Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082353

ARTICLE

TriLVM-UNet: Multi-Scale State Space Modeling with Cross-Channel Fusion Attention Mechanism for Precise Medical Image Segmentation

Kexin Zhang[1], **Lihua Liu**[1,*], **Yuting Xue**[1], **Tao Zhou**[2], **Fengshuai Yue**[1] **and Ruifeng Du**[1]

[1]School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, China
[2]School of Computer Science and Engineering, North Minzu University, Yinchuan, China

*Corresponding Author: Lihua Liu. Email: liulihua@snut.edu.cn

Received: 14 March 2026; Accepted: 08 June 2026; Published: 23 July 2026

ABSTRACT: Traditional Mamba-UNet integrations employ four-stage architectures, replacing conventional five-stage UNets with VMamba blocks for global dependency modeling. Unlike Transformers, which suffer from quadratic complexity and high memory consumption in self-attention, Mamba-UNet achieves efficient global modeling through linear-complexity state space modeling. This paper proposes TriLVM-UNet, a lightweight three-stage architecture that integrates parameter-efficient VMamba blocks and enhances cross-stage feature interaction via an improved skip-attention bridge (SAB) module inspired by UltraLight VM-UNet. The model incorporates a Lightweight Vision Mamba (LVM) layer for high-resolution feature extraction, alongside multi-scale dilated convolution (MSDC) and convolutional block attention module (CBAM) for enhanced feature fusion. Evaluated on the 3D ACDC dataset against six baseline models, TriLVM-UNet achieves 98.57% accuracy. The GitHub repository is available at: https://github.com/730432ch/TriLVM-UNet.

KEYWORDS: Medical image segmentation; visual state space model; TriLVM-UNet; lightweight architecture

1 Introduction

With the rapid advancement of medical imaging technology and artificial intelligence methodologies, the role of Clinical Decision Support Systems (CDSS) in modern healthcare has become increasingly prominent. CDSS provides clinicians with diagnostic support, treatment recommendations, and prognostic assessment through the comprehensive analysis of multi-source medical data. Among these data sources, medical imaging plays a critical role in disease screening, diagnosis, and therapeutic efficacy evaluation. However, raw medical imaging data are often characterized by complex anatomical structures, high dimensionality, and significant information redundancy, making them difficult to utilize directly for automated decision analysis. Consequently, medical image segmentation, as a fundamental technique for extracting key anatomical structures and lesion regions from medical images, plays a central role in the development and application of CDSS.

In recent years, deep learning methods have achieved remarkable progress in medical image segmentation. Among these, the Transformer [1] has drawn attention for its long-range dependency modeling, but suffers from high computational complexity and memory consumption in 3D imaging. To address these issues, the Mamba model [2], a lightweight Selective State Space Model proposed in 2024, achieves efficient global modeling with linear complexity by introducing an input-dependent parameter generation mechanism that dynamically adapts state transitions to the input content, enabling flexible capture of structural

variations. For vision tasks, VMamba [3] (2024) extends Mamba to the visual domain, retaining CNN-like local feature extraction while incorporating selective state space modeling for cross-scale interaction, outperforming Transformer-based methods. Subsequently, U-Mamba [4] (2023) first introduced Mamba to medical image segmentation by integrating it into the nnUNet [5] framework, mitigating the limited receptive field of CNNs and the overhead of Transformers. Later, Mamba-UNet [6] (2024) adopted an encoder-decoder structure based on VMamba with U-Net [7] skip connections, supporting end-to-end 3D training and achieving promising results. Its innovation lies in combining selective state space models with U-Net, balancing efficiency and modeling capacity with a simple structure. Leveraging linear complexity, it processes full 3D volumes directly, avoiding context fragmentation from sliding windows and capturing cross-slice consistency. Moreover, its dynamic parameter mechanism enables adaptive perception of local anatomy, demonstrating robustness in complex segmentation tasks such as multi-region brain tumor and cardiac chamber segmentation.

Subsequently, various improved models based on Mamba and UNet have been proposed. Zhang et al. introduced VM-UNet-V2 [8] in 2024, which incorporates a Visual State Space (VSS) module to capture global contextual information and a Semantics and Detail Infusion (SDI) module to fuse low-level and high-level features, thereby alleviating the limitations of traditional CNNs in long-range dependency modeling while reducing the high computational complexity associated with Transformers. Wang et al. proposed Weak-Mamba-UNet [9] in 2024, constructing a weakly supervised medical image segmentation framework that combines CNN, Visual Transformer, and VMamba architectures, achieving high segmentation efficiency in 3D segmentation tasks with annotated data, demonstrating improved efficiency compared to Mamba-UNet in 3D scenarios. Addressing computational efficiency concerns, EfficientVMamba [10], proposed in 2025, employs an ES2D scanning mechanism to significantly reduce computational overhead and memory consumption while maintaining global modeling capability. Subsequently, SCSegamba [11] further designed a Structure-Aware Visual State Space module (SAVSS), combining lightweight Gated Bottleneck Convolution (GBC) and a Structure-Aware Scanning Strategy (SASS), along with a multi-path serpentine scanning strategy [3,12,13], effectively preserving multi-directional texture continuity in complex background scenarios. Zhao et al. proposed MM-UNet [14] in 2024, systematically analyzing the spatial structure preservation challenges faced by state space models when processing 3D medical images and highlighting that directly applying 1D sequence modeling methods to 3D images may lead to loss of spatial relationships. At the network architecture level, Wu et al. introduced H-vmunet [15] in 2025, which replaces traditional convolutional modules with high-order Vision Mamba modules while retaining the U-Net skip connection structure, and achieves feature alignment through SAB and CAB channels. Xing et al. proposed SegMamba [16] in 2024, effectively capturing long-range dependencies across different scales in 3D volumetric data from a state space modeling perspective, significantly improving processing speed while maintaining segmentation accuracy. Meanwhile, Transformer-based medical image segmentation methods have also achieved important progress. TransUNet [17] combines the Transformer self-attention mechanism with the U-Net structure, achieving a Dice score of 88.39% on the BTCV multi-organ segmentation task. SwinMM [18] introduces a multi-view rotation consistency constraint, achieving a Dice score of 86.18% on the WORD dataset using only 20% of annotated data. UNETR++ [19] significantly reduces model parameters while improving accuracy through paired attention modules. Although the Extended Kalman Filter [20] and Unscented Kalman Filter [21] have alleviated nonlinear issues to some extent, performance degradation risks remain under strong nonlinearities and complex noise conditions. Currently, most Mamba-based medical image segmentation models, such as UltraLight VM-UNet [22], MHA-UNet [23], HF-UNet [24], LightM-UNet [25], and MALUNet [26], generally adopt four-stage or five-stage network architectures. The four-stage design of Mamba-UNet provides an important reference for

subsequent research. MobileMamba [27] employs a TriLVM-UNet progressive feature extraction mechanism in medical image recognition tasks, achieving a favorable balance between performance and efficiency. Semi-Mamba-UNet [28] validates the effectiveness of the Mamba architecture under limited annotation conditions on 3D medical datasets such as MRI and CT. Swin-UNet [29] successfully applies a pure Transformer structure to medical image segmentation tasks for the first time, achieving significant improvements over ViT [30] in 3D window modeling.

Although UltraLight-VM-UNet, as a lightweight framework that integrates Mamba and UNet, has achieved high segmentation accuracy on the ISIC2D skin dataset, its performance on pseudo-3D datasets in the form of ACDC slices still requires further improvement. Most Mamba-UNet hybrid frameworks adopt a four-stage or five-stage architecture, and UltraLight-VM-UNet is no exception. Focusing on the three-stage model and the introduction of multi-scale depthwise convolution (MSDC) and an convolutional block attention module (CBAM), this work presents the following content.

To address the above issues, this paper proposes a three-stage feature-enhanced state space model that stacks consecutive 2D slices along the channel dimension, constructing a multi-channel 2D image where each channel represents one slice. A single-channel 2D image is replicated three times to form an RGB three-channel image, thereby adapting the input to pretrained models. Compared to previous works, we introduce an LVM Layer containing two Mamba blocks, which employs convolution for feature enhancement. The encoder and decoder achieve multi-level and multi-scale information fusion through the Spatial Attention Bridge (SAB) module [15,22,26,31].

The main innovations are as follows:

1. A novel medical image segmentation method called TriLVM-UNet is designed. This method adopts a three-layer network architecture, achieving a favorable balance between performance and efficiency.
2. An LVM Layer containing two Mamba blocks is designed. This layer consists of two parallel Mamba blocks, adopting a parallel mode to reduce memory consumption while maintaining performance. One LVM Layer is used in the first stage, two LVM Layers in the second stage, and three LVM Layers in the TriLVM-UNet to ensure proper channel matching.
3. Three-dimensional medical images are employed as segmentation targets. Multi-scale Depthwise Convolution (MSDC) [32] and an enhanced convolutional block attention module (CBAM) [32] are incorporated into the channel bridge module to improve segmentation accuracy and ensure experimental feasibility.
4. The TriLVM-UNet model achieves an accuracy (Acc) of 98.57%, specificity (Sp) of 99.15%, and sensitivity (Se) of 95.28% on the ACDC dataset.

The remaining parts of this article are organized as follows: Section 2 elaborates on the related work, Section 3 describes the model architecture, Section 4 details the experimental design, Section 5 presents the comparative results, and the Section 6 provides the conclusion.

2 Related Work

2.1 Lightweight State-Space Model

State Space Models (SSMs) have provided a theoretical foundation for efficient long-sequence modeling and critical support for visual tasks such as medical image segmentation. Modern SSM research originates from the Structured State Space Sequence Model (S4) [33] proposed in 2021. Based on HiPPO theory, S4 discretizes continuous-time state-space equations and uses structured state transition matrices with FFT acceleration to achieve efficient convolution kernel computation, delivering breakthrough performance on long-sequence tasks such as speech and long-text modeling. However, its complex structure, high

parameter coupling, and training instability lead to significant engineering costs, limiting deployment in resource-constrained scenarios. To address these issues, the lightweight models S4D [34] and S5 [35] were introduced in 2022. S4D simplifies the state transition matrix to a diagonal form, markedly reducing parameters and computation while improving training stability. S5 further imposes orthogonal parameter constraints and adopts a more efficient kernel generation strategy, enhancing numerical stability and efficiency without sacrificing representational capacity. Although these methods lower the engineering barriers of early SSMs and impart lightweight characteristics, they still face limitations in model generality and complex-task performance.

The Mamba model, proposed in 2023, propelled lightweight SSMs into the mainstream by introducing the concept of Selective SSM. With an input-dependent selection mechanism, it enables adaptive filtering of critical information, maintaining linear time complexity while acquiring attention-like selective modeling capabilities. This design addresses the representational limitations of traditional SSMs, reduces memory footprint during inference, and supports streaming computation, making it particularly suitable for edge devices and resource-constrained platforms. In language modeling, Mamba achieves performance comparable to Transformers with clear advantages in ultra-long sequences, marking a milestone for lightweight SSMs. Subsequently, several improved models emerged. Mamba-2 [36] (2024) enhanced throughput and training stability through optimized memory access and computation scheduling, advancing the engineering maturity of SSMs. Jamba [37] integrated Mamba with Transformer modules into a hybrid architecture, achieving a better balance between performance and efficiency. TinyViM [38] and Vision-Mamba extended the Mamba architecture to vision tasks with targeted optimizations for mobile and lightweight applications, validating the feasibility and potential of SSMs in computer vision. Furthermore, DualMamba [39] (2024) introduced a dual-path conditional computation architecture that dynamically allocates computation based on input features, offering a novel efficiency optimization approach distinct from traditional static compression methods and reflecting a paradigm shift toward dynamic, intelligent scheduling in lightweight design.

2.2 Mamba-Based Medical Image Segmentation Network

With the development of Mamba and its vision variants, researchers have explored their potential in medical image segmentation. UltraLight VM-UNet (2025) is among the smallest Vision Mamba-based segmentation models, with only 0.049M parameters and 0.060 GFLOPs. It achieves high accuracy on skin lesion datasets such as ISIC2017, ISIC2018, and PH^2 while maintaining extremely low computational cost, and its PVM Layer enables systematic extreme compression of the Mamba module for ultra-lightweight medical applications. MHA-UNet (2023) introduces a multi-high-order attention interaction mechanism that allows skin lesion segmentation without negative samples, and employs an interpretable reasoning module for lesion classification. It outperforms state-of-the-art methods on these datasets and offers a highly interpretable joint segmentation–classification framework. HF-UNet (2021) applies a multi-task learning framework with a dual-branch structure and a TCL module to balance task-specific and shared features, and adds a boundary regression task to improve prostate boundary detection in low-contrast CT images. It achieves state-of-the-art prostate segmentation and provides a scalable paradigm for boundary-sensitive multi-task medical image segmentation.

