Electronic Research Archive

ERA, 34(9): 6933–6956.
DOI: 10.3934/era.2026301
Received: 22 May 2026
Revised: 07 July 2026
Accepted: 24 July 2026
Published: 12 August 2026

http://www.aimspress.com/journal/era

Research article

WMSM-Net: Wavelet-enhanced Mamba with spatial-channel mutual-feedback attention network for skin lesion segmentation

Xinrong Fu[1], Jianhua Song[1,2,*], Xinying Huang[2], Yu Chen[1] and Hao Liu[1]

[1] College of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China
[2] Key Laboratory of Light Field Manipulation and System Integration Applications in Fujian Province, Minnan Normal University, Zhangzhou 363000, China

* **Correspondence:** Email: songjianhua@mnnu.edu.cn.

Abstract: Accurate skin lesion segmentation remains challenging due to vague boundaries and heavy artifact interference in complex backgrounds. To address these issues, we introduce WMSM-Net, a frequency-space collaborative network built upon the visual state space (Mamba) architecture. WMSM-Net proposes a Wavelet-enhanced Visual State Space Block (W-VSSB) to extract high-frequency boundary details and mitigate edge loss. Concurrently, a Bidirectional Mutual-Feedback Spatial-Channel Attention Bridge (MF-SCAB) was constructed to deliver two-way guidance across spatial and channel dimensions, enhancing the perception of subtle lesions. Feature transitions were further optimized via dynamic gated fusion to suppress background noise. Extensive experiments demonstrated that WMSM-Net achieves state-of-the-art performance, yielding Dice coefficients of 91.01%, 91.08%, and 94.06% for the ISIC 2017, ISIC 2018, and PH2 datasets, respectively. The code is publicly available at https://github.com/fuxinrong2024/WMSM-Net.

Keywords: skin lesion segmentation; Mamba model; wavelet enhancement; attention mechanism; frequency-domain and spatial-domain collaboration

1. Introduction

The year-on-year rise in melanoma incidence makes early and accurate diagnosis crucial for improving patient survival rates [1]. Although dermoscopy provides high-resolution images of lesions, common image artefacts such as varying scales, blurred boundaries, and hair obstruction pose

significant challenges for manual interpretation [2,3]. Consequently, the development of efficient automatic lesion segmentation algorithms to aid clinical decision-making has become an urgent priority in the field of intelligent medical image analysis.

As a core component of computer-aided diagnosis, lesion segmentation provides a scientific basis for subsequent pathological analysis and treatment planning. Although deep learning has made significant progress in the field of medical imaging [4,5], methods remain inadequate when it comes to handling complex background noise and modeling long-range dependencies. In particular, current algorithms struggle to fully meet the standards required for high-precision clinical diagnosis, particularly in terms of boundary detail recovery and model generalization.

Traditional convolutional neural networks (CNNs) have played a significant role in medical image segmentation due to their exceptional ability to extract local features. Represented by the classic U-Net framework, this model achieves an initial fusion of semantic and fine-scale information through the combination of an encoder and a decoder with skip connections [6]. Subsequent researchers have optimized the architecture across dimensions; for instance, UNet++ reduces semantic discontinuities through nested dense skip connections [7], nnU-Net enhances generalization performance via an adaptive configuration strategy [8], U-Net v2 focuses on the redesign of feature propagation paths [9], and MALUNet explores the application potential of multi-attention mechanisms in lightweight models [10]. Concurrently, our team has also conducted a series of optimization studies. Among these, FM-Unet introduced a feedback regulation mechanism to optimize feature representation of the network, significantly improving the accuracy of biomedical image segmentation [11]; MCNMF-Unet constructed a hybrid structure combining convolutional layers and multi-layer perceptrons, enabling more efficient fusion and utilization of multi-scale features [12].

With the rise of Vision Transformer architectures, the advantages of self-attention mechanisms in modeling global contexts have become increasingly apparent. Representative methods such as TransUNet, Swin-Unet, and TransFuse have enhanced the models' ability to capture global dependencies through hybrid architectures, hierarchical window attention, or dual-branch parallel structures, respectively [13–15]. Relevant research conducted by our team also keeps pace with this technological trend. Among these, HCT-Unet, by constructing a CNN-Transformer hybrid encoding structure, effectively balances the modeling of local details and global dependencies, achieving good results in multi-objective medical image segmentation [16]. DPUSegDiff, through the fusion of a dual-path U-Net and a diffusion model, provides a new generative approach for complex medical image segmentation, and offers a reference for this paper in addressing noise interference and improving segmentation accuracy [17]. However, CNNs are constrained by their receptive field size, while Transformer models face limitations such as excessive computational complexity and insufficient perception of fine-grained local textures; this means they face challenges when handling medical scenarios that demand extremely high edge resolution [18].

