outperforming recent methods by 4.52% in Dice. Validation on the EPDB dataset further confirms its generalizability (Dice, 86.40%). This work contributes to bridging the gap between local feature sensitivity and global context awareness by integrating structured convolutions and superpixel-based transformers, offering a robust tool for clinical decision support.

Keywords Structured convolution · Spatial shift · Superpixel · Vision transformer · Lung adenocarcinoma nodule · Image segmentation

1 Introduction

Lung cancer [1–5] is one of the most prevalent and fatal cancers worldwide, with lung adenocarcinoma as the predominant subtype, representing more than 50% of cases [6]. According to the WHO classification [7], lung adenocarcinoma is categorized into distinct stages—atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC)—each displaying unique imaging characteristics on computed tomography (CT) scans [8]. In the early stages, lesions typically appear as ground-glass opacities with clear, regular borders, making detection challenging. In contrast, advanced stages reveal solid, irregular masses with spiculated or lobulated shapes [9]. The manual diagnosis of CT images often risks missing subtle details, making accurate staging difficult. Recent advancements in machine learning and deep learning, including the exploration of diffusion models [10–12], have introduced automated classification techniques for lung nodules [13]. However, unsegmented CT data introduces excessive computational demands and redundant information, impairing the effectiveness of classification algorithms and decreasing model accuracy [14]. Consequently, precise nodule segmentation has become essential, as it underpins accurate classification and provides clinicians with reliable diagnostic insights.

Recent years have seen rapid progress in pulmonary nodule segmentation techniques, shifting from traditional methods [15, 16] to deep learning approaches [17–19]. Traditional unsupervised methods, including morphological fuzzy mathematics, threshold segmentation, and fuzzy clustering [20, 21], are computationally efficient yet often lack segmentation accuracy. In contrast, deep learning-based methods, such as the widely adopted U-Net [22, 23], leverage a symmetrical encoder-decoder structure with skip connections to combine shallow encoder features with deep decoder features for enhanced segmentation accuracy. Building on this, models like CDP-ResNet [24] employ a dual-path residual network for multi-view feature extraction and edge-based

Wu Yuke, Shi Yunyu, Chen Xinyi, Wang Zhenglei, Xu YuQing, and Wang ShuoHong contributed equally to this work.

Extended author information available on the last page of the article

Published online: 20 August 2025

voxel sampling to capture small nodules. Dense-UNet [25] mitigates class imbalance issues by using dense connections and an innovative loss function to prevent overfitting and gradient vanishing. The CRF-3D-UNet [26] further enhances segmentation by integrating conditional random fields with 3D-UNet for spatial and contextual data fusion.

Despite the success of existing methods (e.g., U-Net variants) in local feature extraction, they suffer from three key limitations: (1) Limited receptive fields of conventional convolutions hinder long-range dependency modeling [27, 28]; (2) spatial context loss during multi-scale feature fusion; (3) high computational costs and sensitivity to fine-grained details when directly applying Transformers to medical images. Vision transformer (ViT) [29, 30] exemplifies this approach by effectively modeling global dependencies, though it struggles with fine-grained details and requires significant computational resources and data. Thus, hybrid architectures combining the strengths of convolutional networks and transformers offer an optimal path forward, where the efficiency of UNet in local feature extraction complements the global comprehension of ViT.

To address these limitations, we propose S^{3}TU-Net, a CNN-Transformer hybrid that integrates multi-space and multi-view fields. S^{3}TU-Net leverages the structured convolutional advantages of CNNs and the global semantic representation capabilities of superpixel-based transformers. By incorporating multi-dimensional spatial connectors for efficient feature fusion, our model enhances feature extraction, fusion, and global context modeling, ultimately improving segmentation performance.

- To overcome U-Net's limitations in handling multi-depth features, we introduce the DWF-Conv block, employing depth-weighted and deep kernel convolutions to improve feature extraction and restoration in the initial encoder-decoder stages. To mitigate overfitting, we propose the D^{2}BR-Conv block, which integrates DropBlock regularization with dual convolutions to reinforce robust feature learning and enhance generalization.
- Addressing U-Net's challenges in feature fusion, we incorporate multi-dimensional spatial connectors into the skip connections. Specifically, the S^2-MLP Link module combines multi-directional spatial shifting and distributed attention mechanisms to integrate features from varied semantic levels, thereby enhancing fusion performance.
- To improve U-Net's contextual understanding, we propose the RM-SViT module, which fuses global and local features using a multi-branch attention mechanism and superpixel visual transformers. Additionally, residual connections enhance model stability and indirectly improve computational efficiency.

Experimental results demonstrate that our model achieves a DSC of 89.04%, precision of 90.73%, mIoU of 90.70%, and sensitivity of 93.70% on the LIDC-IDRI dataset. Furthermore, validation on the independent EPDB private dataset yields a DSC of 86.40%. These results confirm that S^{3}TU offers high segmentation performance with strong model stability and generalization.

2 Methods

2.1 Network architecture

Figure 1 illustrates the proposed S^{3}TU-Net with a U-shaped encoder-decoder structure. The symmetric S^{3}TU-Net architecture primarily consists of two structured convolution blocks (DWF-Conv/D^{2}BR-Conv) used in the encoder and decoder, fusion residual connections and a multi-branch attention-based superpixel visual transformer (RM-SViT) between the encoder and decoder, and a multi-dimensional spatial connector (S^2-MLP Link) based on multi-directional spatial shifting and distributed attention at the skip connections.

Specifically, the encoder's initial stage employs the structured depth-weighted feature convolution block (DWF-Conv), which consists of two 3×3 convolutional layers, each followed by batch normalization, a scalable ReLU activation unit, and an LKA module composed of multiple deep kernel convolutions. The encoder then undergoes three downsampling stages, each comprising a structured D^{2}BR-Conv block and 2×2 max pooling. The D^{2}BR-Conv includes a 3×3 convolution, DropBlock regularization, batch normalization, and ReLU activation. The RM-SViT, with internal iterative updates, is applied between the encoder and decoder to further enhance feature representation and context understanding. The decoder begins with DWF-Conv, and each upsampling step includes a 2×2 transpose convolution that halves the number of feature channels. The feature maps from the corresponding encoder layer are processed through the S^2-MLP Link module, after which they are concatenated with the upsampled feature maps along the channel dimension. This concatenated result is then passed through the D^{2}BR-Conv block. Finally, the model's output layer employs a 1×1 convolution and Sigmoid activation to generate the segmentation map.The algorithmic workflow of S^{3}TU-Net is presented as follows:

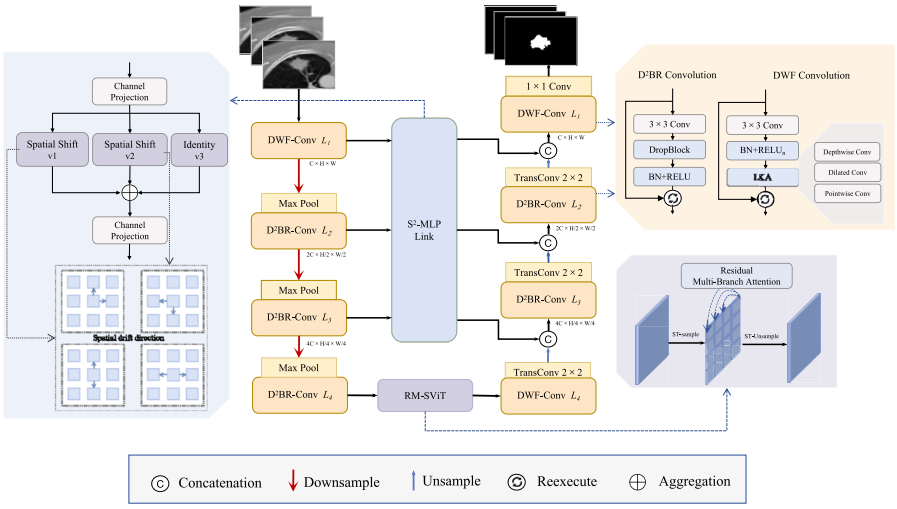

Fig. 1 The overall framework of S^{3}TU-Net. The framework is divided into three broad categories of modules, two novel convolutional modules (DWF-Conv/D^{2}BR-Conv), multi-spatial dimensional connectors (S^2-MLP Link), and residual connection-based superpixel vision transformer (RM-SViT)

Algorithm 1 S^{3}TU-Net segmentation workflow.