LightM-UNet, proposed in 2024, addresses the issues of large parameter counts and high computational complexity in traditional U-Net and its Transformer variants by, for the first time, introducing Mamba as a lightweight core module into medical image segmentation networks, constructing an efficient segmentation model with only approximately 1 million parameters. By designing the RVM Layer, LightM-UNet achieves long-range dependency modeling while realizing extreme parameter compression. On 3D CT liver tumor

segmentation and 2D X-ray lung segmentation datasets, this model achieves or surpasses current state-of-the-art methods in metrics such as DSC and mIoU. Compared to nnU-Net, its parameter count is reduced by 116 times and computational load by 21 times; compared to the contemporaneous U-Mamba, its parameter count is reduced by 224 times while achieving superior performance. Experimental results indicate that LightM-UNet has distinct advantages in segmentation boundary smoothness and small object recognition, validating the application potential of Mamba as a new generation of lightweight vision foundation modules. Furthermore, MALUNet, proposed in 2022, adopts a six-stage U-shaped architecture and significantly compresses the number of channels at each stage (with a minimum of only 8), substantially reducing model parameters and computational complexity while maintaining segmentation performance. Compared to traditional UNet, MALUNet reduces parameters by 44 times and computational complexity by 166 times. On the ISIC2017 and ISIC2018 datasets, this model outperforms most lightweight methods in metrics such as mIoU and DSC, and achieves lower resource consumption while delivering performance comparable to similar models like UNeXt-S (a medical image segmentation network based on convolution and multi-layer perceptrons). Visualization results demonstrate that MALUNet exhibits superior performance in segmentation edge accuracy and detail preservation, offering a solution with practical deployment value for resource-constrained mobile medical devices. Identity-mapping ResFormer [40], proposed in 2025 for pneumonia research, presents a computer-aided diagnostic model that utilizes Multi-Convolution Cascade Residual Modules (MCCRM) and Enhanced Multi-Patch Transformers (EMPT) to extract global attention features from multiple receptive fields. MambaYOLACT [41], also proposed in 2025, designs a Cross-field and Cross-direction Feature Enhancement module (CCFE) to enhance the capability of different feature groups to capture diverse spatial semantic information. Furthermore, it designs a prediction head module based on MambaNet, enabling accurate capture of global image dependencies and highlighting lesion regions. During 2025–2026, attention mechanisms continued to play a pivotal role in hybrid feature extraction. The multi-scale discriminative feature attention network enhanced multi-scale feature discrimination capability through a hybrid Transformer–CNN architecture. The Dynamic Context Channel Attention (DCCA) introduced in Hawk-Net [42] overcame the constraints of conventional static attention mechanisms, enabling dynamic enhancement of contextual information. Furthermore, the Bayesian Spiral Attention Classifier (BASIC, 2025–2026) integrated SE channel attention with an innovative spiral spatial attention, offering new perspectives for multimodal and generalizable medical image analysis.

2.3 State Space Model

The structure of the Selective State Space Model, as shown in Fig. 1, is a mathematical model used to describe the evolution of a dynamic system's state over time. SSM characterizes how a system progresses through time steps using a set of matrices and state variables. This model typically comprises a state equation and an output equation, and can be formulated in either continuous time or discrete time.

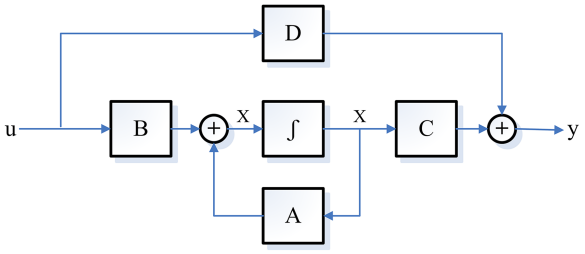

Figure 1: Architecture of the selective state space model.

The State Space Model (SSM) is a mathematical framework for describing dynamic systems. In the SSM, there are three variables related to time t: $\chi(t) \in C^n$ represents n state variables, reflecting the current state of the system. $u(t) \in C^m$ represents m input variables, introduced into the system as external influences. $y(t) \in C^p$ represents p output variables, which are typically the values we wish to observe from the system. In addition, the SSM is constructed from four learnable matrices, whose dimensions correspond to those of the variables: $A \in C^{n\times n}$ represents the state matrix, which governs the evolution of the state vector and influences how the state updates over time. $B \in C^{n\times m}$ represents the control matrix, which governs how the input vector $u(t)$ is mapped to the state $x(t)$. $C \in C^{p\times n}$ represents the output matrix, which governs how the state influences the output. $D \in C^{p\times m}$ represents the feedthrough matrix, which directly maps the input $u(t)$ to the output $y(t)$. These variables satisfy the following equations:

$$x'(t) = Ax(t) + Bu(t)$$

$$y(t) = Cx(t) + Du(t) \tag{1}$$

Discretization is one of the core steps in state space models, as it converts the continuous-time representation of a system into a discrete-time representation, enabling the SSM to be implemented on a computer. We need to approximate the continuous state changes using a time step Δt.

$$x(t + \Delta t) = (1 - a \cdot \Delta t) \cdot x(t) + b \cdot \Delta t \cdot u(t) \tag{2}$$

Here, $x(t)$ is the state of the system, $u(t)$ is the input, and a and b are constants representing the system's decay factor and the influence of the input, respectively. The discretized matrices can be expressed as:

$$x_k = \overline{A}x_{k-1} + \overline{B}u_k$$

$$y_k = \overline{C}x_k \tag{3}$$

where $\overline{\mathrm{A}}$, $\overline{\mathrm{B}}$ and $\overline{\mathrm{C}}$ represent the state transition matrix, input matrix, and output matrix, respectively.

2.4 Pseudo-3D Medical Image Storage Structure

In the field of medical image analysis, the evolution from traditional two-dimensional images to three-dimensional volumetric data represents a paradigm shift from slice-based observation to volumetric analysis (Fig. 2). Three-dimensional medical images, such as CT, MRI, and PET/CT, provide continuous spatial information on human anatomical structures and pathological changes, offering irreplaceable value for precise diagnosis, surgical planning, and treatment evaluation. However, their unique storage structure and large data volumes impose fundamentally different requirements on processing algorithms compared with two-dimensional images. The storage of three-dimensional medical images essentially involves the discretization of continuous anatomical space into a voxel matrix, which primarily manifests in two forms. The first is sequence-based two-dimensional storage, with DICOM serving as the primary format and being the most prevalent in clinical environments. Each three-dimensional scan generates a DICOM series consisting of hundreds of two-dimensional slice images. Each slice is stored as an independent file, containing not only pixel data but also rich metadata, such as patient information, scanning parameters, and crucial spatial positioning and orientation details. The advantage of this structure lies in its high flexibility, facilitating easy viewing and transmission of individual slices. However, for computational analysis, a preprocessing pipeline is essential to reconstruct these discrete slices into a regular three-dimensional grid with either isotropic or anisotropic resolution. The accuracy of this step directly determines the reliability of subsequent analyses.

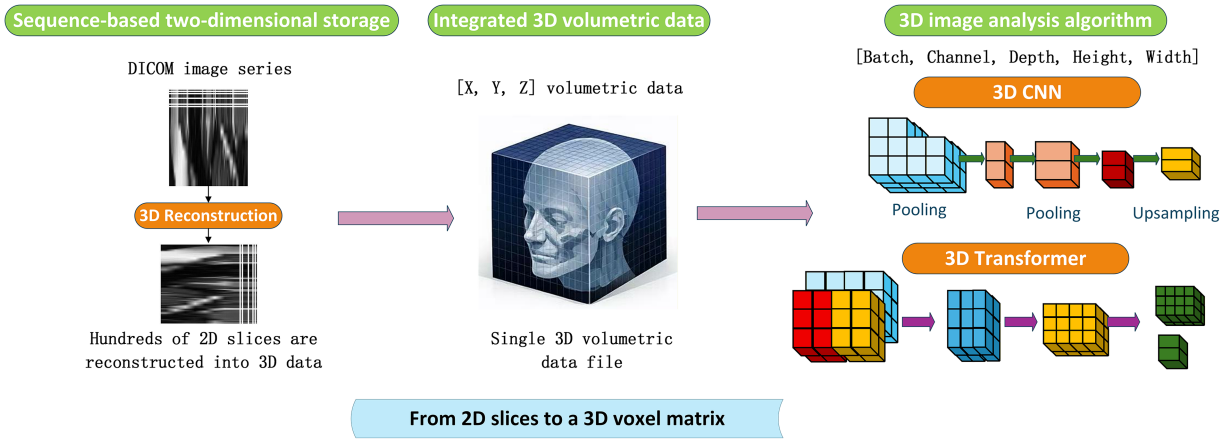

Figure 2: 3D image processing process diagram.

Model construction lies at the core of this pipeline. Depending on the task requirements, three typical paradigms can be adopted. When inter-slice spacing is large or computational resources are limited, the volumetric data can be directly sliced into independent two-dimensional images to train a 2D network, achieving simple yet efficient segmentation or classification. To capture local contextual information, three to five adjacent slices along the Z-axis can be stacked as multi-channel input to form a 2.5D network, which enhances inter-slice continuity without significantly increasing the number of parameters. When the slice spacing is small and GPU memory is sufficient, a 3D convolutional network can be employed to process genuine voxel blocks, fully exploiting three-dimensional spatial information. During training, a combination of Dice loss and cross-entropy is commonly used as the loss function to address foreground–background imbalance. In the inference stage, the entire volume is typically traversed using a sliding-window strategy, and the outputs of individual patches are fused through weighted averaging or voting to obtain the complete predicted volume. Post-processing steps, including thresholding, connected-component analysis, and minimum contour constraints, are applied to remove isolated false positives. The final evaluation employs three-dimensional metrics such as the Dice coefficient and Hausdorff distance. When necessary, the segmentation results are rendered via three-dimensional visualization to verify model coherence in the coronal and sagittal planes. This workflow balances the efficiency of 2D approaches with the integrity of 3D analysis, representing a mainstream solution in medical image analysis.

3 Research Methods

In this section, we first introduce the preliminary concepts related to the TriLVM-UNet model, namely the state-space model and the discretization process. Then, we present a comprehensive overview of the overall architecture of TriLVM-UNet and describe the modeling process of LVM, which serves as the core component of the model. Finally, we analyze the SAB bridge module, the MSDC module, and the CBAM module.

3.1 TriLVM-UNet Network Model

3.1.1 Overall Architecture of the Model

The overall architecture of the proposed TriLVM-UNet model is illustrated in Fig. 3. The network mainly consists of LVM layers, convolutional layers, the SAB bridge module, Multi-scale Depthwise Convolution

(MSDC) [32], the Convolutional Block Attention Module (CBAM) [32], as well as downsampling and upsampling layers.