Moreover, the Mamba architecture, based on selective state-space models, has opened new avenues for efficient visual modeling [19]. Compared to traditional architectures, Mamba is capable of modeling long-sequence dependencies with linear computational complexity, maintaining strong global awareness while reducing computational resource consumption. Although models such as VMamba, Mamba-UNet, and VM-UNet have demonstrated potential in the field of medical imaging [20–25], existing Mamba-based methods still exhibit significant shortcomings when applied to the context of dermatoscopic images. First, their ability to capture texture and high-frequency details at lesion edges is insufficient,

leading to over-segmentation or under-segmentation in low-contrast regions. Second, they lack adequate defense mechanisms against complex noise interference, such as hair occlusion, making it difficult to strike an ideal balance between segmentation accuracy, model lightweighing, and clinical deployment efficiency.

In response to the aforementioned challenges, we propose a skin lesion segmentation network that integrates Discrete Periodic Interpolation (DPI) wavelet enhancement with a spatial channel mutual feedback dual attention mechanism. Based on a U-shaped architecture, the core innovations of the model lie in the design of the Wavelet-enhanced Visual State Space Block (W-VSSB) and the Mutual Feedback Spatial-Channel Attention Block (MF-SCAB). During the encoding phase, we introduce a novel DPI wavelet transform, utilizing multi-scale frequency-domain decomposition to effectively separate low-frequency contours from high-frequency details. This significantly enhances the expressive power of the model for fine structures and lesion textures, while suppressing background noise from the frequency domain. Building upon this, the integration of the Mamba module enables long-range modeling, thereby strengthening the global interaction of multi-scale contextual features.

Furthermore, to address the lack of selectivity exhibited by traditional skip connections in feature fusion, we propose the MF-SCAB module. This module locates lesion regions using guidance information generated in the spatial dimension and establishes cross-dimensional feature dependencies through an adaptive calibration mechanism in the channel dimension, thereby achieving bidirectional dynamic interaction between spatial information and channel features. Through this mutual feedback weighting process, the model can focus on key lesion regions and attenuate irrelevant interference, thereby enhancing edge restoration accuracy in complex scenarios.

In this paper, we provide an in-depth exploration of the synergistic effects of wavelet frequency-domain enhancement, Mamba global sequence modeling, and a dual-attention bridging mechanism when processing low-contrast and high-noise images. The major academic contributions of this paper are as follows:

1) A wavelet-enhanced visual state space module is proposed. By deeply integrating DPI-based discrete interpolated wavelet transform with state space modeling to achieve collaborative representation of frequency-domain information and global contextual features. This design effectively alleviates the information loss caused by conventional downsampling operations and enhances the ability of the model to perceive and discriminate fine-grained lesion details.

2) A spatial-channel feedback dual-attention fusion module is designed. It is employed to enhance multi-scale feature propagation in the skip-connection stage. Through cross-dimensional dynamic feature calibration, the module enables deep complementary interaction between spatial and channel representations, thereby significantly improving feature fusion quality and boundary segmentation accuracy.

3) Extensive validation is conducted for public datasets such as ISIC2017, ISIC2018, and PH2. Experimental results demonstrate that our method outperforms current mainstream algorithms in terms of segmentation accuracy and boundary recovery. Further ablation studies verify the rationality of the proposed modules, and the network achieves outstanding segmentation performance with reliable clinical practicability.

2. Related work

2.1. Attention-based medical image segmentation

The application of attention mechanisms in medical image segmentation has primarily evolved from simple weight gating to complex cross-dimensional interactions. Early studies, such as that by Arora et al., entailed an Attention Gating mechanism to improve U-Net [26], suppressing background noise through adaptive weighting of shallow and deep features. Subsequent works, such as AS-Net [27] and MSCA-Net [28], further explored the synergistic effects of Spatial-Channel Attention, seeking to enhance lesion representation through dual recalibration of spatial position and feature channels.