Require: CT Image Patch $X \in \mathbb{R}^{128\times128}$
Ensure: Segmentation Mask $Y \in \mathbb{R}^{128\times128}$
// Encoder Path
Initialize features $F_{\text{enc}} \leftarrow X$
for $i = 1$ **to** 4 **do**
  if $i = 1$ **then**
    Apply DWF-Conv Block (Fig. 3(b)) to F_{enc}
  else
    Apply D^{2}BR-Conv Block (Fig. 3(a)) to F_{enc}
  end if
  Downsample via MaxPooling (stride=2)
  $F_{\text{enc}} \leftarrow$ Updated Features
end for
// RM-SViT Global Context Modeling
$F_{\text{global}} \leftarrow$ RM-SViT(F_{enc}) (Eq. 3–6)
// Decoder Path
for $j = 4$ **downto** 1 **do**
  Upsample F_{global} via Transposed Conv (stride=2)
  Concatenate with Skip Connection F_{skip} via S^2-MLP Link (Fig. 5
  Apply D^{2}BR-Conv Block to fused features
  $F_{\text{global}} \leftarrow$ Updated Features
end for
// Final Segmentation
$Y \leftarrow \text{Sigmoid}(\text{Conv}_{1\times1}(F_{\text{global}}))$
Compute Loss: $\mathcal{L} = \alpha\mathcal{L}_{\text{BCE}} + (1-\alpha)\mathcal{L}_{\text{Dice}}$
return Y

2.2 Structured convolutional modules

Traditional U-Net and its various improved versions often suffer from severe overfitting during training. This issue is typically addressed through data augmentation, L2 regularization, or Dropout [31], which randomly drops a portion of neurons. In this work, we adopt a more generalized approach—DropBlock [32]. This spatial regularization technique randomly removes contiguous regions from feature maps, forcing the model to make correct predictions even with missing local information. DropBlock controls the size and number of dropped blocks using parameters like block_size and y.block_size, effectively preventing overfitting in convolutional networks. Batch normalization is also employed to accelerate the training process, stabilize gradient flow, and prevent issues like vanishing or exploding gradients. Additionally, to enhance feature representation, we incorporate deep large kernel convolutions and dilated convolutions, which capture a wider range of features without increasing computational costs [33]. We also draw inspiration from the FreeU model [34] and Squeeze-and-Excitation Networks (SE-Net) [35], which apply feature re-weighting: FreeU improves feature fusion by re-weighting the features between skip connections and backbone feature maps, while SE-Net enhances important features and suppresses irrelevant ones by channel-wise re-weighting after each convolutional layer using global average pooling and fully connected layers (Fig. 2a).

Based on these insights, we designed two structured convolutional blocks: the Deep Weighted Feature Convolution (DWF-Conv) and the Double Drop Convolution (D^{2}BR-Conv). DWF-Conv is used at the beginning stages of both the encoder and decoder. It leverages LKA to focus on a broader range of features and utilizes scalable ReLU to enhance feature expression, aiding in the comprehensive capture of global information and the effective restoration of the overall image structure. D^{2}BR-Conv is employed multiple times in the middle stages of the U-shaped network, utilizing DropBlock regularization to enforce the learning of more robust features. As shown in Fig. 3a, D^{2}BR-Conv consists of a DropBlock, a batch normalization (BN) layer,

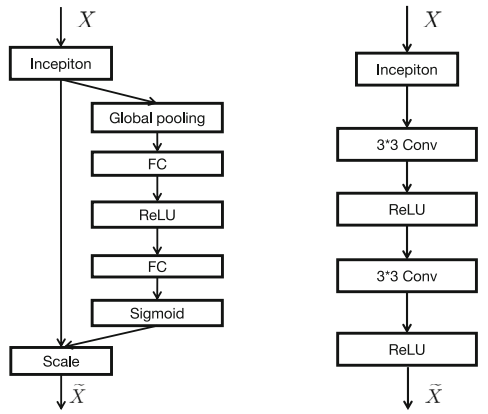

Fig. 2 The architecture of traditional convolutional block. **a** The convolution module in Squeeze-and-Excitation network. **b** The traditional convolution module in the UNet network

and a ReLU activation unit following each convolutional layer. As shown in Fig. 3b, each convolutional layer in the DWF-Conv is immediately followed by BN, a flexible ReLU unit with adjustable feature weighting parameters, and multiple deep large kernel attention (LKA) layers. This approach enhances model performance without additional computational costs by introducing large kernel convolutions and adjustable weight parameters at specific layers. Unlike the original convolutional blocks in U-Net (Fig. 2b), these structured convolutional blocks mitigate overfitting while accelerating network convergence.

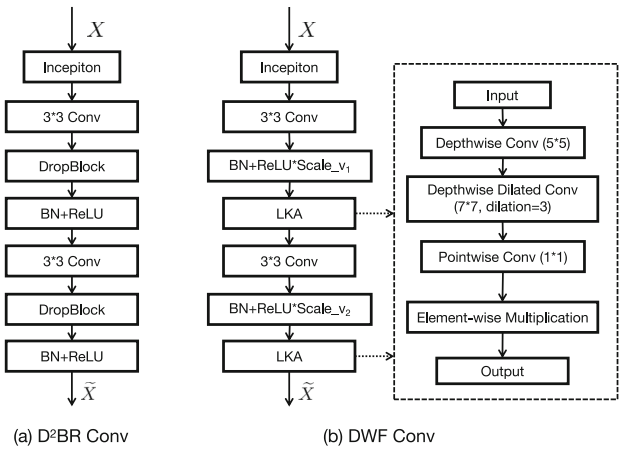

Fig. 3 The architecture of newly proposed convolutional block. **a** The convolutional block named D^2BR, which is composed of 3×3 Conv, DropBlock, BN, and ReLU. **b** A convolutional block named DWF, which is composed of a combination of 3×3 Conv, BN, LKA module, and ReLU with different scaling weight values. The LKA module contains multiple large kernel convolutions, depth convolutions, and pointwise convolutions that can expand the receptive field

2.3 RM-SViT module

To enhance the network's ability to model global context information, we propose the residual and multi-branch attention based superpixel vision transformer (RM-SViT) module, which integrates residual connections and multi-branch attention with superpixel visual transformers. Integrated between the encoder and decoder of the U-shaped network, the RM-SViT module (Fig. 4) iteratively samples visual tokens through sparse relational learning. It then applies residual multi-branch attention (RMBA) on the superpixels, merging the features before mapping them back to the original tokens.