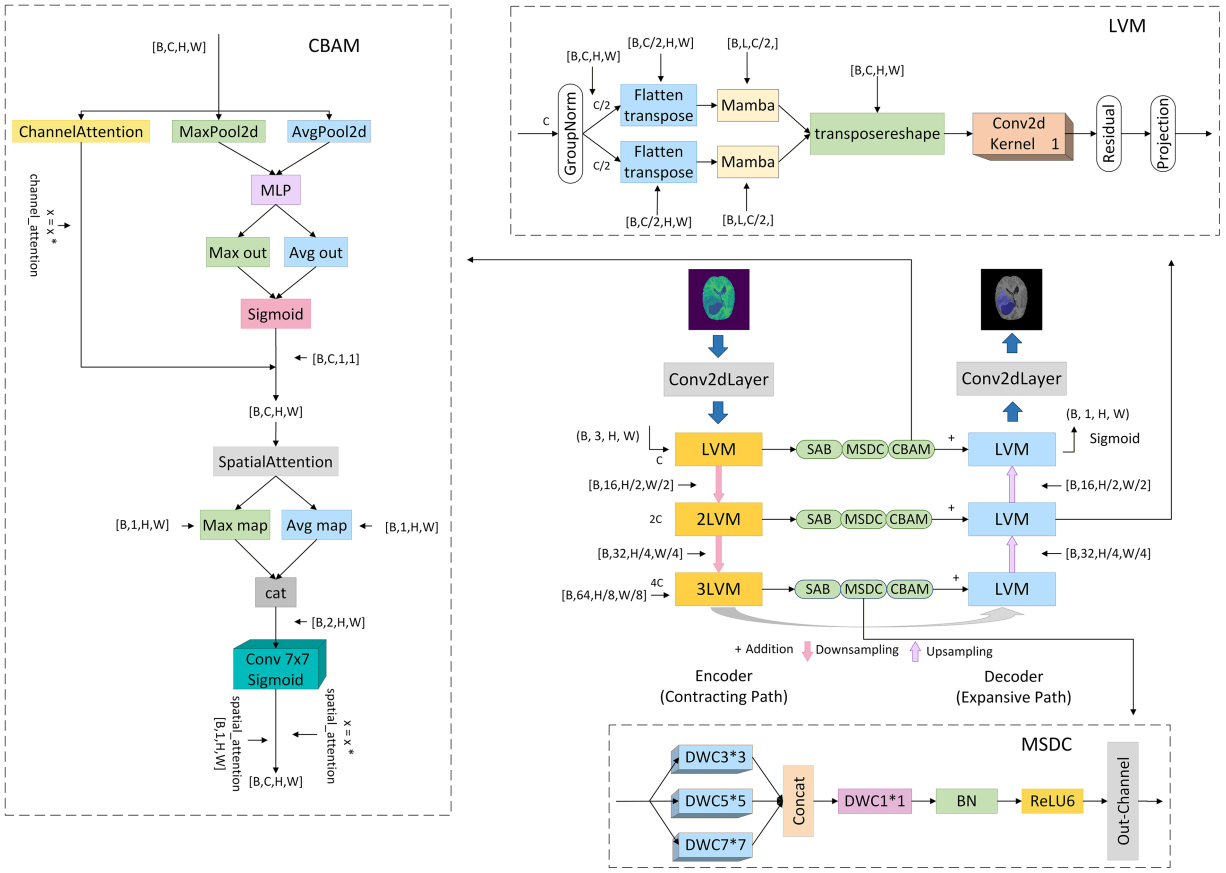

Figure 3: Architecture of the TriLVM-UNet.

The TriLVM-UNet model takes an input tensor of size (B, C, H, W) = (2, 3, 256, 256), where the channel dimensions at each stage are set to [16, 32, 64]. The proposed architecture integrates Multi-scale Depthwise Convolution (MSDC), Mamba blocks, and the Convolutional Block Attention Module (CBAM), thereby achieving a balance between computational efficiency and segmentation performance.

The core design of the TriLVM-UNet model lies in transforming the original four-stage architecture into a three-stage medical image segmentation model by integrating the framework of UltraLight-VM-UNet and the three-layer feature structure of MobileMamba. In the encoder, the GELU activation function is employed. In the following sections, the structure and function of each stage are described in detail.

Stage-1: It aims to extract basic local texture and edge information from the raw input while completing the initial spatial scale compression. Given the input features $X \in R^{B\times C_{in}\times H\times W}$, First, a 3 × 3 convolutional layer is applied to downsample the input, reducing the spatial resolution to (H/2, W/2) and mapping the number of channels to half of the target stage's channel count. This operation reduces computational complexity while completing preliminary feature embedding. Subsequently, a second 3 × 3 convolutional layer is introduced to expand the channel dimension to the full channel count of Stage One while preserving the spatial resolution, enhancing feature representation capability through normalization and nonlinear activation. On this basis, the features are fed into an LVM, which introduces efficient long-range dependency

modeling capabilities without altering the feature map resolution or channel dimension, thereby compensating for the limitations of traditional convolutions in global modeling. Finally, a residual connection mechanism is employed to perform element-wise addition between the output of the LVM and the output of the second convolutional layer, forming the final output features of Stage One $s_1 \in R^{B\times C_1\times \frac{H}{2}\times \frac{W}{2}}$. This design effectively integrates local convolutional features with globally modeled information, while also helping to stabilize the training process of deep networks.

Stage-2: Stage Two takes the output of Stage One as input, aiming to further enhance the semantic representation capability of features while performing the second spatial downsampling. First, a 3 × 3 convolutional layer is used to compress the spatial resolution of the feature map from (H/2, W/2) to (H/4, W/4), while projecting the number of channels to half of the target channel count for Stage Two, achieving a balance between computational efficiency and representational power. Subsequently, a primary 3 × 3 convolutional layer expands the channel dimension to the full channel count while maintaining the spatial scale, forming a stable mid-level semantic representation. To enhance the capability for modeling complex structures and long-range dependencies, two consecutive LVMs are introduced in series at this stage. Each module maintains consistent input and output channel dimensions, allowing features to progressively refine semantic relationships at a unified scale. Finally, Stage Two employs a residual summation strategy, adding the output of the primary convolutional layer to the output of the second LVM, resulting in the stage's output features $s_2 \in R^{B\times C_2\times \frac{H}{4}\times \frac{W}{4}}$. This structure effectively enhances the discriminative power of mid-level features while ensuring stable gradient propagation.

Stage-3: Stage Three serves as the high-level stage of the encoder, primarily responsible for abstracting high-level semantic information and constructing stronger global contextual representations. The input features first undergo a 3 × 3 downsampling convolution, which further compresses the spatial resolution to (H/8, W/8) and adjusts the number of channels to half of the target channel count for Stage Three. Subsequently, a 3 × 3 primary convolutional layer expands the channel dimension to the full C_3, providing sufficient feature capacity for subsequent high-level semantic modeling. In this stage, three consecutive LVMs are stacked to progressively enhance the global modeling capability of the features and facilitate information interaction across spatial positions. This multi-layer stacking design enables the model to efficiently capture long-range dependencies at a lower resolution. Finally, a residual connection is employed to add the output of the primary convolutional layer to the output of the third LVM, forming the final encoded features of Stage Three $s_3 \in R^{B\times C_3\times \frac{H}{8}\times \frac{W}{8}}$.

The bridge module takes as input the feature maps s_1, s_2, and s_3 from different stages of the encoder, with spatial resolutions of (H/2, W/2), (H/4, W/4), and (H/8, W/8), and corresponding channel dimensions of C_1, C_2, and C_3, respectively. By incorporating the SAB module, which integrates convolution and attention mechanisms, it performs cross-stage resampling and fusion of features at different scales. Combined with multi-scale feature modeling and attention enhancement mechanisms, this enables sufficient cross-scale information interaction. The features output by the bridge module maintain the same spatial resolution and channel dimension as their corresponding inputs, thereby enhancing multi-scale semantic representation capabilities without disrupting the feature hierarchy. In the decoding stage, the network begins progressive reconstruction starting from the lowest-resolution feature s_3. First, an LVM is used to map the feature channels from C_3 to C_2, and the feature map is restored to the same spatial resolution as s_2 through a 2× upsampling operation, obtaining the intermediate decoding feature d_2. Subsequently, d_2 is concatenated with the encoder feature s_2 at the corresponding scale along the channel dimension, and another LVM is employed to compress the concatenated channel count from $2C_2$ to C_1. Finally, a 2× upsampling operation is performed again to obtain the decoded feature d_1, which is aligned with s_1 in spatial resolution, providing

a feature representation that combines semantic information and spatial details for high-resolution segmentation prediction. d_1 is then concatenated with s_1, and subsequently passed through a series of convolutions and upsampling operations. The final output is a feature map with the same resolution as the input image, which is then processed using a sigmoid activation function to produce the final output. In the TriLVM-UNet architecture, the input to Stage One is an image of dimensions $H \times W \times C$. Following downsampling, the spatial resolution is halved and the channel depth doubled, resulting in a feature map of size $\frac{H}{2} \times \frac{W}{2} \times 2C$. This serves as the input to Stage Two, where an analogous downsampling operation yields a representation of $\frac{H}{4} \times \frac{W}{4} \times 4C$. Stage Three further processes this representation, producing a final downsampled feature map of $\frac{H}{8} \times \frac{W}{8} \times 8C$. Departing from the conventional four-stage Mamba–UNet hybridization paradigm, the TriLVM-UNet model demonstrates excellent performance on the ACDC dataset, achieving an accuracy (Acc) of 94%. These results provide a valuable reference for future studies in medical image segmentation.

3.1.2 Dual-Branch Parallel LVM Module

As a core component of TriLVM-UNet, the LVM module consists of two parallel Mamba blocks (Fig. 4). In most architectures that integrate Mamba with U-Net, only a single Mamba block is embedded within the U-shaped structure. By contrast, the proposed LVM module introduces two parallel Mamba blocks to enhance feature modeling capability. The module first applies group normalization and then splits the feature channels into two branches, each fed into one of the Mamba blocks. The features are subsequently reshaped into a sequence format (B, H × W, C/2) for Mamba processing. After the parallel Mamba blocks, the outputs are concatenated along the channel dimension and passed through a projection layer to produce the final feature representation.

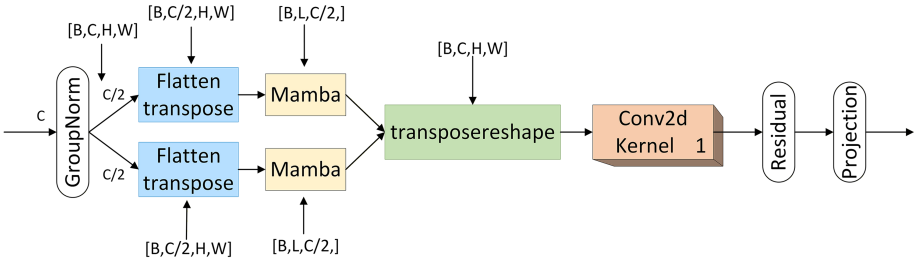

Figure 4: Architecture of the LVM module.

Group Normalization (GN) is a normalization technique used in deep neural networks, designed to mitigate the performance degradation of Batch Normalization (BN) when training with small mini-batches. GN normalizes features by dividing the channels into groups, making it independent of the batch size and thereby ensuring stable performance even when the batch size varies. It operates independently of the batch dimension, which further stabilizes training. Moreover, GN behaves consistently during both training and inference, eliminating the need for the moving average statistics required by BN and thus enhancing the model's generalization capability. Given an input $x \in R^{\{B\times C\times H\times W\}}$, GN satisfies the following formula:

$$GN(x) = \gamma \cdot \frac{x - \mu_{group}}{\sqrt{\sigma^2_{group} + \varepsilon}} + \beta \tag{4}$$

here, the input features are divided into G = 4 groups. μ_{group} and σ_{group} denote the mean and standard deviation of each group; γ and β are learnable scaling and shifting parameters; and ε is a small constant for numerical stability. The normalized feature map then undergoes dual-branch Mamba channel splitting and spatial flattening, satisfying the following formula:

$$x_1, x_2 = split\left(x_{norm}, 2, \text{dim} = 1\right)$$

$$x_i, seq = reshape\left(x_i\right)$$

$$x_1, x_2 \in R^{\{B\times C/2\times H\times W\}} \tag{5}$$

Subsequently, it passes through the Mamba state space (see Eqs. (1)–(3)), followed by a residual connection. $\hat{x}_i \in R^{\{B\times L\times C/2\}}, L = H \times W$ Satisfies the following formula:

$$x_{out} = Mamba\left(x_{seq}\right) + \alpha \cdot x_{seq} \tag{6}$$

where, $\alpha = self.skip\backslash\_scale$: learnable scaling factor, initialized to 1.0, allows the model to automatically adjust the contribution of the residual connection. Subsequently, dual-branch merging and deserialization are performed.

$$x_{mamba} = concat\left(\left[x_{mamba1}, x_{mamba2}\right], \text{dim} = 2\right)$$

$$x_{mamba} \in R^{B\times H\times W\times C}$$

$$x_{out} = reshape\left(x_{mamba}\right) \in R^{B\times H\times W\times C} \tag{7}$$

Finally, the features are concatenated along the channel dimension to reconstruct a planar feature map, which is subsequently passed through a projection convolution to generate the final output. As the core component of the TriLVM-UNet model, the LVM layer not only serves as a bridge connecting traditional vision models with modern sequence models, but also effectively balances three critical dimensions: modeling capability, computational efficiency, and practical deployment requirements, thereby providing a novel solution for lightweight medical image segmentation.

3.1.3 Multi-Scale Feature Interaction Enhancement SAB Module

The SAB is a core module within the TriLVM-UNet architecture, constructing an efficient multi-scale feature interaction bridge between the encoder and decoder. The SAB module integrates a Multi-scale Depthwise Convolution (MSDC) module and a Convolutional Block Attention Module (CBAM) to enhance feature interactions. The process begins with multi-scale feature extraction via MSDC, followed by normalization, a GELU activation, and attention refinement via CBAM. By addressing three critical issues inherent in the traditional U-Net architecture—namely the information bottleneck, scale inconsistency, and semantic gap—this module substantially improves segmentation performance. Structurally, the module implements a bidirectional feature interaction mechanism. Through adaptive resampling, feature maps of different resolutions are scale-aligned, ensuring that each stage receives complementary information from the other stages. This design overcomes the limitations of traditional unidirectional information flow, enabling the full fusion of shallow detail features with deep semantic features.

