Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 1, No. 1, 2026

Liver Tumor Segmentation via State Space Modeling and Multi-Scale Detail Enhancement

Yuhang Feng

Henan Polytechnic University, Jiaozuo 454000, China

Abstract: Accurate liver tumor segmentation from CT images is challenging because tumors present blurred boundaries, irregular shapes, and substantial scale variation. To address these issues, we propose a liver tumor segmentation network that combines state-space modeling with multi-scale detail enhancement (SSMDENet). SSMDENet follows an encoder-decoder architecture and introduces a Global-Local Modeling (GLM) block as the basic feature extractor. Within GLM, the Spatial Dependency Modeling (SMD) module captures long-range spatial dependencies to encode global anatomical context, while the Multi-scale Feature Enhancement (MFE) module uses parallel depthwise convolutions and channel recalibration to strengthen local boundary and texture information. In this way, the proposed network jointly models global semantics and local details. Experiments on the LiTS2017 and 3DIRCADB-01 datasets demonstrate the effectiveness of the method. SSMDENet achieves Dice scores of 85.79% and 84.12% on LiTS2017 and 3DIRCADB-01, respectively, outperforming several representative segmentation methods. Ablation studies further confirm the complementary benefits of the SMD and MFE modules. These results indicate that SSMDENet is an effective and robust solution for liver tumor segmentation.

Keywords: Liver Tumor Segmentation; State-space Model; Multi-scale Feature Enhancement.

1. Introduction

Accurate delineation of liver tumors from CT images is of great clinical value for diagnosis, treatment planning, and treatment-response assessment [1-3]. Nevertheless, automatic liver tumor segmentation remains difficult because tumors often exhibit blurred boundaries, large scale variation, irregular shapes, and low contrast with surrounding liver parenchyma. Encoder-decoder architectures represented by U-Net [4] have achieved strong performance in medical image segmentation, and many variants, such as U-Net++ [5], Attention U-Net [6], and CE-Net [7], further improve feature aggregation and contextual modeling. However, most convolution-dominant models still focus on local neighborhood interactions and therefore have limited ability to explicitly model long-range dependencies, which can reduce performance on small lesions, complex boundaries, and globally ambiguous cases.

Recent advances in state-space models have opened a new direction for long-range dependency modeling in vision [8-11]. Compared with self-attention, state-space modeling can capture long-range interactions with lower computational complexity, which is especially attractive for high-resolution medical features. For liver tumor segmentation, global context helps the network understand the relationship between a lesion and the surrounding liver anatomy, thereby improving localization in challenging cases. Nevertheless, global modeling alone is insufficient for recovering fine boundaries and local multi-scale structures, particularly for small lesions, weak-contrast margins, and morphologically complex tumors. A practical model should therefore couple efficient global dependency modeling with effective local detail enhancement.

Motivated by these observations, we propose State Space and Multi-scale Detail Enhancement Network (SSMDENet) for liver tumor segmentation. The network follows an encoder-decoder pipeline and uses a Global-Local Modeling (GLM) block as the basic feature extractor. Within each GLM block, the Spatial Dependency Modeling (SMD) module captures long-range spatial dependencies to strengthen global contextual perception, while the Multi-scale Feature Enhancement (MFE) module complements this representation with local multi-scale detail enhancement. By integrating global dependency modeling with local detail recovery, SSMDENet improves both semantic understanding and boundary reconstruction, leading to more accurate and robust segmentation.

2. Related Work

Research on liver and liver-tumor segmentation has evolved from pure convolutional architectures to hybrid global-context models. Early encoder-decoder networks, especially U-Net, established the dominant paradigm for medical image segmentation. Later variants such as UNet++, Attention U-Net, CE-Net, and H-DenseUNet [12] improved skip connections, attention mechanisms, contextual aggregation, or hybrid 2D/3D feature fusion, and achieved better performance on liver CT data. More recently, Transformer-based models such as TransUNet[13], UCTransNet [14], and Swin-Unet [15] introduced stronger global interaction modeling and achieved competitive results. However, Transformer-based segmentation often incurs substantial computational overhead, while purely convolutional models may still lack sufficiently strong long-range modeling.