The execution process of the RM-SViT module begins by unfolding the feature tensor F_{enc} extracted by the encoder into non-overlapping local patches, F_{unfold}, and then dividing them into initial superpixels S_0. The superpixels are initialized by averaging the features within each grid area. If the grid size is $h \times w$, the number of superpixels is given by Eq. 1:

$$m = \frac{H}{h} \times \frac{W}{w}. \tag{1}$$

This method ensures an even distribution of superpixels across the image, providing a solid starting point for iterative updates. For each iteration t, the association Q_{ij}^t between feature X_i and superpixel S_j is calculated using the following Eq. 2:

$$Q_{ij}^t = \text{Softmax}\left(\frac{X_i (S^{t-1})_j^T}{\sqrt{d}}\right). \tag{2}$$

where d is the number of channels C. Subsequently, the super token S is updated as the weighted sum of tokens, as in Eq. 3:

$$S = (\hat{Q}^t)^T X. \tag{3}$$

where $\hat{Q}^t$ is the column-normalized version of Q^t. After several iterations, multi-branch self-attention is applied to adjust the final superpixel S, capturing global context dependencies. In this Eq. 4:

$$Q = q(S), \quad K = k(S), \quad V = v(S). \tag{4}$$

Scaled dot-product attention is used to compute the attention weights, normalized by Softmax, and then a weighted sum of values V is performed along the last dimension, as in Eq. 5:

$$\text{Attn}(S) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{5}$$

The result of the weighted sum is then projected through a convolutional layer and added to the residual connection. The

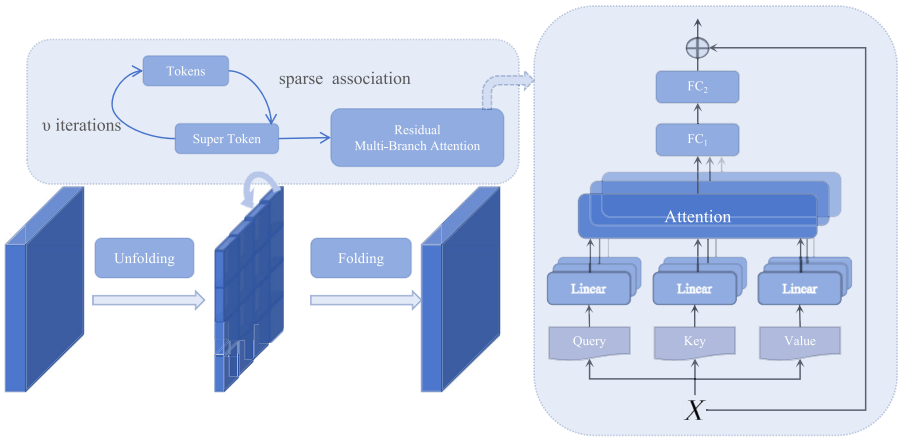

Fig. 4 The architecture of RM-SViT module. The encoder expands the feature tensor, divides "Tokens" into "Super tokens" by sparse association learning, then adjusts the final "Super Token" by applying multi-branch self-attention based on residual connection after corresponding rounds of iteration, and finally maps the expanded local block back to the original Token space

output is finally obtained by combining the adjusted features with the residual connection, as in Eq. 6:

$$\text{output} = \text{LayerNorm}(\text{Conv2d}(\text{Attn}(S)) + \text{residual}). \quad (6)$$

2.4 S^2-MLP Link module

The skip connections enhance the transmission of information between multi-scale feature maps, thereby improving the feature-capturing ability of deep networks [36]. To further gain spatial perception across different dimensions and understand complex positional relationships, we introduce a multi-dimensional spatial connector at the skip connections, namely the spatial-shift mlp (S^2-MLP Link) module. Spatial shift mlp methods [37, 38] combine the inductive bias advantages of MLPs with spatial shifting to enable patch communication, thus achieving high recognition accuracy.

As a multi-dimensional spatial connector, the S^2-MLP Link module, as shown in Fig. 5, consists of an MLP as the patch embedding layer, a spatial shifting module, and a SplitAttention module. First, the MLP$_1$ expands the feature map's channel dimension $[b, c, h, w]$ to three times its original size, splitting it into three parts (F_1, F_2, F_3). Spatial shifts are applied to F_1 and F_2, while F_3 remains unchanged. The parts are then stacked into a tensor $[b, 2, h * w, c]$. The split attention module calculates and applies attention weights to the stacked features. Finally, the MLP$_2$ restores the weighted features, producing the output feature map.

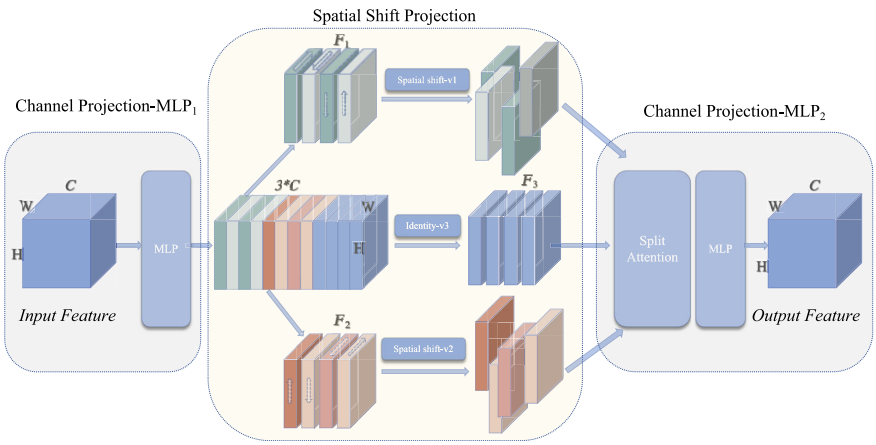

Fig. 5 The architecture of S^2-MLP Link module. Firstly, MLP is used to expand the channel c of the feature map into $3 \times c$ and divide it into three parts (F_1, F_2, F_3) along the channel dimension. F_1 and F_2 are spatially shifted according to different directions, and F_3 remains unchanged. Then, split attention is used for weighting calculation, and finally MLP is used for recovery

2.4.1 MLP layer

In vision transformers (ViT), patch embedding divides the input image into small patches, converting each into a high-dimensional embedding vector. The MLP plays a crucial role in both patch embedding and the final classification head [39]. The initial MLP rearranges the dimensions of the input feature map and expands each pixel block into a high-dimensional vector with three times the number of channels. The subsequent MLP linearly transforms the attention-enhanced feature map, restoring the original channel count.

2.4.2 Spatial shift block

In the S^2-MLP proposed by Yu et al. [37], the spatial shift concept divides the c channels into four parts, each shifted in different directions—up, down, left, and right. In the S^2-MLP Link, the spatial shift module similarly shifts different parts of the channels in various directions, enhancing the capture of spatial contextual information and improving the model's performance and generalization in complex visual tasks. In Eqs. 7 and 8, the first 1/4 of the channels are shifted left (up) by one row (column), filled with the next row's (column's) values; the next 1/4 to 1/2 of the channels are shifted right (down) by one row (column), filled with the previous row's (column's) values; the 1/2 to 3/4 of the channels are shifted up (left) by one column (row), filled with the next column's (row's) values; and the final 3/4 of the channels are shifted down (right) by one column, filled with the previous column's values. The formulas below illustrate the spatial shift operations for the different channel groups.