The core of the module employs a collaborative design that integrates Multi-scale Depthwise Convolution (MSDC) with an attention mechanism. MSDC utilizes parallel depthwise separable convolutions with kernel sizes of 3 × 3, 5 × 5, and 7 × 7 to capture feature representations across multiple receptive fields with low computational cost. Concurrently, the integrated Convolutional Block Attention Module (CBAM) comprises two components—channel attention and spatial attention—enabling adaptive selection of salient feature channels and focusing on critical spatial regions, thereby achieving intelligent feature enhancement. Functionally, this module acts as a post-processing unit for the encoder's output, providing the decoder

with feature inputs that are richer in information, semantically stronger, and spatially well-aligned. By preserving original feature information through residual connections, it facilitates mutual enhancement of cross-scale features: deep features inject semantic understanding into shallow features, while shallow features supplement spatial details for deep features. This design is particularly suited for tasks requiring precise boundary delineation and multi-scale object recognition, such as medical image segmentation and remote sensing image analysis. The SAB module substantially enhances feature representation in a parameter-efficient manner, improves gradient flow, and increases the model's adaptability and segmentation accuracy in complex scenes, providing an effective strategy for improving U-Net-like architectures.

3.1.4 Multi-Scale Depthwise Convolution Module (MSDC)

The MSDC is a core convolutional module within the TriLVM-UNet architecture, positioned within the model's skip connections. Unlike the originally proposed MSDC, instead of immediately concatenating the outputs of the three convolutional operations, we first apply a 1×1 convolution for enhanced feature extraction, followed by a ReLU6 activation function, and then perform concatenation before generating the final output. Multi-scale depthwise convolution facilitates enhanced feature extraction by capturing contextual information at multiple scales, enabling hierarchical representation from local details to global context. The depthwise separable structure substantially reduces both the parameter count and computational complexity, allowing the model to improve feature representation while maintaining a lightweight design. This makes it particularly well-suited for processing high-resolution medical imaging data. The MSDC module employs a parallel multi-path architecture to capture heterogeneous features synchronously, followed by adaptive fusion through channel concatenation and a 1×1 convolution. This enables the model to simultaneously perceive both local details and global structures. Such a design effectively mitigates common challenges in medical imaging, including blurred boundaries and structural deformations.

The MSDC module is embedded within the stage-aware bridge, enabling cross-scale feature interaction across different levels of the encoder. Shallow high-resolution features and deep semantic features are effectively fused through the MSDC, enhancing the model's ability to discriminate multi-scale anatomical structures. Fig. 5 illustrates the structure of the MSDC module:

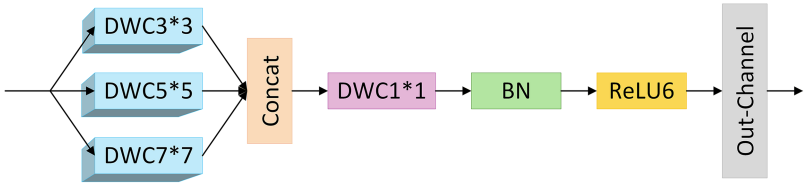

Figure 5: Architecture of the MSDC module.

Here, DWC represents depthwise convolution, BN denotes batch normalization, and ReLU6 is the activation function. The convolutions satisfy the following formulas, respectively:

$$y_c^{(1)}(i,j) = \sum_{m=-1}^{1} \sum_{n=-1}^{1} \omega_c^{(1)}(m,n) \cdot x_c(i+m, j+n) \tag{8}$$

$$y_c^{(2)}(i,j) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} \omega_c^{(2)}(m,n) \cdot x_c(i+m, j+n) \tag{9}$$

$$y_c^{(3)}(i,j) = \sum_{m=-3}^{3} \sum_{n=-3}^{3} \omega_c^{(3)}(m,n) \cdot x_c(i+m, j+n) \tag{10}$$

where $y_c^{(k)}$ represents the output of the k-th scale at the c-th channel, $\omega_c^{(k)}$ represents the convolution kernel of the k-th scale at the c-th channel, and (i, j) represent the output spatial location coordinates. Multi-scale fusion satisfies the following formula:

$$\Upsilon = concat\left(\Upsilon^{(1)}, \Upsilon^{(2)}, \Upsilon^{(3)}\right) \in R^{B\times(3\times C_{in})\times H\times W} \tag{11}$$

where $\Upsilon^{(k)}$ represents the output of the k-th scale, and B denotes the batch size, C_{in} represents the number of input channels.

3.1.5 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) plays a crucial role in feature selection and enhancement within the TriLVM-UNet model. By employing both channel attention and spatial attention, CBAM adaptively refines feature representations. Sequential integration of channel and spatial attention substantially improves the model's ability to perceive and prioritize salient features in medical images. The channel attention submodule employs a dual-path aggregation of global average pooling and max pooling, generating channel weights via a shared Multi-Layer Perceptron (MLP) to dynamically calibrate the importance of feature channels. This effectively suppresses redundant information while emphasizing semantic features relevant to the segmentation target. The spatial attention submodule aggregates average and maximum features along the channel dimension, producing a spatial weight map through a convolutional layer to guide the model in focusing on salient regions corresponding to target anatomical structures. Embedded within the enhanced stage-aware bridge, this module is positioned after multi-scale depthwise convolution for feature extraction and is responsible for the refined recalibration of fused cross-stage features. Fig. 6 illustrates the structure of the CBAM:

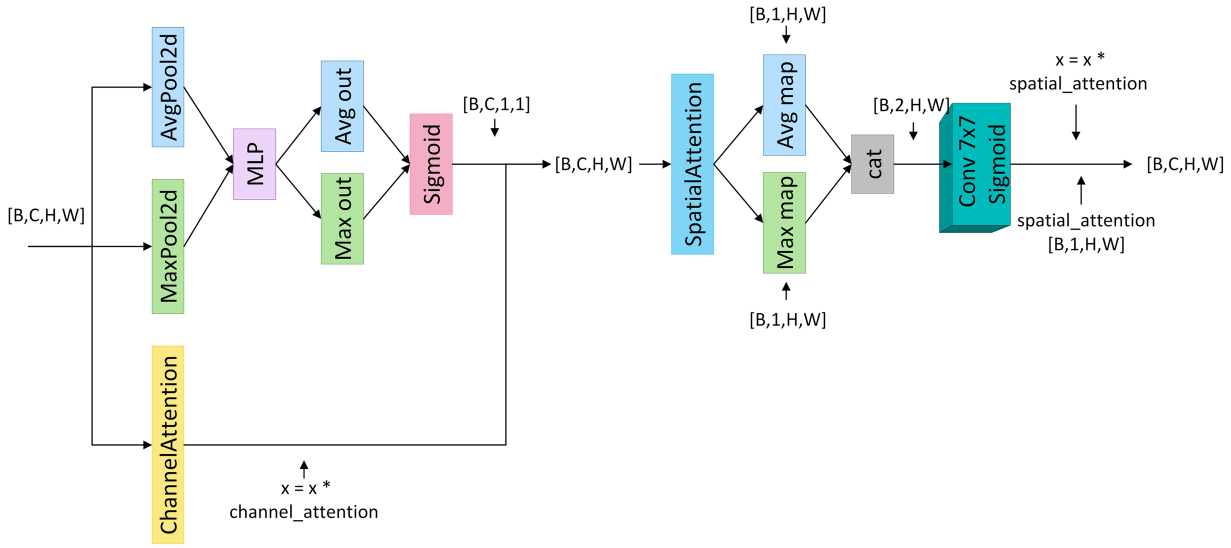

Figure 6: Architecture of the CBAM module.

Channel Attention and Spatial Attention are represented, respectively. The structural diagrams of the two attention mechanisms are provided below. Fig. 7 is the channel attention architecture.

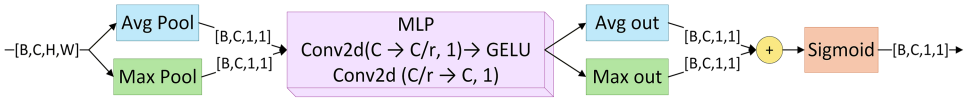

Figure 7: Architecture of the channel attention module.

Here, r is typically an integer greater than 1, used to reduce the number of channels in the intermediate layer, thereby lowering computational cost and parameter count. GELU and Sigmoid are activation functions. Fig. 8 is the spatial attention architecture.

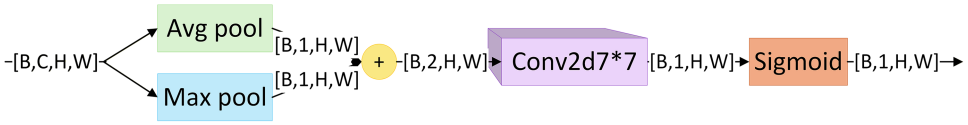

Figure 8: Architecture of the spatial attention module.

The channel attention mechanism branch first undergoes average pooling and max pooling, followed by convolution and activation functions, and finally multiplies with the feature map. The involved formulas are as follows:

$$F_{\text{avg}}^{c} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c\left(i, j\right) \tag{12}$$

$$F_{max}^{c} = \max_{i,j} X_c(i, j) \tag{13}$$

$$M_c\left(F\right) = \sigma\left(W_1 ReLU\left(W_0 F\right)\right) \tag{14}$$

$$M_c\left(X\right) = \sigma\left(\begin{array}{c} MLP\left(AvgPool\left(X\right)\right) + \\ MLP\left(MaxPool\left(X\right)\right) \end{array}\right) \tag{15}$$

where $W_0 \in R^{C/r \times C}$ represents the weights of the first fully connected layer, $W_1 \in R^{C \times C/r}$ represents the weights of the second fully connected layer, r denotes the reduction ratio, and σ is the sigmoid activation function.

The spatial attention mechanism performs channel-wise average pooling and max pooling, followed by concatenation, convolution, and an activation function, and finally multiplies with the feature map, satisfying the following formula:

$$F_{\text{avg}}^{8}\left(i, j\right) = \frac{1}{C} \sum_{c=1}^{C} X_c\left(i, j\right) \tag{16}$$

$$F_{max}^{8}(i, j) = \max_{c} X_c(i, j) \tag{17}$$

4 Experimental Design

4.1 Dataset

We evaluated the effectiveness of the TriLVM-UNet model for medical image segmentation using three publicly available datasets: the ACDC cardiac 3D MRI dataset, the LiTS liver tumor 3D CT dataset, and the

BraTS brain tumor 3D MRI dataset. For each dataset, 60% were used for training, 20% for validation, and 20% for testing (Table 1).

Table 1: Detailed descriptions of the three datasets.

Name	Imaging Modality	Number of Classes	Dataset Size	Training Set/ Validation Set/ Test Set	Data Availability
ACDC MRI Dataset	Cardiac cine-MRI	4	200	140/20/40	Public
LiTS CT Dataset	Contrast-enhanced abdominal CT	3	201	131/-/70	Partially public
BraTS MRI Dataset	Multimodal brain MRI	4	1470	1470/-/Private	Partially public

ACDC MRI Dataset [43]: The ACDC dataset is a cardiac MRI database, annotated for three anatomical structures: the left ventricle blood pool, the right ventricle blood pool, and the myocardium. It contains data from 100 patients, comprising a total of 200 3D volumes. The training set includes 70 patients (140 3D volumes), the validation set consists of 10 patients (20 3D volumes), and the test set comprises 20 patients (40 3D volumes).

LiTS CT Dataset [44]: The LiTS dataset is a publicly available abdominal CT image database that includes two anatomical structures: the liver and liver tumors. It comprises a total of 201 contrast-enhanced CT scans. The training set contains 131 cases with publicly available annotations, whereas the validation and test sets, comprising 70 cases, do not have publicly released labels. This dataset is widely used for liver and tumor segmentation tasks.

BraTS MRI Dataset [45]: The BraTS dataset is a multimodal brain MRI database, annotated for three tumor sub-regions: necrotic/non-enhancing tumor core, peritumoral edema, and enhancing tumor. It comprises a total of 1470 (3D) scans. The combined training and validation set includes all 1470 cases with publicly available annotations, whereas the test set labels are private and not publicly released. This dataset is widely used for brain tumor segmentation tasks.