State-space models provide an efficient alternative for long-range representation learning. Mamba and its vision extensions, such as VMamba, show that scan-based state-space operators can model global dependencies with linear complexity. These ideas have recently been introduced into medical imaging through frameworks such as VM-UNet and SegMamba, demonstrating the promise of state-space modeling for dense prediction tasks. Nevertheless, many existing state-space segmentation models mainly emphasize efficient global context encoding, while the restoration of local boundaries and multi-scale lesion details remains

comparatively underexplored. For liver tumor segmentation, where lesion size and appearance vary substantially, this gap is particularly important. Our SSMDENet addresses it by explicitly coupling state-space-based global modeling with multi-scale local enhancement.

3. Method

(1) Overall Architecture of SSMDENet

As illustrated in Figure 1, SSMDENet adopts a typical encoder-decoder architecture consisting of a shallow feature extractor, an encoder, a decoder, skip connections, and an output projection layer. Different from conventional U-Net-like networks, our backbone uses the GLM block as the core modeling unit so that global context modeling and local detail enhancement can be performed within the same hierarchical feature extraction process. This design is well suited to liver tumor segmentation, where targets exhibit large scale variation, blurry boundaries, and complex morphology.

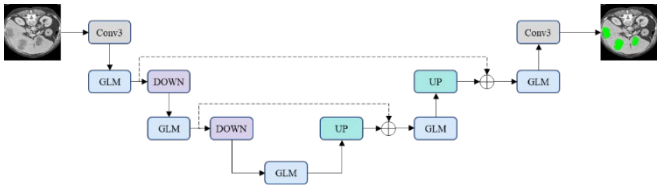

Figure 1. Overall architecture of SSMDENet.

Given an input CT slice, the network first applies a 3×3 convolution to obtain shallow features containing basic texture, edge, and intensity variation information. These shallow features are then fed into the encoder for progressive semantic abstraction. During encoding, GLM blocks and downsampling operations are stacked alternately to perform multi-level feature modeling while gradually enlarging the receptive field. As the network goes deeper, spatial resolution decreases and channel dimensionality increases, enabling deeper features to encode richer semantics and stronger contextual information for multi-scale representation of both the liver and tumor regions.

In the decoder, upsampling is used to progressively recover spatial resolution. The current decoding features are aligned with the corresponding high-resolution encoder features through skip connections, and the two branches are fused by element-wise addition to compensate for spatial details lost during downsampling. The fused features are then refined by another GLM block to strengthen the representation of tumor boundaries, local texture, small lesions, and irregular structures. After multi-stage decoding, an output projection layer maps the high-dimensional feature representation to the final tumor segmentation mask.

(2) Global-Local Joint Modeling Module

To further improve feature representation in the backbone, the GLM block is inserted at multiple stages of both the encoder and decoder. As shown in Figure 2, this block contains two serial submodules. The first is SMD, which models long-range spatial dependencies in 2D medical images and enhances the perception of global anatomy and lesion context. The second is MFE, which uses multi-scale convolution and feature recalibration to supplement local details and improve the segmentation of blurred boundaries, scale variation, and irregular tumor regions. Both submodules are embedded with residual connections, which stabilize feature propagation, facilitate optimization, and improve robustness. This "global modeling first, local enhancement second" strategy enables SSMDENet to integrate global semantics and local details more effectively than conventional convolution-only segmentation blocks.

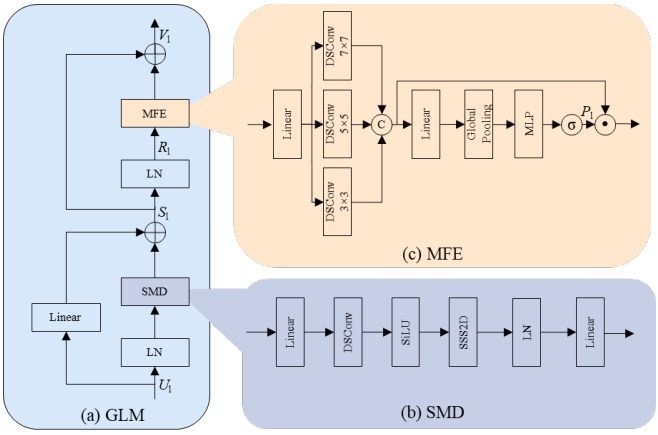

Figure 2. Structure of the GLM module

Let the input feature to the l-th GLM block be $U_l \in R^{H_l \times W_l \times C_l}$ where H_l, W_l, and C_l denote the height, width, and channel dimension, respectively. GLM first performs two-branch processing. One branch applies a linear projection to produce a base representation, while the other branch normalizes the input and feeds it into the SMD module for spatial dependency modeling. The outputs of the two branches are added element-wise to obtain an intermediate feature. This feature is then normalized and sent to the MFE module for local multi-scale enhancement, and finally combined with the residual connection to generate the output of the GLM block. The procedure can be expressed as follows in Eq. (1) and Eq. (2).