$$\mathrm{SS1}(x) = \begin{cases} x[:, : w-1, :, : \frac{c}{4}] & \text{if } x[:, 1:, :, : \frac{c}{4}] \\ x[:, 1:, :, \frac{c}{4} : \frac{c}{2}] & \text{if } x[:, : w-1, :, \frac{c}{4} : \frac{c}{2}] \\ x[:, :, : h-1, \frac{c}{2} : \frac{3c}{4}] & \text{if } x[:, :, 1:, \frac{c}{2} : \frac{3c}{4}] \\ x[:, :, 1:, \frac{3c}{4} :] & \text{if } x[:, :, : h-1, \frac{3c}{4} :] \end{cases} \tag{7}$$

$$\mathrm{SS2}(x) = \begin{cases} x[:, :, : h-1, : \frac{c}{4}] & \text{if } x[:, :, 1:, : \frac{c}{4}] \\ x[:, :, 1:, \frac{c}{4} : \frac{c}{2}] & \text{if } x[:, :, : h-1, \frac{c}{4} : \frac{c}{2}] \\ x[:, : w-1, :, \frac{c}{2} : \frac{3c}{4}] & \text{if } x[:, 1:, :, \frac{c}{2} : \frac{3c}{4}] \\ x[:, 1:, :, \frac{3c}{4} :] & \text{if } x[:, : w-1, :, \frac{3c}{4} :] \end{cases} \tag{8}$$

2.4.3 Split attention

Split attention is derived from the ResNest model proposed, where feature maps are finely divided, transformed, fused within groups, and then weighted and summed using attention mechanisms. A residual connection then produces the final feature map with diverse information. This paper adopts the core idea: leveraging multi-head attention and global context to perform weighted fusion on input feature maps, enhancing the diversity and accuracy of feature representation. Additionally, feature reshaping and normalization steps ensure the model's stability in complex tasks. The main processing steps of this module are as follows:

First, the input tensor x_{all} is reshaped to the form (b, k, h, w, c), where b is the batch size, k is the number of attention heads, h and w are height and width, and c is the number of channels. The tensor is further reshaped to (b, k, n, c), where $n = h \times w$, to facilitate matrix operations. According to Eq. 9, the input tensor is summed over the spatial and head dimensions and averaged to obtain the intermediate representation a:

$$a = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{n} \sum_{j=1}^{n} x_{\mathrm{all},i,j} \right) = \frac{1}{kn} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{\mathrm{all},i,j}. \tag{9}$$

The intermediate representation a is then passed through two layers of MLP and a GELU activation function to compute a higher-dimensional representation $\hat{a}$. First, as in Eq. 10, a is reduced to $c/2$ dimensions through the first MLP, activated by GELU, and then expanded to kc dimensions by the second MLP:

$$\hat{a} = \mathrm{MLP}_2(\mathrm{GELU}(\mathrm{MLP}_1(a))) \in \mathbb{R}^{(b,kc)}. \tag{10}$$

Next, as in (11), $\hat{a}$ is reshaped to (b, k, c) and normalized using the Softmax function to obtain the attention weights $\bar{a}$:

$$\bar{a} = \mathrm{Softmax}(\hat{a}) \in \mathbb{R}^{(b,k,c)}. \tag{11}$$

Once the attention weights $\bar{a}$ are obtained, they are expanded by one dimension to be element-wise multiplied with the input tensor, producing the attention-weighted matrix. The weighted feature maps are then reshaped to generate the final output feature map $\hat{X}$ as in Eqs. 12, 13, and 14:

$$\mathrm{attention} = \bar{a} \in \mathbb{R}^{(b,k,1,c)}, \tag{12}$$

$$\mathrm{out}_{ijkc} = x_{\mathrm{all},ijkc} \cdot \mathrm{attention}_{ijkc}, \tag{13}$$

$$\hat{X} = \sum_{k=1}^{K} \mathrm{out}_{ijkc} \rightarrow \hat{X} \in \mathbb{R}^{(b,h,w,c)}. \tag{14}$$

The design choices of these components are systematically validated through ablation studies (see Sect. 4.1).

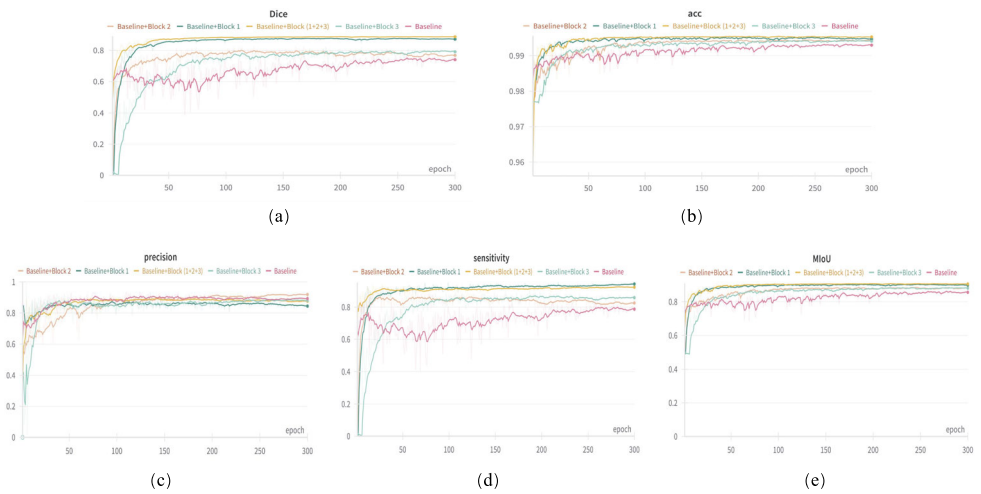

Fig. 6 Comparison results of performance. Various performance comparison results on LIDC-IDRI dataset (Baseline/+Block$_1$/+Block$_2$/+Block$_3$), The performance metrics from graph **a** to graph **e** are Dice, accuracy, precision, sensitivity, and MIoU

Specifically, (1) DWF-Conv enhances global feature extraction in early encoding stages via depth-weighted convolutions and large-kernel attention (LKA), achieving a 9.41% Dice improvement when applied alone (Table 3); (2) D^{2}BR-Conv mitigates overfitting through DropBlock regularization and dual-convolution structures, improving accuracy (Acc) by 0.15% on the EPDB dataset; (3) RM-SViT optimizes global context modeling via superpixel iteration and multi-branch attention, with a single iteration boosting Dice by 4.95%; and (4) the S^2-MLP Link module strengthens cross-level feature fusion through spatial shifts and split attention, increasing mIoU by 2.23% with only 0.8% additional parameters. These results demonstrate the synergistic contributions of all components to the overall performance.

3 Experiment and analysis

This section covers three parts: the two datasets used in the experiments, the evaluation metrics employed, and the implementation details (Fig. 6).

3.1 Datasets

We evaluated S^{3}TU-Net on two datasets: the public LIDC-IDRI [40] and the private EPDB from Shanghai Electric Power Hospital. LIDC-IDRI is a public dataset from the Lung Image Database Consortium and Image Database Resource Initiative. From 1010 distinct patients, we extracted 1303 radiologist-annotated nodules across 6474 CT slices, with strict patient-level partitioning (908 patients/5717 slices for training, 102 patients/757 slices for testing). The EPDB dataset includes 481 pathologically confirmed lung adenocarcinoma patients (AAH:112, MIA:158, IAC:142, AIS:69) diagnosed between September 2016 and October 2023, with 1198 annotated CT slices serving as an independent validation set. All annotations were verified by three experienced radiologists using consensus reading. Table 1 provides details. Preprocessing steps, including lung parenchyma segmentation, ROI extraction, and image enhancement, were applied. The final input was 128×128 patches centered on the lung nodules.