4.2 Experimental Environment

To evaluate the effectiveness and competitiveness of the TriLVM-UNet model in medical image segmentation, we selected 3D datasets that have demonstrated superior performance in various segmentation tasks. During data preprocessing, a transformation pipeline was applied to the training set, including random cropping to 256 × 256 pixels, whereas the validation and test sets were resized to a fixed size. Both training and evaluation sets underwent the same normalization procedure. The initial learning rate was set to 0.01, with a weight decay of 1×10^{-8}, and the cross-entropy loss function was employed. Training was performed for a maximum of 30 epochs, and an early stopping strategy was implemented to terminate training if the validation accuracy plateaued for a predetermined number of consecutive epochs. Model validation was conducted at an interval of 30 epochs. The total number of epochs was set to 30, and testing was performed immediately after the completion of the final epoch. Regarding the preprocessing pipeline, all images were resized to 256 × 256 pixels. First, the corresponding subdirectory was accessed according to the data split. All patient folders were traversed, and for each patient, every temporal NIfTI image was paired with its corresponding label. For each three-dimensional volume, slices were extracted sequentially along the slice dimension. Slices containing fewer than 50 foreground pixels were discarded as largely empty, and the slice

indices could be further restricted by a predefined range. The retained slices were stored in a memory cache together with their patient identifier, slice index, original spatial spacing, and original shape. When a sample was fetched, the slice image and label were converted to tensors; the image remained single-channel, while the label was merged into a binary mask according to the target class set. All foreground classes were mapped to 1 and the background to 0, yielding a floating-point mask of shape (1, H, W). During training, three types of online augmentation were applied probabilistically: random rotation within ±15° (bilinear interpolation for the image and nearest-neighbor interpolation for the label), random horizontal flipping, and random brightness scaling with a factor uniformly drawn from [0.9, 1.1]. Following augmentation, slice-wise adaptive normalization was performed. Specifically, the mean and standard deviation were computed exclusively over pixels with intensity greater than zero; Z-score normalization was applied, the values were clipped to [−3.0, 3.0], and subsequently linearly rescaled to [0, 1]. Finally, the single-channel image was replicated three times along the channel dimension to construct a pseudo-RGB three-channel input. Each sample was ultimately returned as a dictionary containing the three-channel image, binary mask, patient identifier, slice index, original shape, and related metadata.

Hardware Configuration: Experiments were conducted on a system running Ubuntu 20.04.3, equipped with two NVIDIA A800 80 GB PCIe GPUs. Software Environment: The model was implemented using the PyTorch deep learning framework (Python 3.8, PyTorch 1.13.0, CUDA 11.7). Training for 30 epochs took 40.99 min. Memory utilization stabilized at 1.49% with total memory of 80 GB. GPU utilization stabilized between 385.46% and 466.33%.

4.3 Evaluation Metrics

To comprehensively quantify model performance, six standard evaluation metrics were adopted in this study: the Dice similarity coefficient (DSC) measures the overlap between the predicted segmentation and the ground truth, defined as the ratio of twice the true positives to the sum of the predicted positives and the true positives, with a higher DSC indicating more reliable overlap; the Intersection over Union (IoU), also known as the Jaccard index, quantifies the ratio of the intersection to the union of the predicted and ground truth regions, calculated as true positives divided by the sum of true positives, false positives, and false negatives, with a higher IoU signifying better spatial agreement; accuracy (Acc) measures the overall correctness of the prediction results, defined as the proportion of correctly predicted voxels among all voxels, with higher Acc indicating more reliable overall segmentation; sensitivity (Se), also called recall, reflects the model's ability to identify positive samples, defined as the proportion of true positives among actual positive samples, with higher Se indicating a stronger ability to detect positive samples; specificity (Sp) evaluates the model's ability to exclude negative samples, defined as the proportion of true negatives among actual negative samples, with higher Sp indicating a stronger ability to exclude negative samples; and Param denotes the total number of trainable parameters, reflecting model complexity and computational cost, with a lower Param indicating a more lightweight and efficient model.

$$IoU = \frac{TP}{TP + FP + FN} \tag{18}$$

$$\mathrm{DSC} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{19}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{20}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

$$Param = C_o \times (k_w \times k_h \times C_i + 1) \tag{23}$$

5 Comparative Results

5.1 TriLVM-UNet Model Performance

5.1.1 Objective Experimental Results

To comprehensively and objectively evaluate the performance of the TriLVM-UNet model, we conducted comparative experiments on three datasets. (Note that certain earlier UNet models reported outdated metrics, which are excluded from our comparison.) The compared models are broadly classified into UNet-based and Mamba-based architectures. Table 2 presents a comparative evaluation of the TriLVM-UNet model against other models on the ACDC dataset. Specifically, the TriLVM-UNet model achieves an Acc of 98.57%, Sp of 99.15%, and Se of 95.28% on this dataset. Here, Acc, Se, and Sp denote accuracy, sensitivity, and specificity, respectively.

Table 2: Performance comparison of the TriLVM-UNet model with reference models on the ACDC dataset.

Dataset	Model	IoU (%)	DSC (%)	Acc (%)	Sp (%)	Se (%)	Param (M)
ACDC	UltraLight VM-UNet [22]	36.95	53.96	87.63	92.91	53.74	**0.038**
	MALUNet [26]	37.38	54.42	98.30	99.10	55.43	0.175
	H-vmunet [15]	16.28	–	57.29	56.62	61.58	1.832
	UNet [7]	–	–	95.31	**99.91**	–	4.096
	VM-UNet V2 [8]	54.92	68.42	89.06	88.38	89.48	22.77
	Swin-Unet [29]	24.57	–	98.19	98.99	75.76	27.80
	TriLVM-UNet	**90.95** ± 0.24	**95.26** ± 0.13	**98.57** ± 0.03	99.15 ± 0.11	**95.28** ± 0.73	0.068

Note: Bold values indicate the best results.

Table 3 presents the evaluation results of the TriLVM-UNet model and reference models on the LiTS dataset. Specifically, the TriLVM-UNet model achieves an Acc of 99.46% and a Sp of 99.68% on the LiTS dataset.

Table 3: Performance comparison of the TriLVM-UNet model with reference models on the LiTS dataset.

Dataset	Model	IoU (%)	DSC (%)	Acc (%)	Sp (%)	Se (%)	Param (M)
LiTS	UltraLight VM-UNet [22]	38.07	55.15	96.56	98.04	57.63	**0.038**
	MALUNet [26]	47.12	64.05	96.03	97.89	64.15	0.175
	H-vmunet [15]	57.94	73.37	97.06	98.87	68.36	1.832
	UNet [7]	–	–	98.00	**100**	–	4.096
	VM-UNet V2 [8]	49.40	66.13	97.73	99.16	60.30	22.77
	Swin-Unet [29]	75.22	84.08	96.98	99.40	79.03	27.80
	TriLVM-UNet	**84.81** ± 4.63	**91.73** ± 2.76	**99.46** ± 0.18	99.68 ± 0.12	**92.80** ± 2.27	0.068

Note: Bold values indicate the best results.

Table 4 presents the evaluation results of the TriLVM-UNet model and reference models on the BraTS dataset. Specifically, the TriLVM-UNet model achieves an Acc of 98.59% and a Sp of 99.14% on the BraTS dataset.

Table 4: Performance comparison of the TriLVM-UNet model with reference models on the BraTS dataset.

Dataset	Model	IoU (%)	DSC (%)	Acc (%)	Sp (%)	Se (%)	Param (M)
BraTS	UltraLight VM-UNet [22]	50.63	67.22	98.11	99.06	66.63	**0.038**
	MALUNet [26]	72.48	84.05	**99.80**	99.89	**85.45**	0.175
	H-vmunet [15]	51.17	67.70	98.42	99.68	56.72	1.832
	UNet [7]	–	–	97.08	**100**	–	4.096
	VM-UNet V2 [8]	50.80	67.38	98.14	99.10	65.96	22.77
	Swin-Unet [29]	**76.37**	**84.82**	98.36	99.31	83.10	27.80
	TriLVM-UNet	63.48 ± 6.26	77.51 ± 4.88	98.59 ± 0.34	99.14 ± 0.24	80.92 ± 4.33	0.068

Note: Bold values indicate the best results.

ACDC Dataset Result Analysis: On the ACDC dataset, TriLVM-UNet demonstrates comprehensive and outstanding advantages. Its IoU reaches as high as 90.95%, which is approximately 36 percentage points higher than the second-best VM-UNet V2 at 54.92%, and its DSC reaches 95.26%, far exceeding VM-UNet V2′s 68.42%, indicating that the model achieves excellent segmentation overlap and contour agreement for the target region. The Acc reaches 98.57%, surpassing the classic UNet's 95.31% and Swin-Unet's 98.19%, while the Sp is 99.15%, second only to UNet (99.91%) which lacks IoU and DSC metrics, yet still at a top level, demonstrating an extremely low false positive rate for the background. Most critically, the Se reaches as high as 95.28%, not only far exceeding the second-best VM-UNet V2′s 89.48% but also nearly 20 percentage points higher than Swin-Unet's 75.76%, proving its exceptionally strong ability to capture positive samples with minimal risk of missed detection. Parameter efficiency is another core advantage of this model. TriLVM-UNet has only 0.068M parameters, approximately 1/335 of VM-UNet V2 (22.77M) and also far smaller than Swin-Unet (27.80M) and UNet (4.096M), achieving the highest accuracy while being extremely lightweight, making it highly suitable for resource-constrained scenarios. Furthermore, the standard deviations of all metrics are relatively low, reflecting highly stable performance and insensitivity to data fluctuations. Overall, with an extremely low parameter count, TriLVM-UNet achieves comprehensive leadership in segmentation accuracy, sensitivity, specificity, and stability on the ACDC dataset, balancing the dual demands of high performance and lightweight deployment.

LiTS Dataset Result Analysis: On the LiTS dataset, TriLVM-UNet achieves an IoU of 84.81% and a DSC of 91.73%, which are 9.59 and 7.65 percentage points higher than those of the second-best model in the table, Swin-Unet (IoU 75.22%, DSC 84.08%), respectively. Its Se is 92.80%, 13.77 percentage points higher than Swin-Unet's 79.03%, indicating more complete identification of lesion areas and fewer missed detections. The Acc is 99.46% and Sp is 99.68%, both at high levels, and the Sp is close to that of UNet (100%), which has the highest Sp, demonstrating good background suppression capability. The number of parameters is only 0.068M, far lower than other models, such as Swin-Unet's 27.80M and VM-UNet V2′s 22.77M, and even smaller than lightweight models like MALUNet (0.175M) and UltraLight VM-UNet (0.038M, but its IoU is only 38.07%), achieving the highest segmentation accuracy while maintaining a low parameter count. The small standard deviations of all metrics in cross-validation indicate that the model has stable performance and good generalization. Overall, TriLVM-UNet achieves the best results in core segmentation metrics such as IoU, DSC, and Se, while having an extremely small number of parameters, striking an effective balance between accuracy and lightweight design.

BraTS Dataset Result Analysis: On the BraTS dataset, TriLVM-UNet, with an extremely low parameter count of only 0.068M, achieves competitive overall performance while revealing a gap in segmentation overlap compared with the best models. Its Acc reaches 98.59%, surpassing Swin-Unet's 98.36%; its Sp is 99.14%, close to MALUNet's 99.89% and UNet's 100%, indicating excellent control of background misclassification; its Se is 80.92%, which, although lower than MALUNet's 85.45% and Swin-Unet's 83.10%, is still significantly better than UltraLight VM-UNet's 66.63% and H-vmunet's 56.72%, demonstrating that its ability to recognize positive samples is leading among lightweight models. In addition, the standard deviations of all metrics are relatively small, reflecting good robustness and stability. However, the model's weakness lies in the precise agreement of segmentation: its IoU is 63.48%, clearly behind Swin-Unet's 76.37% and MALUNet's 72.48%; its DSC is 77.51%, also lower than Swin-Unet's 84.82% and MALUNet's 84.05%, indicating insufficient overlap and contour consistency between the predicted regions and the ground truth, with noticeable over-segmentation or under-segmentation. Overall, on BraTS, TriLVM-UNet achieves the best accuracy and stability among models of comparable size with its extremely low parameter count, but compared with MALUNet and Swin-Unet, there is still clear room for improvement in segmentation completeness and precise contour delineation. Figs. 9 and 10 show the bar chart and box plot of the three datasets, respectively.