$$S_l = \mathrm{Linear}(U_l) + \mathrm{SMD}\big(\mathrm{LN}(U_l)\big) \quad (1)$$

$$V_l = S_l + \mathrm{MFE}\big(\mathrm{LN}(S_l)\big) \quad (2)$$

Through this design, GLM completes global contextual modeling and local detail enhancement within a single feature extraction unit, thereby improving discrimination of liver tumor regions. Compared with a standard convolutional block, GLM can not only establish long-range spatial dependencies, but also strengthen local information such as boundaries, textures, and scale variation, making it more suitable for liver tumor segmentation with blurred margins and irregular shapes.

(3) Spatial Dependency Modeling Module

The Spatial Dependency Modeling module is designed to capture long-range dependencies in 2D medical image features and to compensate for the limited receptive field of conventional convolutions. For liver tumor segmentation, lesion appearance depends not only on local intensity and edge cues but also on the anatomical context of the liver. Introducing explicit spatial dependency modeling is therefore important for accurate feature extraction.

In SMD, the input feature is first normalized and linearly projected for channel transformation. A depthwise convolution is then used to capture local neighborhood information, followed by a SiLU activation to enhance nonlinearity. On top of this local preprocessing, a two-dimensional state-space modeling unit (SSS2D) is introduced to further model spatial interactions between distant positions, and a final normalization and linear projection are used to obtain the module output. This design follows a "local preprocessing + global dependency modeling" scheme, which

preserves nearby structural information while explicitly modeling long-range spatial relationships.

$$\widehat{U}_l = \mathrm{SiLU}\left(\mathrm{DSConv}\left(\mathrm{Linear}\left(\mathrm{LN}(U_l)\right)\right)\right) \quad (3)$$

$$\mathrm{SMD}(U_l) = \mathrm{Linear}\left(\mathrm{LN}\left(\mathrm{SSS2D}\left(\widehat{U}_l\right)\right)\right) \quad (4)$$

Here, LN denotes layer normalization, Linear denotes linear projection, DWConv denotes depthwise convolution, σ denotes the nonlinear activation function, and SSS2D denotes the two-dimensional state-space modeling process. By combining local preprocessing with global dependency modeling, SMD preserves local structures while extending contextual perception across distant regions.

To better describe spatial dependencies in 2D medical images, SMD employs a two-dimensional state-space strategy inspired by recent visual state-space models [16]. Specifically, the preprocessed feature map is serialized into multiple spatially continuous one-dimensional sequences using several scan routes, and each sequence is modeled by a state-space unit. The outputs from different scan routes are then remapped to the 2D feature space and fused. Compared with simple row-wise or column-wise unfolding, this multi-route serialization better preserves spatial continuity and is therefore more suitable for jointly modeling local neighborhood information and long-range dependencies.

Let X denote the locally preprocessed feature map. It is converted into M one-dimensional sequences through M scanning routes, and the process can be written as follows:

$$S_l^{(m)} = \mathcal{P}_{sp}^{(m)}\left(\widehat{U}_l\right), m = 1,2,\dots,M \quad (5)$$