Table 1 Detailed information of LIDC-IDRI and EPDB datasets

Category	LIDC-IDRI	EPDB
Data type	Pulmonary nodules	Lung adenocarcinoma subtypes
Source	National Cancer Institute	Shanghai Electric Power Hospital
Total images	6474	1198
Patients	1010	481
Image size	128×128	128×128
Data partition		
Train	5717 (908 patients)	–
Test	757 (102 patients)	–
Validation	–	1198 (481 patients)
Label information		
Positive/negative	1303 nodules	–
Subtypes	–	AAH/MIA/IAC/AIS
Class distribution	–	
AAH	–	23%
MIA	–	32%
IAC	–	30%
AIS	–	15%

Table 2 System configuration information

Configuration name	Details
Operating system	Linux ai 5.15.0-86-generic
CPU model	Intel(R) Core(TM) i7-10700K CPU @ 3.80GHz
GPU model	NVIDIA GeForce RTX 3080
Memory information	20GB
Python version	3.8.10
Cuda version	12.2
PyTorch version	1.13.1

3.2 Evaluation metrics

This study assesses segmentation performance using common metrics in medical image segmentation: Dice similarity coefficient (DSC), accuracy (Acc), sensitivity (Sen), precision (Pre), and supplementary metrics like IoU or AUC. DSC measures the similarity between the predicted mask and the ground truth, Acc measures the percentage of correctly classified pixels, Sen evaluates the model's ability to detect true positives, and Pre indicates the proportion of true positives among predicted positives. A and B represent the prediction and ground truth; N is the number of pixels in the 3D patch; p_i is the predicted probability for pixel i; and g_i is the ground truth label for pixel i. Equations 15, 16, and 17 are as follows:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|} = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}, \tag{15}$$

$$\text{Sensitivity} = \frac{|A \cap B|}{|B|} = \frac{\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} g_i}, \tag{16}$$

$$\text{Precision} = \frac{|A \cap B|}{|A|} = \frac{\sum_{i=1}^{N} p_i \cdot g_i}{\sum_{i=1}^{N} p_i}. \tag{17}$$

Table 3 Performance comparison on LIDC-IDRI dataset (Max values from 5 independent runs, reflecting optimal performance under ablation settings)

Methods	DSC %	Acc %	mIoU %	Pre %	Sen %
Baseline	77.44	99.33	86.41	93.14	86.94
+ Block1	86.85	99.52	90.31	87.40	**95.35**
+ Block2	82.39	99.41	88.53	**93.25**	88.56
+ Block3	82.35	99.40	88.64	90.34	90.53
+ ALL (S^3TU-Net)	**89.04**	**99.53**	**90.70**	90.73	93.70

The bold fonts indicate the best overall results in this table

3.3 Implementation details

The dataset was divided into training, testing, and independent validation sets to more accurately evaluate the model's segmentation performance and generalization. For the specific dataset partitioning, please refer to Section 3.1. The LIDC dataset was split with a nearly 9:1 ratio, with 5717 images for training and 757 for testing. The EPDB dataset's 1198 images were used as an independent validation set to objectively assess the model's generalization and prepare for subsequent lung adenocarcinoma nodule classification. We have employed conventional data augmentation techniques during the training phase, including random horizontal flipping (with a probability of 50%) and random rotation (with an angle range of $\pm 15°$), to enhance the model's robustness against variations in nodule morphology. These operations are used in conjunction with DropBlock (with a block size of 7), and the two are optimized respectively for the global data distribution and local feature robustness.The S^3TU-Net model was trained using the Adam optimizer with a combined binary cross-entropy and Dice loss function. The initial learning rate was set at 0.001, dynamically adjusted using a scheduler, with one warm-up epoch. The batch size was 16, with 300 epochs, and the DropBlock size was set to 7. The implementation was based on the PyTorch framework, with the hardware and software configurations detailed in Table 2.

4 Experimental results

4.1 Ablation study

To demonstrate that each component of the proposed S^3TU-Net enhances lung nodule segmentation performance, we conducted ablation experiments on the LIDC-IDRI dataset and validated the results on the EPDB dataset. Using the UNet model as the baseline, we evaluated the segmentation performance of Baseline+Block_1 (structured convolution), Baseline+Block_2 (RM-SViT), and Baseline+Block_3 (S^2-MLP Link) on both datasets, as shown in Tables 3 and 4. Figure 6 presents the corresponding performance compari-

Table 4 Performance metrics on EPDB dataset (average)

Methods	DSC %	Acc %	AUC %	Pre %	Sen %
+ Block1	85.01	98.38	89.04	94.54	78.36
+ Block2	81.55	98.14	86.30	95.49	72.85
+ Block3	83.13	98.44	86.94	**96.38**	74.04
+ ALL (S^3TU-Net)	**86.40**	**98.53**	**90.00**	94.79	**80.27**

The bold fonts indicate the best overall results in this table

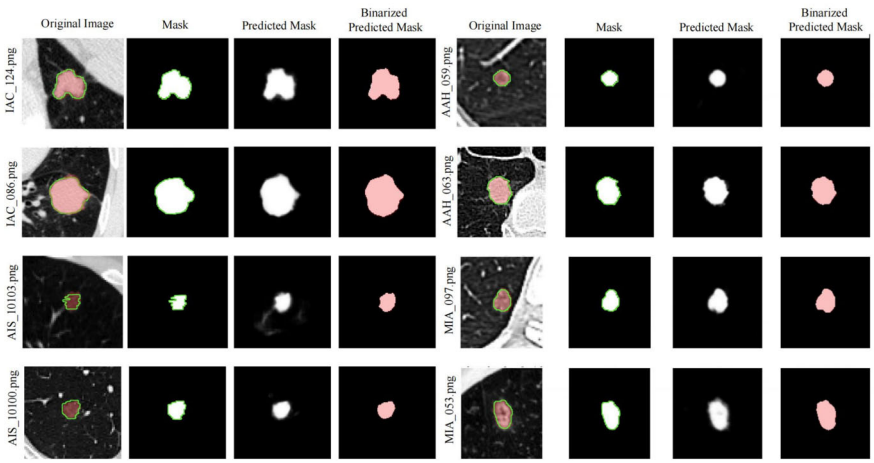

Fig. 7 The example images from the EPDB dataset. Randomly shown are the original images, annotated masks, segmentation results, and binarized segmentation results of stage IV lung adenocarcinoma (AAH/MIA/IAC/AIS)

son on the LIDC-IDRI dataset. Since the segmentation task is a precursor to lung adenocarcinoma classification, which involves various irregular nodule shapes, the EPDB dataset was used to test the model's generalization ability. Figure 7 shows the segmentation results of the four types of lung adenocarcinoma (AAH, MIA, IAC, AIS) on sample nodules from the EPDB dataset.

4.1.1 Baseline+Block$_1$

When replacing traditional convolutions with two types of structured convolutions, performance improvements were most notable. The DSC reached 86.85%, mIoU was 90.31%, and sensitivity peaked at 95.35%. Compared with the baseline model, the DSC and sensitivity are increased by 9.41%. This demonstrates the effectiveness of constructing the backbone using novel structured convolutional blocks with distinct functions (Fig 8).