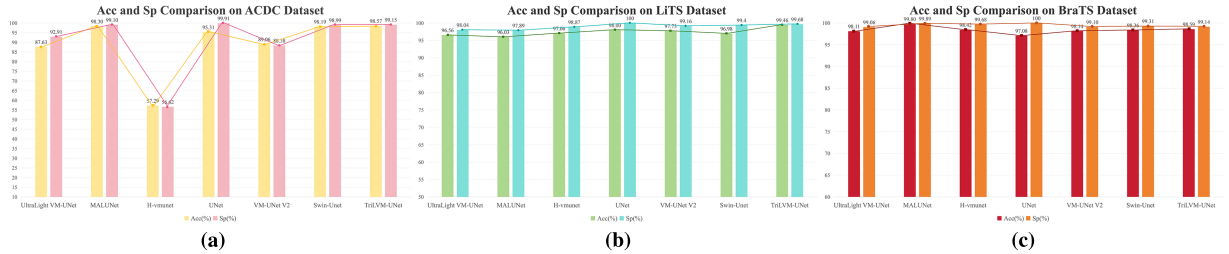

Figure 9: Model analysis diagrams for the three datasets. (**a–c** show the bar charts of ACDC, LiTS, and BraTS, respectively).

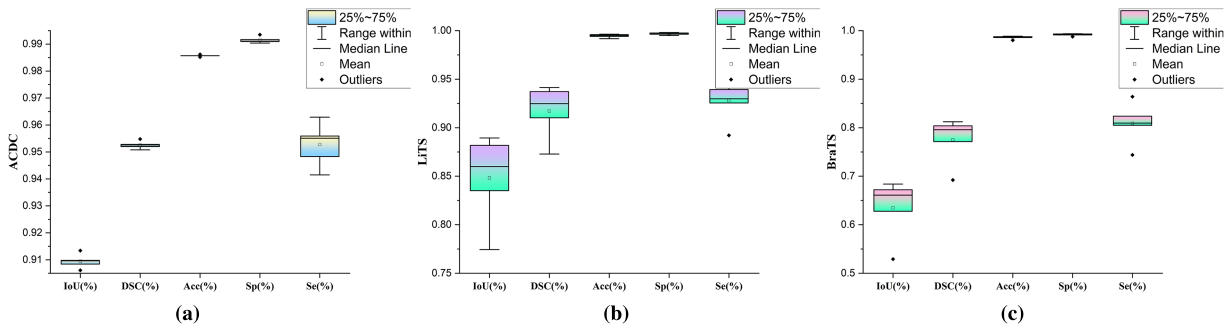

Figure 10: Box plot for analysis of three datasets. (**a–c** show the box plots for ACDC, LiTS, and BraTS, respectively).

5.1.2 Subjective Experimental Results

On the ACDC dataset, the models show relatively distinct performance differences. UltraLight VM-UNet and MALUNet have very small parameter counts, but their segmentation accuracy is limited, with weak capability in identifying target regions and frequent missed detections, making it difficult to adequately meet the clinical need for segmentation completeness. The performance of H-vmunet is even weaker, with low overlap between the segmentation results and the ground truth annotations; its ability to capture the foreground and distinguish the background is insufficient, leading to suboptimal segmentation outcomes. The classic UNet achieves acceptable overall classification accuracy and background specificity, but due to the

absence of some key segmentation metrics, its performance in boundary delineation or small target capture is difficult to evaluate. VM-UNet V2 has a considerably larger number of parameters, which brings relatively balanced segmentation performance and some improvement in target coverage and boundary alignment; however, the increased parameter count diminishes its lightweight nature. Swin-Unet incorporates the Transformer architecture, yet its segmentation accuracy is unsatisfactory, with low target overlap. Although it maintains high specificity, its sensitivity is only moderate, suggesting that the potential of the self-attention mechanism was not fully exploited in this medical image segmentation task, and the large parameter size further affects its practicality. In contrast, TriLVM-UNet achieves good segmentation performance with a small parameter scale. It provides relatively complete target coverage, and its segmentation boundaries closely adhere to the true contours; both missed detections and false positives are kept at low levels, while a good balance between sensitivity and specificity is maintained. This indicates that the model possesses strong discriminative ability in feature extraction and lesion localization, can capture anatomical variations at a low computational cost, and achieves a favorable trade-off between lightweight design and segmentation accuracy, demonstrating good subjective visual quality and clinical application potential. As the ACDC dataset is stored as image slices, representative segmentation examples are visualized in Fig. 11.

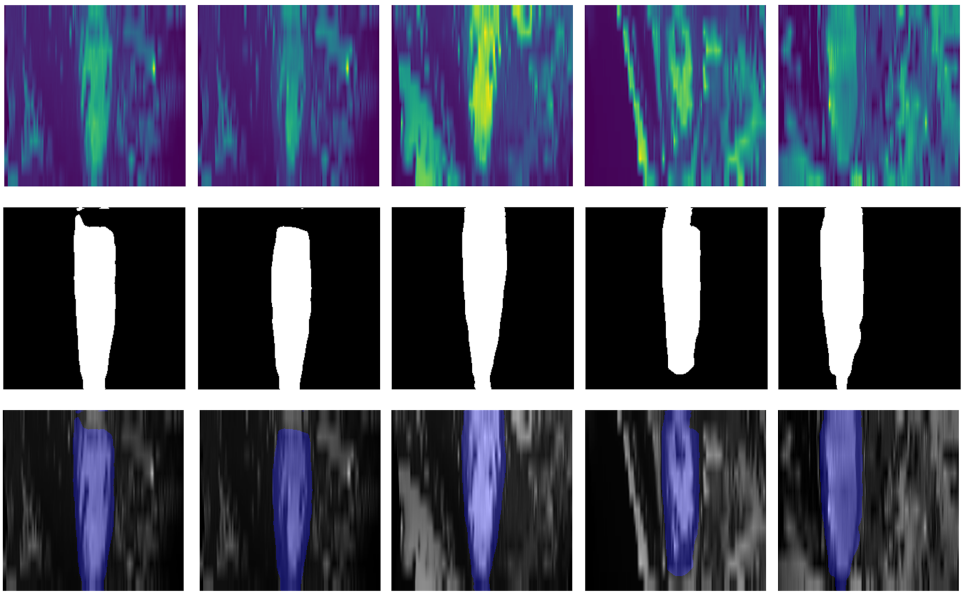

Figure 11: ACDC segmentation visualization.

The segmentation results of TriLVM-UNet on the LiTS dataset demonstrate several advantages. In terms of overall segmentation accuracy, the model clearly outperforms other methods in the table in both regional overlap and composite score, indicating that its predictions are highly consistent with the ground truth in spatial location and contour shape, and that it can localize lesion areas with relatively high precision. The significantly higher sensitivity shows that the model covers lesions more completely and effectively controls missed detections, maintaining good recognition capability even in regions with blurred boundaries or low contrast. At the same time, the specificity remains at a high level, close to that of the best-performing model in background suppression, reflecting fewer false positives in normal tissue regions and an effective ability to distinguish lesions from the background. Notably, in terms of parameter count, TriLVM-UNet is considerably lower than most methods in the table, and is even more compact than several purpose-built lightweight networks. This means that while maintaining extremely low computational

complexity and storage requirements, it achieves the best segmentation performance in the current list, striking a favorable balance between accuracy and resource consumption. In contrast, some models with fewer parameters suffer from clearly insufficient segmentation accuracy, while others with higher accuracy involve a large number of parameters, which is not conducive to practical deployment. Moreover, the small fluctuation ranges of the metrics in cross-validation indicate that the model performs consistently across different data subsets, with relatively reliable generalization ability and stability. Overall, the model exhibits satisfactory performance in accurately capturing lesions, suppressing irrelevant backgrounds, reducing missed detections, and maintaining computational efficiency. These characteristics make it particularly suitable for application scenarios that require fast and precise segmentation, offering an effective reference solution for lightweight medical image segmentation. Fig. 12 shows the visualization of LiTS segmentation.

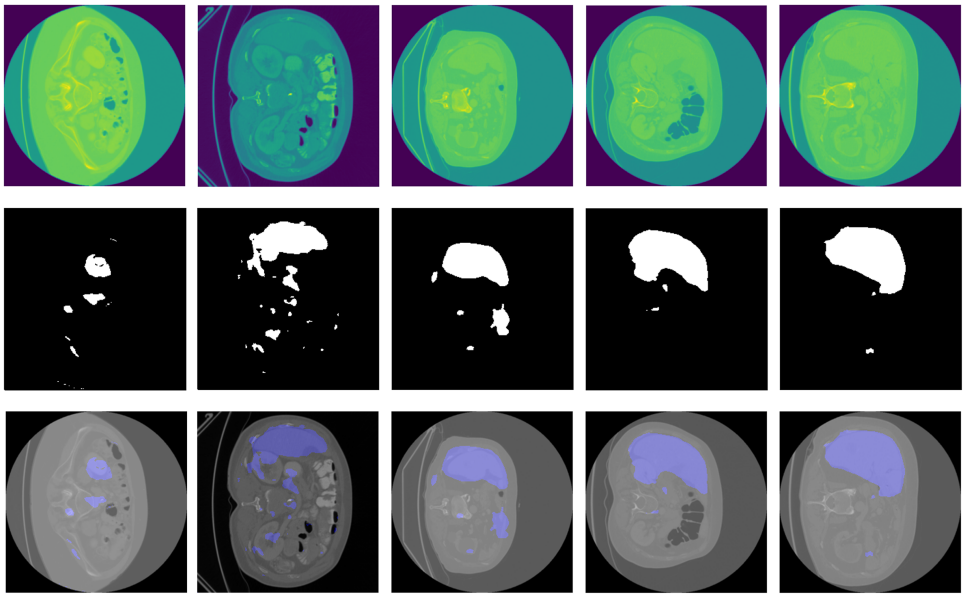

Figure 12: LiTS segmentation visualization.

On the BraTS dataset, the models exhibit a clear performance hierarchy. UltraLight VM-UNet and VM-UNet V2 both deliver unsatisfactory segmentation, with poor tumor coverage and boundary adherence; the large parameter count of VM-UNet V2 brings no gain, resulting in low cost-effectiveness. H-vmunet performs similarly, with weak target capture and notable missed detections. UNet achieves excellent background specificity but lacks overlap-based evaluation, making tumor completeness and boundary precision unassessable. MALUNet stands out among lightweight models, attaining markedly improved overlap, well-controlled false positives, and relatively clear boundaries. Swin-Unet achieves the best overall segmentation with the highest overlap and tightest boundaries, though at the highest deployment cost. TriLVM-UNet demonstrates good global segmentation efficiency with an extremely low parameter scale, yet its boundary delineation still requires improvement; moreover, performance fluctuates across samples, indicating that stability needs to be strengthened. In summary, MALUNet offers a favorable trade-off between efficiency and accuracy, Swin-Unet pursues maximal segmentation quality, and TriLVM-UNet maintains competent global segmentation but must enhance edge precision and consistency. Fig. 13 shows the visualization of BraTS segmentation.

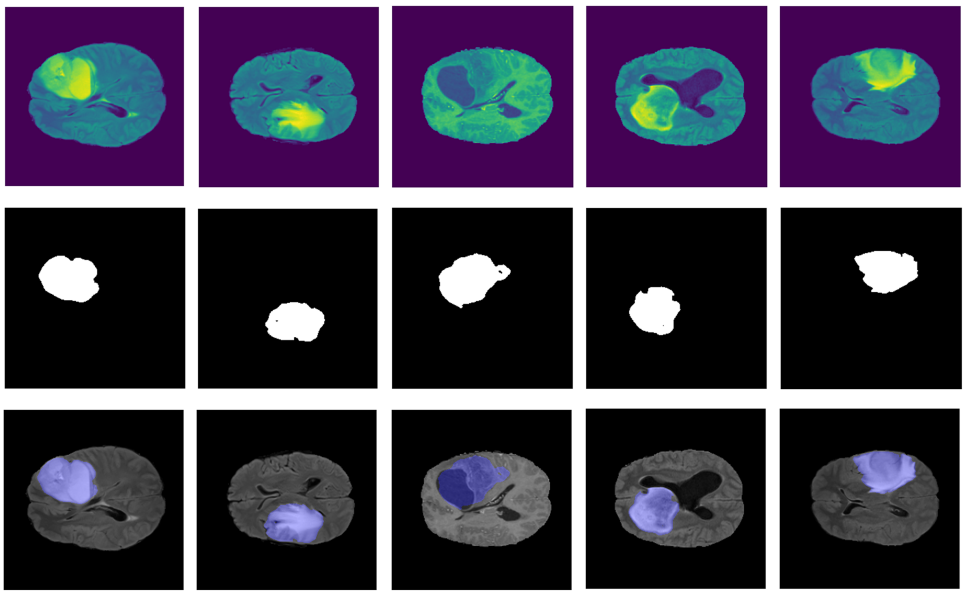

Figure 13: BraTS segmentation visualization.