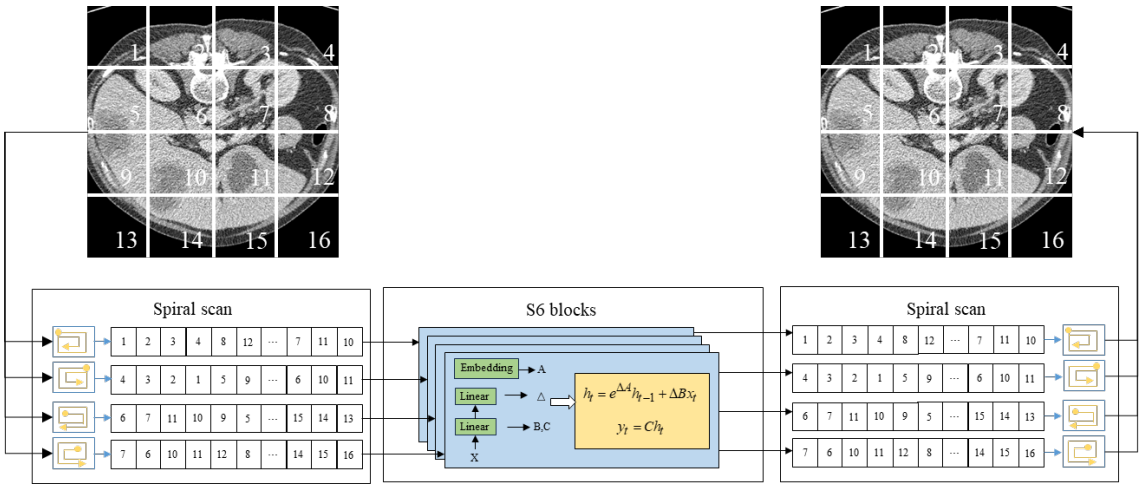

Figure 3. Two-dimensional state-space modeling process.

where $\mathcal{P}_{sp}^{(m)}(\cdot)$ denotes the serialization operation associated with the m-th scan route, $S_l^{(m)} \in R^{L\times C}$ denotes the resulting one-dimensional sequence, and L denotes the sequence length. Through this operation, the original 2D feature map is reorganized into several sequences with different traversal orders, which serve as the input to the subsequent state-space modeling.

For the sequence $S_l^{(m)} = \{s_t^{(m)}\}_{t=1}^{L}$ generated by the m-th route, a state-space unit recursively encodes the sequence as follows in Eq. (6) and Eq. (7).

$$h_t^{(m)} = A_m h_{t-1}^{(m)} + B_m s_t^{(m)} \quad (6)$$

$$y_t^{(m)} = C_m h_t^{(m)} \quad (7)$$

where h_t^m denotes the hidden state at time step t for the m-th scan route, y_t^m denotes the corresponding output, and A, B, and C denote the state transition, input projection, and output projection matrices, respectively. This recursive mechanism progressively aggregates contextual information along the scan path and establishes long-range interactions with relatively low computational cost.

After sequence modeling, the one-dimensional outputs obtained from different scan routes are remapped to the 2D space and fused to obtain the final state-space response, which can be expressed as follows:

$$\widetilde{Y_l^{(m)}} = \mathcal{P}_{sp}^{(m)-1}\left(Y_l^{(m)}\right), m = 1,2,\dots,M \quad (8)$$

$$\mathrm{SSS2D}\left(\widehat{U}_l\right) = \mathrm{Fuse}\left(\widetilde{Y_l^{(1)}}, \widetilde{Y_l^{(2)}}, \dots, \widetilde{Y_l^{(M)}}\right) \quad (9)$$

where $\mathcal{P}_{sp}^{(m)-1}(\cdot)$ denotes the inverse mapping associated with the m-th scan route and Fuse(·) denotes multi-route feature fusion. Through the procedure of sequence serialization, state-space encoding, inverse remapping, and feature fusion, the 2D state-space unit establishes long-range dependencies while preserving spatial continuity, thereby enhancing global contextual representation.

For liver tumor segmentation, where lesions often have blurred boundaries, irregular shapes, and low contrast with surrounding tissues, local convolutions alone are insufficient to characterize the relationship between tumors and the global liver structure. By embedding the above two-dimensional state-space modeling strategy into SMD, the network encodes the input feature from multiple scan routes, enhances perception of global context, and provides stronger support for subsequent local detail enhancement and final prediction.

(4) Multi-Scale Feature Enhancement Module

Although SMD effectively models global spatial dependencies and improves the perception of liver anatomy and lesion context, global modeling alone is not enough to fully recover local details. In liver tumor segmentation, especially for small lesions, weak-contrast boundaries, and morphologically complex regions, the network still requires stronger local structural representation. We therefore design the Multi-scale Feature Enhancement module to enhance intermediate features at multiple local scales and improve the description of boundaries, textures, and scale variation.

Liver tumors often exhibit significant scale differences in CT images, and local structures at different receptive fields contribute differently to precise segmentation. Smaller

receptive fields are beneficial for capturing fine-grained edges and textures, whereas larger receptive fields provide richer regional shape information. Based on this observation, MFE first applies a linear projection to the input feature and then sends the resulting feature to three depthwise-convolution branches with different kernel sizes to extract multi-scale local structural information. The process can be written as follows in Eq. (10) and Eq. (11).

$$R_l = \text{Linear}(U_l) \tag{10}$$

$$P_l = \text{Linear}\left(\text{Concat}\left(\text{DSConv}_{3\times3}(R_l),\text{DSConv}_{5\times5}(R_l),\text{DSConv}_{7\times7}(R_l)\right)\right) \tag{11}$$

where $DSConv_k(\cdot)$ denotes depthwise convolution with kernel size k and Concat(·) denotes channel concatenation. Through these parallel branches, the module simultaneously aggregates fine texture cues, local boundary structures, and larger-range shape information, yielding a richer multi-scale representation.