4.1.2 Baseline+Block$_2$

RM-SViT was developed to address limitations in capturing long-range dependencies and global context during local feature extraction. Experiments showed that the size of the super token coverage (stoken_size) and the number of iteration updates strongly affect segmentation performance and computational complexity. A smaller grid size (8×8) captures local features more precisely but increases computational cost, while a larger grid size (16×16) focuses on global features and reduces complexity. Higher iteration counts capture more complex relationships between image patches but significantly increase processing time. Without iterations,

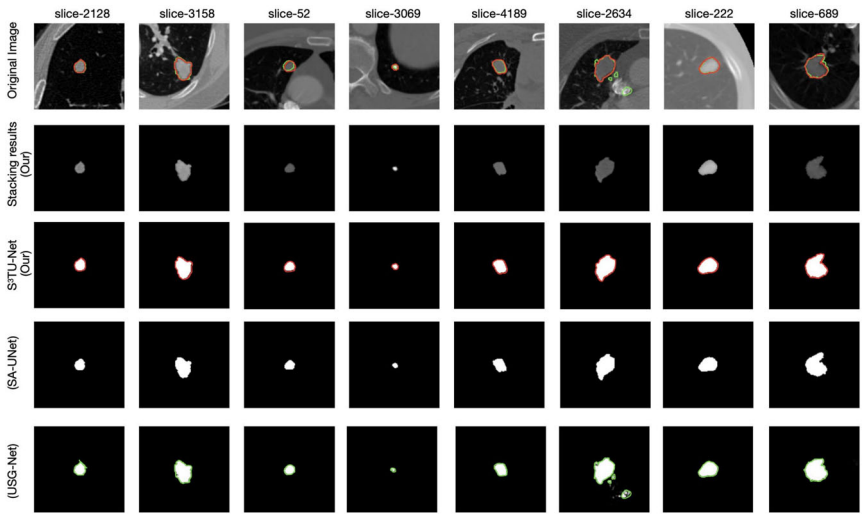

Fig. 8 Qualitative comparison (data were randomly extracted from LIDC-IDRI). (1) Original image. (2) and (3) Segmentation results and display results of S^{3}TU-Net. (4) and (5) The segmentation results of the comparison model

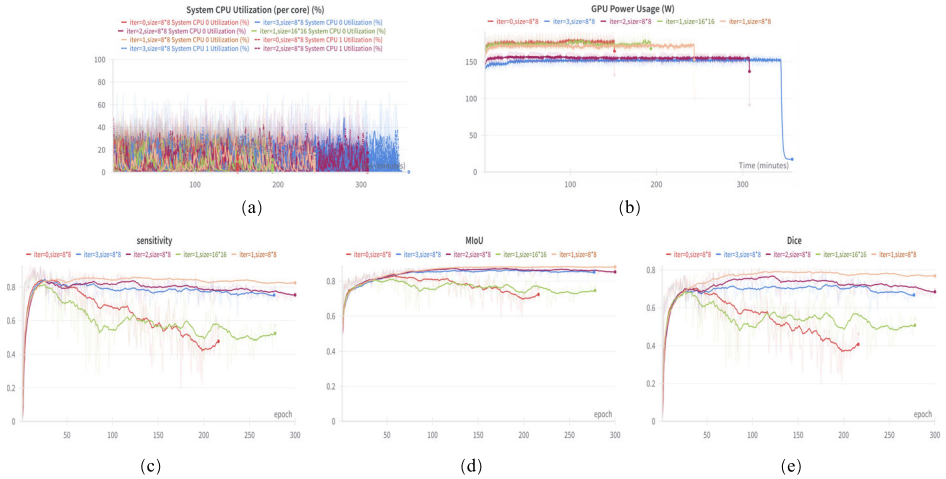

Fig. 9 Performance comparison on Block$_2$. iter=0/1/2/3 and size=8×8/16×16. **a** System CPU utilization. **b** GPU power consumption. **c** Sensitivity. **d** MIoU. **e** Dice. Left to right, top to bottom: CPU/GPU usage and GPU energy consumption data were obtained using PyTorch Profiler for GPU utilization, kernel time, and memory; NVML via pynvml for real-time GPU power; and Linux perf with psutil for CPU metrics. Per NVIDIA docs, GPU power accuracy error <5%, and CPU data was sampled every 100 ms to balance granularity and overhead

performance drops due to potential misalignment of super tokens across different semantic regions. Figure 9 shows the performance comparison results, including the system CPU utilization, GPU power consumption, sensitivity, MIoU, and Dice. The final conclusion is that a single iteration with a smaller grid size is best suited for processing 128×128 input images. Excessive or insufficient iterations affect Dice accuracy, wasting computation time despite low GPU energy consumption. An 8×8 grid size outperforms 16×16 in Dice, mIoU, and sensitivity under similar GPU and CPU consumption.

4.1.3 Baseline+Block$_3$

The S^2-MLP Link optimizes skip connections by leveraging spatial information to enhance feature fusion and aid gradient flow. Compared to baseline, with only a slight increase in parameters, it achieves a DSC of 82.35%, an accuracy of 99.40%, mIoU of 88.64%, and sensitivity of 90.53%. This further demonstrates that MLP-based visual architectures combined with spatial shifts can achieve higher performance with less inductive bias.

4.2 Comparative experiments

To evaluate the segmentation performance of S^{3}TU-Net, this study compares it with several advanced and commonly used open source methods for lung nodule segmentation, including YNet [41], UNet++ [42], R2U-Net [43], Attention-UNet [44], UNet3++ [45], USG-Net [47], and SA-UNet [46]. Table 5 presents the results on the LIDC-IDRI dataset. The best-performing model, USG-Net, achieved a DSC of 84.52%, followed by SA-UNet with a DSC of 84.35%. In contrast, the proposed S^{3}TU-Net model excels across all metrics, with a maximum DSC of 88.87%, MIoU of 91.14%, sensitivity of 93.48%, and precision of 91.97%, while its accuracy is comparable to that of other methods. Compared to the top-ranked models, USG-Net and SA-UNet, our model outperforms them in DSC by 4.52% and 4.69%, in mIoU by 5.26% and 2.0%, and in sensitivity by 6.82% and 3.16%, respectively. Table 6 shows the computational complexity and performance comparison. According to the results, compared with other models, S^{3}TU-Net demonstrates an excellent efficiency-performance balance: (1) The number of parameters (2.87 million) is lower than that of all baseline models (for example, UNet++ has 3.66 million parameters); (2) the amount of floating-point operations is reduced (59.8 billion times, while UNet3++ has 195.8 billion times); (3) the inference speed (38.4 ms) is faster than that of the real-

Table 5 Performance comparison of different methods

Methods	DSC %	Acc %	mIoU %	Pre %	Sen %
Y-Net [41]	79.53	92.39	83.71	83.77	78.54
UNet++ [42]	82.27	96.28	85.37	82.93	84.98
R2U-Net [43]	82.10	98.11	76.18	84.00	86.33
Attention-UNet [44]	80.21	98.78	82.84	83.78	86.91
UNet3++ [45]	81.34	99.80	83.64	81.14	84.50
SA-UNet [46]	84.35	99.24	88.70	88.20	90.54
USG-Net [47]	84.52	99.02	85.44	86.78	86.88
S^{3}TU-Net	89.04	99.53	90.70	90.73	93.70

*Note: The red, blue and green rows represent the 1st, 2nd, 3rd places, respectively

Table 6 Computational complexity and performance comparison

Model	Params (M)	FLOPs (G)	Time (ms)	Dice (%)
Y-Net [41]	28.5	58.2	35.9	79.53
U-Net [22, 23]	31.0	65.3	42.7	82.27
UNet++ [42]	36.6	161.2	68.9	84.35
R2U-Net [43]	39.2	183.5	72.4	82.10
Attention-UNet [44]	34.8	89.7	55.3	80.21
UNet3++ [45]	40.1	195.8	76.1	81.34
SA-UNet [46]	34.1	78.5	53.2	84.52
USG-Net [47]	33.7	74.6	49.8	84.52
S^3TU-Net	28.7	59.8	38.4	89.04

time-oriented Y-Net (35.9 ms), and the Dice coefficient is 9.51% higher than that of Y-Net. This is attributed to the parameter-efficient D^2BR-Conv module, the sparse attention mechanism in RM-SViT, and the low-cost spatial fusion of S^2 Link.