TriLVM-UNet achieved relatively good segmentation performance with a low parameter count on all three datasets. On the ACDC dataset, the model provided relatively complete coverage of the target region, with segmentation boundaries closely fitting the ground truth contours. Both missed detections and false positives were kept at low levels, and a good balance between sensitivity and specificity was maintained, achieving a favorable trade-off between lightweight design and accuracy. On the LiTS dataset, the model demonstrated strong target capture capability with few missed detections and high sensitivity, while also maintaining a low false positive rate and high specificity; however, the evaluation metrics exhibited some fluctuation across different samples, indicating that the stability and consistency of the segmentation could be further improved. On the BraTS dataset, the model achieved relatively complete target coverage, good boundary adherence, and acceptable missed-detection control, but the segmentation results also exhibited certain sample-to-sample fluctuations, and the stability still requires improvement. Overall, while maintaining an extremely lightweight parameter scale, TriLVM-UNet exhibits strong target capture capability and relatively complete region coverage, with segmentation boundaries that closely match the ground truth contours. Nevertheless, on some datasets the segmentation consistency is insufficient, leaving room for further improvement in the model's robustness.

5.2 Ablation Study

Our study conducts systematic ablation experiments focusing on the key modules of the proposed network architecture, aiming to validate, from a structural perspective, the individual contributions and synergistic interactions of each functional unit to overall performance improvement. The ablation strategy follows a module-by-module removal approach combined with structural comparative analysis. By selectively removing or modifying specific modules while keeping the overall network architecture consistent, we observe corresponding changes in model performance, thereby revealing the sources of performance gains. First, an ablation study is conducted on the LVM module in the encoding stage. As a critical component responsible for modeling global dependencies within the network, the primary function of the LVM is to address the limitations of traditional convolutions in capturing long-range relationships. By constructing comparative schemes that remove or attenuate the LVM, we can assess the extent of degradation in structural

coherence and semantic representation in the absence of global modeling capability, thereby validating the importance of global contextual information for medical image segmentation tasks. Subsequently, an independent ablation is conducted on the SAB module within the bridge layer. The SAB primarily facilitates structural alignment and semantic enhancement of cross-scale features, functioning as a critical structure-aware unit that bridges the encoding and decoding stages. By removing the SAB while preserving multi-scale fusion and attention mechanisms, we can evaluate the impact of the structure-guiding mechanism on boundary detail recovery and spatial continuity, thereby confirming its essential role in semantic transition and feature recombination. Finally, a joint ablation is conducted on the MSDC and CBAM modules. This combined module is responsible for multi-scale feature integration and attention enhancement across both channel and spatial dimensions, thereby improving the model's perception of lesion regions at different scales. By removing these modules while retaining the SAB, we can assess the contribution of the multi-scale enhancement mechanism to fine-grained feature representation and small-target recognition. Table 5 presents the design rationale of the ablation study.

Table 5: Ablation study design.

No.	Module	Ablation	Purpose of Design
Scheme 1	Encoder Stage: LVM + LVM + LVM Bridge Layer: SAB + MSDC + CBAM Decoder Stage: Conv + LVM	LVM	To investigate the key role of the LVM module in multi-scale feature extraction and long-range dependency modeling within the encoding path, we reduced the number of stacked LVM modules and observed the resulting performance degradation, thereby validating the necessity of the LVM module for deep feature learning.
Scheme 2	Encoder Stage: LVM + 2LVM + 3LVM Bridge Layer: MSDC + CBAM Decoder Stage: Conv + LVM	SAB	To evaluate the contribution of the SAB in cross-stage feature fusion, the SAB is designed to dynamically integrate feature maps from different levels, enhancing the interaction of multi-scale contextual information. By removing this module, we can assess whether the model's ability to fuse multi-scale information is degraded, thereby verifying the necessity of the SAB.
Scheme 3	Encoder Stage: LVM + 2LVM + 3LVM Bridge Layer: SAB Decoder Stage: Conv + LVM	MSDC + CBAM	To validate the role of MSDC and CBAM in refining features and enhancing responses in important regions, MSDC captures multi-scale local details through depthwise convolutions with different receptive fields, while CBAM emphasizes key information via channel and spatial attention. By removing these two components, we can observe changes in the model's sensitivity to boundary details and lesion regions.

Scheme 1 Result Analysis: To validate the effectiveness of each module in the proposed model, we conducted experiments by reducing the number of LVM layers (Table 6). Specifically, the two LVM layers in Stage Two and the three LVM layers in Stage Three were each replaced with a single LVM layer for experimentation using the ACDC dataset. The results are as follows: On the ACDC dataset, compared with the ablated model (with a single LVM layer), TriLVM-UNet achieved improvements of 19.10 percentage points in DSC, 5.67 percentage points in Acc, 4.88 percentage points in Sp, and 11.20 percentage points in Se, demonstrating a clear overall enhancement. On the LiTS dataset, compared with the ablated model, TriLVM-UNet showed an improvement of 35.02 percentage points in DSC, 6.10 percentage points in Acc, 5.06 percentage points in Sp, and 19.46 percentage points in Se, with sensitivity exhibiting a particularly notable increase. In contrast, on the BraTS dataset, all metrics of TriLVM-UNet decreased relative to the ablated model, indicating certain limitations in the model's generalization capability. These results demonstrate that

stacking LVM layers has a more pronounced effect on segmentation performance for the ACDC and LiTS datasets than for the BraTS dataset.

Table 6: Comparative analysis of ablation scheme 1 and the baseline model.

Dataset	Model	DSC (%)	Acc (%)	Sp (%)	Se (%)
ACDC	TriLVM-UNet	**95.26**	**98.57**	**99.15**	**95.28**
	Scheme 1	76.16	92.90	94.27	84.08
LiTS	TriLVM-UNet	**91.73**	**99.46**	**99.68**	**92.80**
	Scheme 1	56.71	93.36	94.62	73.34
BraTS	TriLVM-UNet	77.51	98.59	99.14	80.92
	Scheme 1	**92.88**	**99.70**	**99.85**	**92.66**

Note: Bold values indicate the best results.

Scheme 2 Result Analysis: To validate the effectiveness of each module in the proposed model, we conducted a second ablation experiment by removing the SAB block (Table 7). Specifically, the SAB module was removed from the skip connections, while retaining only the MSDC and CBAM modules for the experiment using the ACDC dataset. The results are as follows: On the ACDC dataset, compared with the ablated model (with the SAB module removed), TriLVM-UNet achieved improvements of 18.66 percentage points in DSC, 5.32 percentage points in Acc, 4.12 percentage points in Sp, and 13.44 percentage points in Se, demonstrating an overall performance enhancement. On the LiTS dataset, compared with the ablated model, TriLVM-UNet achieved improvements of 28.93 percentage points in DSC, 4.85 percentage points in Acc, 3.95 percentage points in Sp, and 16.02 percentage points in Se, demonstrating an overall performance enhancement. On the BraTS dataset, compared with the ablated model, TriLVM-UNet achieved improvements of 15.43 percentage points in DSC, 1.08 percentage points in Acc, 0.81 percentage points in Sp, and 10.91 percentage points in Se, demonstrating a substantial overall enhancement. These results illustrate the importance of the SAB module for segmentation performance across all three datasets.

Table 7: Comparative analysis of ablation scheme 2 and the baseline model.

Dataset	Model	DSC (%)	Acc (%)	Sp (%)	Se (%)
ACDC	TriLVM-UNet	**95.26**	**98.57**	**99.15**	**95.28**
	Scheme 2	76.60	93.25	95.03	81.84
LiTS	TriLVM-UNet	**91.73**	**99.46**	**99.68**	**92.80**
	Scheme 2	62.80	94.61	95.73	76.78
BraTS	TriLVM-UNet	**77.51**	**98.59**	**99.14**	**80.92**
	Scheme 2	62.08	97.51	98.33	70.01

Note: Bold values indicate the best results.

Scheme 3 Result Analysis: To validate the effectiveness of each module in the proposed model, we conducted a third ablation experiment by removing the MSDC and CBAM blocks (Table 8). Specifically, the MSDC and CBAM modules were removed from the skip connections, while retaining only the SAB module for the experiment using the ACDC dataset. The results are as follows: On the ACDC dataset, compared with the ablated model (with the MSDC and CBAM modules removed), TriLVM-UNet achieved improvements of 20.69 percentage points in DSC, 5.80 percentage points in Acc, 4.15 percentage points in Sp, and 16.77 percentage points in Se, demonstrating an overall performance enhancement. On the LiTS

dataset, compared with the ablated model, TriLVM-UNet showed improvements of 45.65 percentage points in DSC, 10.56 percentage points in Acc, 10.22 percentage points in Sp, and 12.76 percentage points in Se, indicating a substantial overall improvement. On the BraTS dataset, compared with the ablated model, TriLVM-UNet achieved improvements of 17.56 percentage points in DSC, 1.50 percentage points in Acc, 1.38 percentage points in Sp, and 6.21 percentage points in Se, also demonstrating an overall enhancement. These results illustrate that the MSDC and CBAM modules contribute positively to segmentation metrics across all three datasets, highlighting that these modules are integral and inseparable components of the TriLVM-UNet model.

Table 8: Comparative analysis of ablation scheme 3 and the baseline model.

Dataset	Model	DSC (%)	Acc (%)	Sp (%)	Se (%)
ACDC	TriLVM-UNet	**95.26**	**98.57**	**99.15**	**95.28**
	Scheme 3	74.57	92.77	95.00	78.51
LiTS	TriLVM-UNet	**91.73**	**99.46**	**99.68**	**92.80**
	Scheme 3	46.08	88.90	89.46	80.04
BraTS	TriLVM-UNet	**77.51**	**98.59**	**99.14**	**80.92**
	Scheme 3	59.95	97.09	97.76	74.71

Note: Bold values indicate the best results.

Based on the results of the three ablation experiment tables, the complete design of TriLVM-UNet outperforms each ablated scheme in most cases, confirming the necessity of the synergistic cooperation among the LVM layers, the SAB module, and the MSDC and CBAM modules. On the ACDC and LiTS datasets, reducing the number of LVM layers, removing the SAB module, or removing the MSDC and CBAM modules all led to a marked decline in segmentation performance, with the weakening of boundary adherence and target capture capability being the most prominent, reflecting that hierarchical feature representation, spatial attention refinement, and multi-scale channel semantic enhancement are all indispensable for accurate segmentation on these two datasets. In contrast, the BraTS dataset exhibited a different trend: reducing the number of LVM layers caused several metrics to surpass those of the full model, suggesting that excessively deep LVM layers may introduce feature redundancy or overfitting on this dataset; however, removing the SAB module or the MSDC and CBAM modules still resulted in performance degradation, indicating that spatial localization refinement and discriminative feature learning also contribute positively to BraTS segmentation. Overall, the contribution of each module to segmentation performance exhibits a certain degree of task dependence: the depth of the LVM needs to be flexibly adjusted according to data characteristics, while the gains from the SAB, MSDC, and CBAM modules are more stable. In summary, TriLVM-UNet, through its complementary multi-module design, achieves effective adaptation to segmentation tasks on different datasets within a lightweight framework, balancing accuracy and generalization. Fig. 14 is the analysis diagrams of ablation 1–3, respectively.

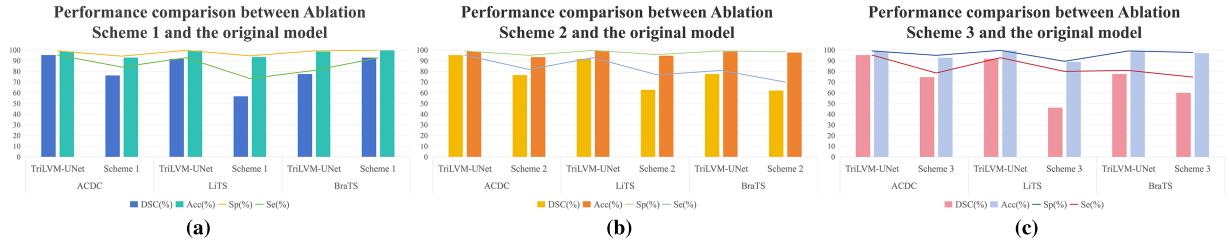

Figure 14: Performance analysis chart of ablation scheme 1–3 and the original model. (**a**–**c** are the bar and line analysis charts for Scheme 1, Scheme 2, and Scheme 3, respectively).