However, attention-based models face technical challenges when processing images of skin lesions. On the one hand, while most methods, such as MALUNet [10] or GFANet [2], have introduced multi-level attention fusion, the interaction between spatial and channel dimensions is often achieved through simple parallel or serial superposition, lacking deep mutual-feedback interaction logic. This results in the model struggling to achieve precise boundary localization in low-contrast scenarios where lesions and normal tissue are highly similar. On the other hand, attention mechanisms often rely on extremely high computational costs to perceive global context; how to realize dynamic spatial-channel feature calibration under limited computational overhead remains a pressing issue. The MF-SCAB module designed in this paper aims to address this gap by constructing bidirectional mutual feedback paths to enable the coordinated evolution of spatial guidance and channel responses.

2.2. State space models and frequency-domain learning

The emergence of selective state-space models, namely the Mamba architecture, has provided a new approach to addressing the limited receptive fields of convolutional neural networks and the excessive computational complexity of Transformers. Since the introduction of the Visual State Space (VSS) module in VMamba [20], Mamba-based models such as VM-UNet [25] and ASP-VMUNet [29] have demonstrated outstanding performance in modeling long-range dependencies. In the task of dermatological lesion segmentation, works such as SkinMamba [30] and MambaU-Lite [31] have attempted to capture morphological changes in lesions through multi-scale state modeling. A set of U-Net improvement frameworks combining attention mechanism, wavelet transform and multimodal fusion for abdominal organ segmentation also offer important inspirations for the frequency-spatial collaborative optimization strategy proposed in this paper [32–35].

Although Mamba possesses a natural advantage in handling large-scale contextual information, its sequence-scanning nature renders the model relatively insensitive to local high-frequency details. Particularly in dermoscopy images, due to hair occlusion or extremely subtle lesion edge textures, purely spatial sequence modeling often results in the loss of critical high-frequency information. Although some researchers have attempted to remedy this by introducing dilated convolutions or edge-guided modules, such methods fail to address the issue of feature loss from the perspective of signal processing fundamentals. In contrast, the W-VSSB module proposed in this paper introduces DPI wavelet transforms to incorporate frequency-domain decomposition into the state-space modeling process. By leveraging the multi-resolution analysis capabilities of wavelet transforms, it enhances the accuracy of the model in capturing lesion edges and textural details. This combination of frequency-

domain enhancement and global modelling endows the model with greater robustness and semantic consistency in complex backgrounds.

3. Methodology

3.1. Architecture overview

The overall architecture of the proposed WMSM-Net is shown in Figure 1. The network adopts a symmetric encoder-decoder architecture and incorporates skip connections to preserve low-level fine-grained features, integrating the VMamba state-space model, MF-SCAB, and the DPI wavelet high-frequency feature enhancement module. Specifically, the encoder comprises two W-VSSB blocks and two VSS blocks, which progressively perform downsampling operations on the input dermatological lesion images to extract multi-scale hierarchical features; the MF-SCAB is responsible for refining the skip-connected features, while the W-VSSB uses DPI wavelets to enhance high-frequency details and suppress noise, thereby addressing the processing requirements of low-contrast images; and the decoder employs a symmetric upsampling layer to restore spatial resolution, fuses the encoder features enhanced by MF-SCAB, and outputs segmentation results matching the input dimensions via a convolutional prediction head. This achieves a balance between global modeling, local boundary preservation, and noise robustness, thereby meeting the requirements of precise dermatological image segmentation tasks.