To facilitate intuitive visualization, eight nodule images were randomly selected from the LIDC-IDRI dataset. Figure 8 presents the segmentation results of the two models with the highest DSC and the S^3TU-Net model. Compared with the classical methods and the more advanced models, our S^3TU-Net model, which combines SViT with spatial interaction mechanisms, consistently achieves smoother and more accurate segmentation results.

5 Conclusion

Lung adenocarcinoma nodules in CT images often exhibit irregular and complex characteristics, posing significant challenges for accurate staging. Precise segmentation is essential for clinicians to focus on critical regions of interest and derive reliable diagnostic insights. While U-Net and its variants have demonstrated strong performance on conventional CT images, their generalization capability diminishes when handling complex adenocarcinoma nodules. To address this, we propose S^3TU-Net, a CNN-Transformer hybrid model that integrates structured convolution blocks, residual-based superpixel visual transformers (RM-SViT), and multi-dimensional spatial connectors (S^2-MLP Link) to enhance feature extraction, feature fusion, and global context understanding. Key components such as DWF-Conv and D^2BR-Conv blocks improve global information representation and mitigate overfitting, while RM-SViT captures long-range dependencies efficiently through sparse correlation and multi-branch attention.Although the S^3TU-Net has demonstrated excellent performance on both the LIDC-IDRI and EPDB datasets, we are also aware of the potential limitations in data diversity and annotation consistency. To address the possible bias issues, as described above, we have implemented the following strategies during the training process: (1) Operations such as random rotation, flipping, and intensity variation have been applied to mitigate the overfitting problem and enhance the generalization ability of the model on nodule data with different morphologies. (2) By adjusting the batch weights, we have ensured the balance of sampling in the EPDB dataset for the four subtypes of adenocarcinoma (AAH/MIA/IAC/AIS). (3) Internal five-fold cross-validation has been conducted on the LIDC-IDRI dataset to verify the stability of the model. Although the EPDB dataset comes from a single center, the nodules in it have been annotated by three radiologists, and the majority voting method has been adopted to reduce the differences among observers. These measures further verify the robustness of the model in dealing with institutional biases. Furthermore, the S^2-MLP Link module facilitates multi-scale feature transmission and minimizes information loss. On the LIDC-IDRI dataset, S^3TU-Net achieved a DSC of 89.04%, precision of 90.73%, IoU of 90.70%, and sensitivity of 93.70%. On the EPDB dataset, it achieved a DSC of 86.40% and an accuracy of 98.53%, demonstrating robust segmentation performance and strong generalization ability.

6 Future work

Future research will focus on developing lightweight hybrid architectures to reduce computational overhead and improve inference speed, enabling real-time processing on resource-constrained devices such as portable medical equipment. Currently, we are expanding the applicability of S^3TU-Net to other medical imaging modalities (such as ultrasound and magnetic resonance imaging) and various disease segmentation tasks (such as skin diseases and prostate cancer). By leveraging its ability to capture fine-grained local details and comprehensive global semantics, we aim to enhance segmentation accuracy and broaden its application scope within the medical field. The research will be carried out in the following four directions: (1) optimize computational efficiency through lightweight architectures (such as pruning redundant parameters and using factorized attention mechanisms) to enable real-time processing on portable devices; (2) extend the model to the field of 3D segmentation by replacing 2D convolutions with 3D kernels and integrating voxel-level attention in RM-SViT, and conduct further validation on multi-modal imaging (such as positron emission tomography-computed tomography fusion (PET-CT fusion)); (3) address the scarcity of annotated data through semi-supervised strategies, including consistency regularization using unlabeled data (with the help of the "Mean Teacher" algorithm) and iterative generation of pseudo-labels; (4) enhance the interpretability of the model by

visualizing the attention maps in RM-SViT and analyzing the contributions of each component with the aid of saliency maps. These efforts aim to improve the scalability, generalization ability, and clinical transparency of the model, ultimately bridging the gap between technological innovation and practical deployment. In addition, to address the interpretability challenges posed by the multi-component architecture of the model, we plan to study methods for visualizing the attention mechanisms and feature importance in the structured convolution blocks (DWF-Conv/D^2BR-Conv), RM-SViT, and S^2-MLP Link modules. This includes analyzing the contribution of each component to segmentation performance and designing saliency maps to highlight the key regions in CT images that influence the model's decisions. This will help improve the interpretability of the model and make it more suitable for clinical applications.

Author Contributions Contributions of the authors have been clearly stated and ranked according to authorship to reflect their contributions.

Funding No funding

Data Availability All data, materials, and codes are available.

Declarations

Ethics Approval and Consent to Participate No ethical approval or consent required for participation.

Consent for Publication All authors have given their published consent.

Conflict of Interest No conflict of interest.

References

1. Chen H-Y, Wang H-M, Lin C-H, Yang R, Lee C-C (2023) Lung cancer prediction using electronic claims records: a transformer-based approach. IEEE J Biomed Health Inform
2. Xiang D, Zhang B, Lu Y, Deng S (2022) Modality-specific segmentation network for lung tumor segmentation in PET-CT images. IEEE J Biomed Health Inform 27(3):1237–1248
3. Li Z, Zhang J, Tan T, Teng X, Sun X, Zhao H, Liu L, Xiao Y, Lee B, Li Y et al (2020) Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the ACDC@ LungHP Challenge 2019. IEEE J Biomed Health Inform 25(2):429–440
4. Hutchinson BD, Shroff GS, Truong MT, Ko JP (2019) Spectrum of lung adenocarcinoma. In: Seminars in ultrasound, CT and MRI, vol 40. Elsevier, pp 255–264
5. Myers DJ, Wallen JM (2023) Lung adenocarcinoma. StatPearls [Internet]
6. Borczuk AC (2016) Prognostic considerations of the new world health organization classification of lung adenocarcinoma. Eur Respir Rev 25(142):364–371
7. Organization WH et al (2015) Who classification of tumours of the lung, pleura, thymus and heart. WHO/IARC Class Tumours 7
8. Shao X, Niu R, Jiang Z, Shao X, Wang Y (2020) Role of PET/CT in management of early lung adenocarcinoma. Am J Roentgenol 214(2):437–445
9. Cohen J, Reymond E, Jankowski A, Brambilla E, Arbib F, Lantuejoul S, Ferretti G (2016) Lung adenocarcinomas: correlation of computed tomography and pathology findings. Diagn Interv Imaging 97(10):955–963
10. Liu W, Liu X, Li H, Li M, Zhao X, Zhu Z (2021) Integrating lung parenchyma segmentation and nodule detection with deep multi-task learning. IEEE J Biomed Health Inform 25(8):3073–3081
11. Shen F, Ye H, Zhang J, Wang C, Han X, Wei Y (2023) Advancing pose-guided image synthesis with progressive conditional diffusion models. In: The twelfth international conference on learning representations
12. Shen F, Ye H, Liu S, Zhang J, Wang C, Han X, Yang W (2024) Boosting consistency in story visualization with rich-contextual conditional diffusion models. arXiv:2407.02482
13. Zhang J, Xia Y, Cui H, Zhang Y (2018) Pulmonary nodule detection in medical images: a survey. Biomed Signal Process Control 43:138–147
14. Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, Goldgof DB (2016) Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography 2(4):388
15. Hu J, Huang Z, Shen F, He D, Xian Q (2023) A bag of tricks for fine-grained roof extraction. In: IGARSS 2023-2023 IEEE international geoscience and remote sensing symposium. IEEE
16. Liu H, Geng F, Guo Q, Zhang C, Zhang C (2018) A fast weak-supervised pulmonary nodule segmentation method based on modified self-adaptive FCM algorithm. Soft Comput 22:3983–3995
17. Xie H, Yang D, Sun N, Chen Z, Zhang Y (2019) Automated pulmonary nodule detection in CT images using deep convolutional neural networks. Pattern Recogn 85:109–119
18. Jain S, Choudhari P, Gour M (2023) Pulmonary lung nodule detection from computed tomography images using two-stage convolutional neural network. Comput J 66(4):785–795
19. Hu J, Huang Z, Shen F, He D, Xian Q (2023) A robust method for roof extraction and height estimation. In: IGARSS 2023-2023 IEEE international geoscience and remote sensing symposium. IEEE
20. Tavakoli MB, Orooji M, Teimouri M, Shahabifar R (2020) Segmentation of the pulmonary nodule and the attached vessels in the CT scan of the chest using morphological features and topological skeleton of the nodule. IET Image Proc 14(8):1520–1528
21. Zhang Y, Chung F-L, Wang S (2020) Clustering by transmission learning from data density to label manifold with statistical diffusion. Knowl-Based Syst 193:105330
22. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp 234–241
23. Krithika Alias AnbuDevi M, Suganthi K (2022) Review of semantic segmentation of medical images using modified architectures of unet. Diagnostics. 12(12):3064
24. Liu H, Cao H, Song E, Ma G, Xu X, Jin R, Jin Y, Hung C-C (2019) A cascaded dual-pathway residual network for lung nodule segmentation in CT images. Physica Med 63:112–121
25. Lu D, Chu J, Zhao R, Zhang Y, Tian G (2022) A novel deep learning network and its application for pulmonary nodule segmentation. Comput Intell Neurosci 2022(1):7124902
26. Hou T, Zhao J, Qiang Y, Wang S, Wang P (2020) Pulmonary nodules segmentation based on CRF 3D-UNet structure. Comput Eng Des 41(6):1663–1669
27. Arregui García X (2023) ViTs vs. CNNs for 3D medical image segmentation: are transformers all you need? Master's thesis
28. Amanatiadis A, Kaburlasos VG, Kosmatopoulos EB (2018) Understanding deep convolutional networks through gestalt theory. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE, pp 1–6