After removing the LVM module, the DSC plummeted to 76.16%, and the sensitivity dropped from 95.28% of the full model to 84.08%, indicating that LVM is the core of hierarchical feature extraction and its absence severely weakens the model's ability to capture target regions. Removing the SAB module reduced the DSC to 76.60%. Although the specificity remained acceptable, the sensitivity declined to 81.84%, demonstrating that spatial attention is crucial for improving boundary adherence and ensuring complete target coverage; its absence leads to a significant increase in missed detections. Simultaneously removing the MSDC and CBAM modules resulted in the lowest DSC, with a sensitivity of only 78.51%, reflecting that once multi-scale features and the channel attention mechanism are stripped away, the model's adaptability to lesion scales and its ability to discriminate key features deteriorate severely. From the perspective of inter-module relationships, these three components form a progressively complementary and synergistic structure. LVM provides hierarchical, multi-granularity base feature representations and serves as the foundation for subsequent modules to function effectively. Building on this, SAB refines the features from the spatial dimension, enhancing lesion edge details and suppressing background interference, thereby compensating for LVM's deficiencies in localization precision. MSDC and CBAM enhance feature semantic discrimination from the multi-scale and channel attention dimensions, respectively: the former aggregates contextual information from different receptive fields, while the latter adaptively calibrates channel responses, effectively mitigating the problems of high missed-detection rates and scale insensitivity when LVM is used alone. These three components do not operate independently or in parallel; LVM ensures information completeness, SAB improves spatial accuracy, and MSDC and CBAM enrich semantic discrimination. All are indispensable and jointly propel the model toward high accuracy and high robustness. Table 9 presents the segmentation performance of different ablation experiment settings and our model on the ACDC dataset.

Table 9: Quantitative ablation study results on the ACDC dataset.

Dataset	Model	LVM	SAB	MSDC	CBAM	Evaluation Metrics			
						DSC	Acc	Sp	Se
ACDC	Scheme 1	√	×	×	×	76.16	92.90	94.27	84.08
	Scheme 2	×	√	×	×	76.60	93.25	95.03	81.84
	Scheme 3	×	×	√	√	74.57	92.77	95.00	78.51
	TriLVM-UNet	×	×	×	×	**95.26**	**98.57**	**99.15**	**95.28**

Note: Bold values indicate the best results.

6 Conclusion

This study proposes the TriLVM-UNet framework, aiming to introduce a novel segmentation model to address the challenges of balancing global dependency modeling and computational efficiency in medical image segmentation. The design of TriLVM-UNet seeks to better meet the demands of real-time clinical applications. Its core contributions lie in the innovative integration of Multi-scale Depthwise Convolution (MSDC) and the Convolutional Block Attention Module (CBAM) to enhance segmentation efficiency. While significantly improving classification performance, the model overcomes the computational complexity limitations of traditional Transformer-based architectures. The design philosophy of TriLVM-UNet, which balances segmentation accuracy and computational efficiency through a progressive refinement strategy, has been effectively validated on the relatively structured task of cardiac segmentation. However, when addressing the heterogeneity of liver tumors and the multimodal complexity of brain tumors, its information propagation mechanisms and feature extraction capabilities still require further optimization. This observation suggests that future research in medical image segmentation should not only pursue superior performance on a single dataset but also place greater emphasis on cross-task and cross-modal adaptability and robustness. Future work will focus on developing more flexible adaptive architectures that allow the model to dynamically adjust its processing pipeline based on task characteristics, exploring more efficient multi-scale feature fusion mechanisms—particularly for irregular object boundaries—and introducing meta-learning or domain adaptation techniques to enhance rapid adaptability to novel medical image data. In the era of precision medicine, segmentation models capable of balancing accuracy, reliability, and generalization will play an increasingly vital role in clinical diagnosis, surgical planning, and treatment evaluation.

Acknowledgement: Not applicable.

Funding Statement: The Key Project of Ningxia Natural Science Foundation under grant 2025AAC020006, and the Regional Program of National Natural Science Foundation of China under grant 62561002. The Shaanxi University of Technology Foundation Project under grant SLGRCQD2137.

Author Contributions: Kexin Zhang: undertook the majority of the work, covering the core aspects from research conception, method design, experimental execution, and data analysis to the drafting of the manuscript. Lihua Liu: primarily responsible for manuscript revision, research supervision, project management, and funding support. Yuting Xue: assisted with formal analysis and investigation. Tao Zhou: participated in validation and resource support. Fengshuai Yue: assisted with data curation. Ruifeng Du: performed some validation work. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
2. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. In: Proceedings of the Conference on Language Modeling 2024; 2024 Oct 7–9; Philadelphia, PA, USA.
3. Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, et al. VMamba: visual state space model. In: Proceedings of the Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 103031–63.

4. Ma J, Li F, Wang B. U-mamba: enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722. 2024.
5. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
6. Wang Z, Zheng JQ, Zhang Y, Cui G, Li L. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. arXiv:2402.05079. 2024.
7. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention; 2015 Oct 5–9; Munich, Germany. p. 234–41.
8. Zhang M, Yu Y, Gu L, Lin T, Tao X. VM-UNET-V2 rethinking vision mamba UNet for medical image segmentation. In: Proceedings of the 20th International Symposium on Bioinformatics Research and Applications; 2024 Jul 19–21; Kunming, China. p. 335–46.
9. Wang Z, Tao T, Ge Y, Chen Z, Chen T, Ye Z, et al. Weak-mamba-UNet: visual mamba makes CNN and vit work better for scribble-based medical image segmentation. IEEE Trans Biomed Eng. 2026: 1–11. doi:10.1109/TBME.2026.3668882.
10. Pei X, Huang T, Xu C. EfficientVMamba: atrous selective scan for light weight visual mamba. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence; 2025 Feb 25–Mar 4; Philadelphia, PA, USA. p. 6443–51.
11. Liu H, Jia C, Shi F, Cheng X, Chen S. SCSegamba: I lightweight structure-aware vision mamba for crack segmentation in structures. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 29406–16. doi:10.1109/CVPR52734.2025.02738.
12. Yang C, Chen Z, Espinosa M, Ericsson L, Wang Z, Liu J, et al. PlainMamba: improving non-hierarchical mamba in visual recognition; 2024 Nov 25–28; Glasgow, UK.
13. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X. Vision mamba: efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria. p. 62429–42.
14. Zhao L, Ma J, Shao Y, Jia C, Zhao J, Yuan H. MM-UNet: a multimodality brain tumor segmentation network in MRI images. Front Oncol. 2022;12:950706. doi:10.3389/fonc.2022.950706.
15. Wu R, Liu Y, Liang P, Chang Q. H-vmunet: high-order vision mamba UNet for medical image segmentation. arXiv:2403.13642. 2024.
16. Xing Z, Ye T, Yang Y, Liu G, Zhu L. SegMamba: long-range sequential modeling mamba for 3D medical image segmentation. Neurocomputing. 2025;624:129447.
17. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. In: Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention; 2024 Oct 6–10; Marrakesh, Morocco.
18. Wang Y, Li Z, Mei J, Wei Z, Liu L, Wang C, et al. SwinMM: masked multi-view with swin transformers for 3D medical image segmentation. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. p. 486–96.
19. Shaker A, Maaz M, Rasheed H, Khan S, Yang M-H, Khan FS. UNETR++: delving into efficient and accurate 3D medical image segmentation. IEEE Trans Med Imaging. 2024;43(9):3377–90. doi:10.1109/tmi.2024.3398728.
20. Revach G, Shlezinger N, Ni X, Escoriza AL, van Sloun RJG, Eldar YC. KalmanNet: neural network aided Kalman filtering for partially known dynamics. IEEE Trans Signal Process. 2022;70:1532–47. doi:10.1109/TSP.2022.3158588.
21. Van der Merwe R, Wan EA. The square-root unscented Kalman filter for state and parameter-estimation. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing; 2001 May 7–11; Salt Lake City, UT, USA. p. 3461–4. doi:10.1109/ICASSP.2001.940586.
22. Wu R, Liu Y, Liang P, Chang Q. UltraLight VM-UNet: parallel vision mamba significantly reduces parameters for skin lesion segmentation. Patterns. 2025;6(11):101298. doi:10.1016/j.patter.2025.101298.
23. Wu R, Liu J, Liu Y, Liang P, Ning G, Chang Q. Only positive cases: 5-fold high-order attention interaction model for skin segmentation derived classification. In: Proceedings of the 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2025 Dec 15–18; Wuhan, China.

24. He K, Lian C, Zhang B, Zhang X, Cao X, Nie D, et al. HF-UNet: learning hierarchically inter-task relevance in multi-task U-Net for accurate prostate segmentation in CT images. IEEE Trans Med Imaging. 2021;40(8):2118–28. doi:10.1109/TMI.2021.3072956.
25. Liao W, Zhu Y, Wang X, Pan C, Wang Y, Ma L. LightM-UNet: mamba assists in lightweight UNet for medical image segmentation. arXiv:2403.05246. 2024.
26. Ruan J, Xiang S, Xie M, Liu T, Fu Y. MALUNet: a multi-attention and light-weight UNet for skin lesion segmentation. In: Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2022 Dec 6–8; Las Vegas, NV, USA. p. 1150–6.
27. He H, Zhang J, Cai Y, Chen H, Hu X, Gan Z, et al. MobileMamba: lightweight multi-receptive visual mamba network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 11–15; Nashville, TN, USA. p. 4497–507.
28. Ma C, Wang Z. Semi-Mamba-UNet: pixel-level contrastive and pixel-level cross-supervised visual mamba-based UNet for semi-supervised medical image segmentation. Knowl Based Syst. 2024;300:112203. doi:10.1016/j.knosys.2024.112203.
29. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the Computer Vision—ECCV 2022 Workshops; 2022 Oct 23–27; Tel Aviv, Israel. p. 205–18.
30. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z, et al. Tokens-to-token ViT: training vision transformers from scratch on ImageNet. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 538–47. doi:10.1109/iccv48922.2021.00060.
31. Wu R, Liang P, Huang X, Shi L, Gu Y, Zhu H, et al. MHorUNet: high-order spatial interaction UNet for skin lesion segmentation. Biomed Signal Process Control. 2024;88:105517. doi:10.1016/j.bspc.2023.105517.
32. Rahman MM, Marculescu R. Medical image segmentation via cascaded attention decoding. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 6211–20. doi:10.1109/wacv56688.2023.00616.
33. Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. In: Proceedings of the Tenth International Conference on Learning Representations 2022; 2022 Apr 25–29; Virtual Event.
34. Gupta A, Gu A, Berant J. Diagonal state spaces are as effective as structured state spaces. Adv Neural Inf Process Syst. 2022;35:22982–94. doi:10.52202/068431-1670.
35. Smith JTH, Warrington A, Linderman SW. Simplified state space layers for sequence modeling. In: Proceedings of the Eleventh International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
36. Dao T, Gu A. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria.
37. Lieber O, Lenz B, Bata H, Cohen G, Osin J, Dalmedigos I, et al. Jamba: a hybrid transformer-mamba language model. In: Proceedings of the 13th International Conference on Learning Representations; 2025 Apr 24–28; Singapore.
38. Ma X, Ni Z, Chen X. TinyViM: frequency decoupling for tiny hybrid vision mamba. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2025; 2025 Oct 19–23; Honolulu, HA, USA. p. 23519–29.
39. Sheng J, Zhou J, Wang J, Ye P, Fan J. DualMamba: a lightweight spectral-spatial mamba-convolution network for hyperspectral image classification. IEEE Trans Geosci Remote Sens. 2025;63:1–15. doi:10.1109/tgrs.2024.3516817.
40. Zhou T, Peng C, Guo Y, Wang H, Niu Y, Lu H. Identity-mapping ResFormer: a computer-aided diagnosis model for pneumonia X-ray images. IEEE Trans Instrum Meas. 2025;74:1–12. doi:10.1109/tim.2025.3534218.
41. Zhou T, Chai W, Chang D, Chen K, Zhang Z, Lu H. MambaYOLACT: you only look at mamba prediction head for head-neck lymph nodes. Artif Intell Rev. 2025;58(6):180. doi:10.1007/s10462-025-11177-y.
42. Hasan M, Hasan MK. Hawk-Net: medical image segmentation and classification using multi-scale convolutional self-attention-based image processor with DK-CNN-Mamba-xAttention Fusion Network. Biomed Signal Process Control. 2026;115(6):109401. doi:10.1016/j.bspc.2025.109401.

43. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging. 2018;37(11):2514–25. doi:10.1109/TMI.2018.2837502.
44. Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G, et al. The liver tumor segmentation benchmark (LiTS). Med Image Anal. 2023;84:102680. doi:10.1016/j.media.2022.102680.
45. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. The RSNA-ASNR-MICCAI BraTS, 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314. 2021.