However, features from different scales do not contribute equally to the segmentation task. Directly using the concatenated representation may introduce redundant information and even weaken the response to critical lesion regions. To address this issue, we further introduce a channel recalibration mechanism. Global average pooling and a multilayer perceptron are used to generate channel weights, which are then applied to adaptively enhance the fused multi-scale feature. The process can be expressed as follows:

$$\text{MFE}(U_l) = P_l \odot \sigma\left(\text{MLP}\left(\text{AP}\ \ (P_l)\right)\right) \tag{12}$$

where GAP denotes global average pooling, MLP denotes a multilayer perceptron, Sigmoid denotes the activation function used to generate channel weights, and ⊙ denotes element-wise multiplication. This design allows the network to adaptively adjust channel importance according to global statistics, highlight features that are more sensitive to tumor boundaries, local textures, and critical shape cues, and suppress irrelevant background responses.

In summary, MFE compensates for the deficiency of SMD in local detail modeling through multi-scale convolutional extraction and adaptive channel recalibration. When facing tumors with blurred boundaries, obvious scale variation, or irregular morphology, the module helps preserve fine details and improves discriminative ability. Combined with SMD, the GLM block therefore achieves a coordinated integration of global contextual modeling and local detail enhancement, providing SSMDENet with a more sufficient and robust feature representation.

4. Experiments and Analysis

(1) Experimental Settings

All models were implemented in PyTorch 1.11.0 and trained on an NVIDIA A100 GPU with 40 GB memory. The training hyperparameters are listed in Table 1.

We used binary cross-entropy (BCE) loss as the supervision signal. BCE measures the discrepancy between pixel-wise prediction probabilities and ground-truth labels and is defined as follows in Eq. (13).

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)], \tag{13}$$

where N denotes the total number of pixels, y_i is the ground-truth label of the i-th pixel (0 for background and 1 for tumor), and p_i is the predicted probability that the pixel belongs to the tumor class.

To verify the effectiveness of the proposed SSMDEUNet, liver tumor segmentation experiments were conducted on the LiTS2017[17] and 3DIRCADB-01[18] datasets. The LiTS2017 dataset contains 201 abdominal 3D CT scans, among which 130 cases provide both image data and corre sponding annotations, while the remaining 70 cases include only image data. Since the official challenge has concluded and the test-set annotations are unavailable, this study employs the 130 annotated cases for training and evaluation. First, invalid slices were removed, including completely blank slices, slices with extremely low variance, or slices containing less than 5% non-zero pixels. Second, intensity normalization was performed by clipping CT values to the range of [−200, 250] HU, which covers the typical range of liver parenchyma, and then linearly scaling the intensities to [0, 1]. Finally, spatial resampling and resizing to 448 × 448 were applied to standardize voxel spacing and image dimensions. After preprocessing, a total of 7,190 valid 2D slices were obtained. Among them, 80% (5,752 slices) were used for training and 20% (1,438 slices) for testing. The training set was further divided at a 4:1 ratio into 4,601 training and 1,151 validation images.

Table 1. Training hyperparameters.

Parameter	Value
Epoch	300
Batch size	8
Learning rate	0.0001
Optimizer	Adam

The 3DIRCADB-01 dataset contains 10 annotated 3D CT scans, each comprising up to 46 tumor lesions. After preprocessing, a total of 568 liver images were obtained, including 363 for training, 91 for validation, and 114 for testing, with an input image size of 512 × 512.

(2) Comparative Experimental Results and Analysis

To evaluate the robustness and effectiveness of SSMDEUNet, we compared it with several state-of-the-art (SOTA) methods, including U-net[1], U-Net++[2], Attention U-Net[3], RIU-Net[19], QAU-Net[20], UCTrans-Net[7], CE-Net[4], Swin-UNet[8], and Fas-Net[21].

The quantitative results on LiTS2017 and 3DIRCADB-01 are reported in Tables 2 and 3, respectively.