29. Dosovitskiy A (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
30. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 10012–10022
31. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
32. Ghiasi G, Lin T-Y, Le QV (2018) Dropblock: a regularization method for convolutional networks. Adv Neural Inf Process Syst 31
33. Guo M-H, Lu C-Z, Liu Z-N, Cheng M-M, Hu S-M (2023) Visual attention network. Comput Vis Media 9(4):733–752
34. Si C, Huang Z, Jiang Y, Liu Z (2024) Freeu: free lunch in diffusion u-net. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 4733–4743
35. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7132–7141
36. Ruan J, Xiang S, Xie M, Liu T, Fu Y (2022) Malunet: a multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp 1150–1156
37. Yu T, Li X, Cai Y, Sun M, Li P (2022) S2-mlp: Spatial-shift mlp architecture for vision. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 297–306
38. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J et al (2021) Mlp-mixer: an all-mlp architecture for vision. Adv Neural Inf Process Syst 34:24261–24272
39. Taud H, Mas J-F (2018) Multilayer perceptron (mlp). Geomatic approaches for modeling land change scenarios. pp 451–455
40. Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, He T, Mueller J, Manmatha R, et al. (2022) Resnest: split-attention networks. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp 2736–2746
41. Mehta S, Mercan E, Bartlett J, Weaver D, Elmore JG, Shapiro L (2018) Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer, pp 893–901
42. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp 3–11
43. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK (2018) Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv:1802.06955
44. Hu J, Huang Z, Shen F, He D, Xian Q (2023) A bag of tricks for fine-grained roof extraction. IEEE
45. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J (2020) Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1055–1059
46. Guo C, Szemenyei M, Yi Y, Wang W, Chen B, Fan C (2021) Sa-unet: spatial attention u-net for retinal vessel segmentation. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp 1236–1242
47. Yang H, Shen L, Zhang M, Wang Q (2022) Uncertainty-guided lung nodule segmentation with feature-aware attention. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 44–54

Yuke Wu Yuke Wu is a postgraduate student at Shanghai University of Engineering Science, majoring in Medical Imaging. Her research mainly focuses on CT images of lung adenocarcinoma. She has published an undergraduate - level paper on a medical system integrating artificial intelligence and medical image processing. Additionally, she is the second author of a SCI - indexed paper in the second quartile on the research of prostate magnetic resonance imaging (MRI) images.

Yunyu Shi Yunyu Shi obtained her Bachelor's and Master's degrees in Computer Science from East China University of Technology in 2004 and 2007 respectively. In 2012, she received her Ph.D. in Computer Applications from Shanghai University. From 2012 to 2014, she conducted post - doctoral research at Shanghai Jiao Tong University. Currently, she is a lecturer at the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. Her research interests include image and video processing and analysis, video compression, and quality assessment.

Xinyi Chen Xinyi Chen is a master's student at Shanghai University of Engineering Science, majoring in Medical Imaging. She/He has published an SCI paper on prostate cancer.

Zhenglei Wang Zhenglei Wang is currently a chief physician in the Department of Medical Imaging at Shanghai Electric Power Hospital. He is engaged in the diagnosis, teaching, and research work of CT and MRI. He is proficient in radiological diagnosis of various systems of the whole body, such as abdominal diseases. He has published more than ten academic papers.

Yuqing Xu Yuqing Xu, a master's degree holder and a senior engineer, works at Shanghai Ideal Information Industry (Group) Co., Ltd. His/Her research area focuses on cloud computing.

Shuohong Wang Shuohong Wang received her Ph.D. degree from Fudan University, China, in 2017 and her B.Eng. degree from East China University of Science and Technology, China, in 2012. She is now a research associate at Harvard University. Her research interests include biomedical image analysis, multi-object tracking and machine learning.

Xiang Liu Xiang Liu, a Doctor of Science from Fudan University, is currently an associate professor at Shanghai University of Engineering Science. His research focuses on computer vision and artificial life, covering areas such as medical image analysis and multi - robot collaboration. He has presided over several national - level and cultural - field research projects. His research achievements have won multiple awards, including those from Shanghai and the automotive industry. He holds two domestic patents and one international patent, and has published four SCI papers and six CCF Class B conference papers. His research findings are widely applied in fields like smart cities and industrial inspection.

Authors and Affiliations

Yuke Wu[1] · Xiang Liu[1] · Yunyu Shi[1] · Xinyi Chen[1] · Zhenglei Wang[2] · YuQing Xu[3] · ShuoHong Wang[4]

✉ Xiang Liu
xliu@sues.edu.cn

Yuke Wu
M325122219@sues.edu.cn

Yunyu Shi
yunyushi@sues.edu.cn

Xinyi Chen
c2257873708@163.com

Zhenglei Wang
hanqi_willis@163.com

YuQing Xu
xuyuqing.sh@chinatelecom.cn

ShuoHong Wang
wangsh@fas.harvard.edu

[1] The College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201600, China

[2] The Department of Medical Imaging, Shanghai Electric Power Hospital, Shanghai 200000, China

[3] The Department of Cloud Network, Shanghai IDEAL INFORMATION Industry Co. LTD, Shanghai 200000, China

[4] The Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University, Cambridge MA, USA