On the LiTS2017 dataset, SSMDENet achieves Dice, Precision, Recall, VOE, and RAVD scores of 85.79%, 86.05%, 85.49%, 21.65%, and 0.12, respectively. Among all compared methods, our model obtains the best Dice and Precision. Compared with the baseline U-Net, SSMDENet improves Dice by 3.28 percentage points and Precision by 3.48 percentage points. Even relative to the strong FasNet model, Dice and Precision are further improved by 0.07 and 0.13 percentage points, respectively. These results indicate that the combination of state-space modeling and multi-scale detail enhancement enables more effective integration of global context and local boundary information. Although SSMDENet does not obtain the best Recall or VOE on this dataset, its leading Dice and Precision suggest a better balance between accurate target delineation and reduction of false positives.

On the 3DIRCADB-01 dataset, the advantage of

SSMDENet is more evident. The proposed model achieves the best Dice, Precision, and VOE among all compared methods, reaching 84.12%, 86.82%, and 23.05%, respectively. Compared with U-Net [1], Dice increases by 8.07 percentage points, Precision by 12.54 percentage points, and VOE decreases by 9.62 percentage points. Compared with FasNet [18], Dice and Precision are further improved by 0.61 and 0.90 percentage points, while VOE is reduced by 0.33 percentage points. Since tumors in 3DIRCADB-01 usually exhibit stronger scale variation and more complex morphology, the clearer gains on this dataset suggest that the proposed global-local collaborative modeling strategy is particularly effective in challenging scenarios. Although Recall and RAVD are not the best, the overall performance across Dice, Precision, and VOE demonstrates the robustness of SSMDENet.

For a more intuitive comparison, we randomly selected five CT images containing liver tumors from LiTS2017 for qualitative analysis. The visual results on LiTS2017 is shown in Figures 4, respectively.

Table 2. Segmentation performance of different methods on the LiTS2017 dataset.

Method	Dice (%)	Precision (%)	Recall (%)	VOE (%)	RAVD
U-Net[1]	82.51	82.57	84.97	25.81	0.29
U-Net++[2]	83.67	84.77	84.10	24.35	**0.07**
Attention U-Net[3]	84.39	84.80	85.54	23.77	0.10
RIU-Net[19]	83.30	83.87	84.44	25.32	0.17
QAU-Net[20]	84.00	84.01	85.51	24.33	0.15
UCTransNet[7]	84.77	84.57	**86.37**	23.14	0.13
CE-Net[4]	84.24	85.36	84.91	**20.02**	0.15
Swin-Unet[8]	84.62	84.75	85.88	23.03	0.12
FasNet[21]	85.72	85.92	85.88	21.33	0.12
SSMDENet (ours)	**85.79**	**86.05**	85.49	21.65	0.12

Table 3. Segmentation performance of different methods on the 3DIRCADB-01 dataset.

Method	Dice (%)	Precision (%)	Recall (%)	VOE (%)	RAVD
U-Net[1]	76.05	74.28	83.13	32.67	0.43
U-Net++[2]	79.43	76.13	87.31	29.54	0.38
Attention U-Net[3]	80.61	77.86	87.77	27.73	0.49
RIU-Net[19]	76.85	78.96	77.62	32.11	0.29
QAU-Net[20]	77.44	75.74	82.87	30.53	0.29
UCTransNet[7]	80.07	78.52	89.39	28.94	0.49
CE-Net[4]	80.29	83.73	80.38	28.62	0.03
Swin-Unet[8]	81.62	84.57	85.64	24.03	0.28
FasNet[21]	83.51	85.92	86.88	23.38	0.22
SSMDENet (ours)	84.12	86.82	85.21	23.05	0.36

(a) Test image

(b) Ground truth

(c)U-Net

(d)U-Net++

(e)AttentionU-Net

(f)RIU-Net

(g)QAU-Net

(h)UCTrans-Net

(i)CE-Net

(j)SwinU-Net

(k)Fas-Net

(l) SSMDENet (ours)

Figure 4. Qualitative results on the LiTS2017 dataset.

As can be observed from the qualitative results, the segmentation maps produced by SSMDENet are generally closer to the ground truth, especially on samples with multiple small lesions, blurred boundaries, and irregular shapes. The proposed method recovers more complete tumor contours while reducing false segmentation in the background. In contrast, several competing methods either miss lesion regions, oversmooth the boundary, or generate erroneous local predictions on difficult cases. The superior qualitative behavior is consistent with the improvements in Dice and Precision observed in the quantitative evaluation, further confirming the effectiveness of the proposed framework.

(3) Ablation Study

To fairly verify the contribution of each proposed component, we constructed a baseline network with the same encoder-decoder topology, the same number of downsampling stages, and the same channel configuration as SSMDENet, but without the proposed modules. Specifically, the GLM block was replaced by a standard convolutional block, while the remaining structure was kept unchanged. Based on this baseline, we progressively introduced the SMD and MFE modules to analyze the contribution of each component. The results are shown in Tables 4 and 5.

Table 4. Ablation results on the LiTS dataset.

Method	Dice (%)	Precision (%)	Recall (%)	VOE (%)	RAVD
Baseline	83.62	84.11	83.38	24.17	0.21
Baseline + SMD	84.96	85.41	84.72	22.63	0.17
Baseline + MFE	84.58	84.87	84.95	23.08	0.18
SSMDENet (ours)	85.79	86.05	85.49	21.65	0.12

Table 5. Ablation results on the 3DIRCADB-01 dataset.

Method	Dice (%)	Precision (%)	Recall (%)	VOE (%)	RAVD
Baseline	81.43	83.76	82.35	26.17	0.52
Baseline + SMD	83.28	85.31	84.18	24.09	0.43
Baseline + MFE	82.91	84.77	84.36	24.58	0.45
SSMDENet (ours)	84.12	86.82	85.21	23.05	0.36

As shown in Table 4, both proposed modules improve segmentation performance on the LiTS dataset, indicating that spatial dependency modeling and multi-scale feature enhancement are both beneficial for liver tumor segmentation. The baseline achieves Dice, Precision, Recall, VOE, and RAVD of 83.62%, 84.11%, 83.38%, 24.17%, and 0.21, respectively. After adding either SMD or MFE, all major metrics improve over the baseline, demonstrating that both global context modeling and local multi-scale enhancement strengthen feature representation.

When SMD is added to the baseline, Dice, Precision, and Recall increase to 84.96%, 85.41%, and 84.72%, respectively, while VOE decreases from 24.17% to 22.63% and RAVD decreases from 0.21 to 0.17. This shows that modeling long-range spatial dependencies helps the network better perceive global liver structure and lesion context, thereby improving overall target discrimination.

When MFE is added to the baseline, Dice, Precision, and Recall reach 84.58%, 84.87%, and 84.95%, respectively, with VOE and RAVD reduced to 23.08% and 0.18. The gain in Recall is particularly notable, suggesting that multi-scale convolution and channel recalibration are effective for preserving local boundary cues and small-scale lesion responses, thereby reducing missed detections.

When both SMD and MFE are introduced, SSMDENet achieves the best results on the LiTS dataset, with Dice, Precision, Recall, VOE, and RAVD of 85.79%, 86.05%, 85.49%, 21.65%, and 0.12, respectively. These results demonstrate strong complementarity between the two modules: SMD focuses on global spatial dependency modeling, whereas MFE emphasizes local detail enhancement. Their joint use improves lesion localization,

boundary recovery, and overall segmentation accuracy simultaneously.

As shown in Table 5, the two core modules also produce clear performance gains on the 3DIRCADB-01 dataset. The baseline achieves Dice, Precision, Recall, VOE, and RAVD of 81.43%, 83.76%, 82.35%, 26.17%, and 0.52, respectively. Compared with LiTS, the lower baseline performance on this dataset indicates that the stronger scale variation, blurred boundaries, and morphological complexity of 3DIRCADB-01 make the task more challenging and therefore provide a more demanding testbed for module validation.

After introducing SMD into the baseline, Dice, Precision, and Recall rise to 83.28%, 85.31%, and 84.18%, respectively, while VOE decreases to 24.09% and RAVD to 0.43. These improvements indicate that SMD enhances the model's ability to understand the spatial relationship between tumors and the overall liver structure, thereby improving global consistency and segmentation accuracy.

After introducing MFE, Dice, Precision, and Recall reach 82.91%, 84.77%, and 84.36%, respectively, while VOE and RAVD decrease to 24.58% and 0.45. Again, the improvement in Recall is especially evident. This suggests that multi-scale local enhancement better preserves boundary information and small-target responses, which helps reduce missed segmentations in complex lesion regions. Although the Dice and Precision gains are slightly smaller than those obtained with SMD alone, MFE still provides clear improvements over the baseline.

When both SMD and MFE are used together, SSMDENet achieves the best performance on 3DIRCADB-01, with Dice, Precision, Recall, VOE, and RAVD of 84.12%, 86.82%, 85.21%, 23.05%, and 0.36, respectively. Compared with the baseline, Dice, Precision, and Recall improve by 2.69, 3.06, and 2.86 percentage points, respectively, while VOE and RAVD decrease by 3.12 percentage points and 0.16. Compared with the SMD-only setting, Dice and Precision are further improved by 0.84 and 1.51 percentage points; compared with the MFE-only setting, they are further improved by 1.21 and 2.05 percentage points. These results confirm that SMD and MFE exhibit stronger synergy on more challenging data, where global structural understanding and local detail recovery are both essential.

In summary, both SMD and MFE improve the baseline network, but in different ways. SMD is more effective for strengthening global spatial dependency modeling, whereas MFE is more effective for supplementing local multi-scale detail information. Their combination yields the best results on both datasets, which verifies the rationality and effectiveness of the SSMDENet design.

5. Conclusion

In this study, we proposed SSMDENet, a liver tumor segmentation network based on state-space modeling and multi-scale enhancement, to address the difficulty of identifying small lesions and delineating complex tumor boundaries. The model follows an encoder-decoder architecture and introduces the GLM block to jointly model global context and local details. Within GLM, the SMD module uses a two-dimensional state-space modeling mechanism to capture long-range spatial dependencies and improve the perception of overall liver anatomy and lesion context, while the MFE module extracts multi-scale local information through depthwise convolution branches with different receptive fields and further emphasizes informative responses through channel recalibration. Experimental results on the LiTS and 3DIRCADB-01 public datasets show that SSMDENet achieves strong segmentation performance. The comparison and ablation studies further demonstrate the effectiveness of the proposed architecture and confirm that the collaboration between global context modeling and local detail enhancement is beneficial for accurate and robust liver tumor segmentation.

References

[1] WEI D, JIANG Y, ZHOU X, et al. A review of advancements and challenges in liver segmentation[J]. Journal of Imaging, 2024, 10(8): 202.

[2] WU C, CHEN Q, WANG H, et al. A review of deep learning approaches for multimodal image segmentation of liver cancer[J]. Journal of Applied Clinical Medical Physics, 2024, 25(12): e14540.

[3] MANJUNATH R, GOWDA Y. Automated segmentation of liver tumors from computed tomographic scans[J]. Journal of Liver Transplantation, 2024, 15: 100232.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, 2015, pp. 234-241.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3-11.

[6] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv:1804.03999, 2018.

[7] Z. Gu et al., "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," IEEE Transactions on Medical Imaging, vol. 38, no. 10, pp. 2281-2292, 2019.

[8] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752, 2023.

[9] Y. Liu et al., "VMamba: Visual State Space Model," arXiv:2401.10166, 2024.

[10] J. Ruan et al., "VM-UNet: Vision Mamba UNet for Medical Image Segmentation," arXiv:2402.02491, 2024.

[11] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range Sequential Modeling Mamba for 3D Medical Image Segmentation," in Proc. MICCAI, 2024.

[12] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid Densely Connected U-Net for Liver and Tumor Segmentation From CT Volumes," IEEE Transactions on Medical Imaging, vol. 37, no. 12, pp. 2663-2674, 2018.

[13] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021.

[14] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," in Proc. AAAI, 2022.

[15] H. Cao et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," in Proc. ECCV Workshops, 2022.

[16] KONG L, DONG J, TANG J, et al. Efficient visual state space model for image deblurring[C]//Proceedings of the computer vision and pattern recognition conference. 2025: 12710-12719.

[17] KONG L, DONG J, TANG J, et al. Efficient visual state space model for image deblurring[C]//Proceedings of the computer vision and pattern recognition conference. 2025: 12710-12719.

[18] BILIC P, CHRIST P, LI H B, et al. The liver tumor segmentation benchmark (lits)[J]. Medical image analysis, 2023, 84: 102680.

[19] SOLER L, HOSTETTLER A, AGNUS V, et al. 3d image reconstruction for comparison of algorithm database[J].

[20] P. Lv, J. Wang, and H. Wang, "2.5D Lightweight RIU-Net for Automatic Liver and Tumor Segmentation from CT," Biomedical Signal Processing and Control, vol. 75, p. 103567, 2022.

[21] L. Hong, R. Wang, T. Lei, X. Du, and Y. Wan, "QAU-Net: Quartet Attention U-Net for Liver and Liver-Tumor Segmentation," in Proc. IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.

[22] R. Singh et al., "FasNet: A Hybrid Deep Learning Model with Attention Mechanisms and Uncertainty Estimation for Liver Tumor Segmentation on LiTS17," Scientific Reports, vol. 15, p. 17697, 2025.