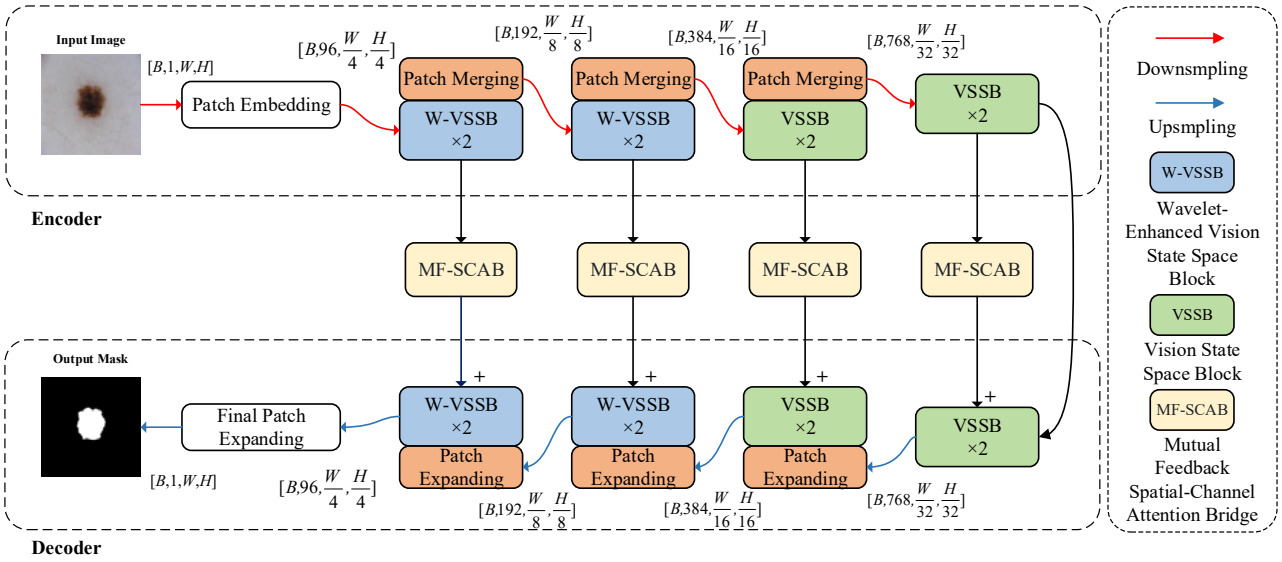


Figure 1. Overall Architecture of WMSM-Net.

To demonstrate the overall hierarchical composition of WMSM-Net, we summarize the entire network into three core stages: Encoder, skip connection bridge and decoder, as listed in Table 1. The table records the core components, primary functions and structural characteristics of each stage. The encoder adopts patch embedding, stacked W-VSSB and VSSB modules together with patch merging downsampling to extract multi-scale global context and high-frequency boundary information via DPI wavelet transform. The proposed MF-SCAB modules are deployed on all skip paths to realize

bidirectional spatial-channel mutual feedback feature optimization. The decoder adopts symmetric patch expanding upsampling layers and integrates W-VSSB and VSSB modules to further refine multi-scale fused features and recover subtle boundary textures lost in upsampling. Finally, a convolutional prediction head outputs binary segmentation masks matching the input image resolution.

Table 1. Three-stage hierarchical composition of WMSM-Net.

Stage	Core components	Main function	Structural property
Encoder	• Patch embedding • VSSB • W-VSSB • Patch merging • downsampling	Stepwise downsampling of input dermoscopic images, multi-scale global context modeling via selective state space, DPI wavelet decomposition extracts high-frequency lesion boundary details to alleviate edge loss	Partial layers equipped with W-VSSB for frequency-domain detail enhancement, standard VSSB for lightweight global sequence modeling, patch merging doubles channel dimensions while halving spatial resolution
Skip connection bridge	• MF-SCAB	Process multi-scale skip transmission features, realize bidirectional spatial-channel mutual feedback calibration; suppress hair and background texture interference, highlight lesion salient regions	Spatial branch generates guidance vectors to dynamically adjust channel responses, adaptive gated fusion balances original and attention-enhanced features, closed-loop cross-dimensional feature interaction
Decoder	• Patch expanding • VSSB • W-VSSB • Final_Patch expanding • upsampling	Gradually restore spatial resolution of high-level semantic features; refine fused multi-scale features; output binary lesion segmentation mask consistent with input size	Symmetric upsampling structure, both W-VSSB and VSSB are embedded inside decoder layers to compensate for lost texture and edge details during upsampling

3.2. Mutual feedback spatial-channel attention block (MF-SCAB)

In medical image segmentation tasks, the skip connection between the encoder and decoder serves as the core pathway for multi-scale feature propagation. However, traditional skip connections support only simple feature concatenation, suffering from shortcomings such as insufficient feature propagation, a disconnect between spatial and channel information, and interference from redundant features. These issues can easily lead to blurred lesion boundaries and loss of detail, making it difficult to meet the requirements for accurate segmentation of low-contrast, weak-boundary medical images such as dermatological lesions.

To address these issues, we propose the MF-SCAB module, with its architecture shown in Figure 2. Centered on a bidirectional mutual feedback mechanism, this module achieves precise fusion of skip-connection features through the synergistic enhancement of spatial and channel attention, thereby providing the decoder with more discriminative enhanced features. The MF-SCAB primarily consists of three components: The Spatial Attention Branch (SAB), the Channel Attention Branch

(CAB), and the mutual feedback fusion module. The two branches exchange information via spatial guidance vectors, and the enhanced features are fused through adaptive gating.

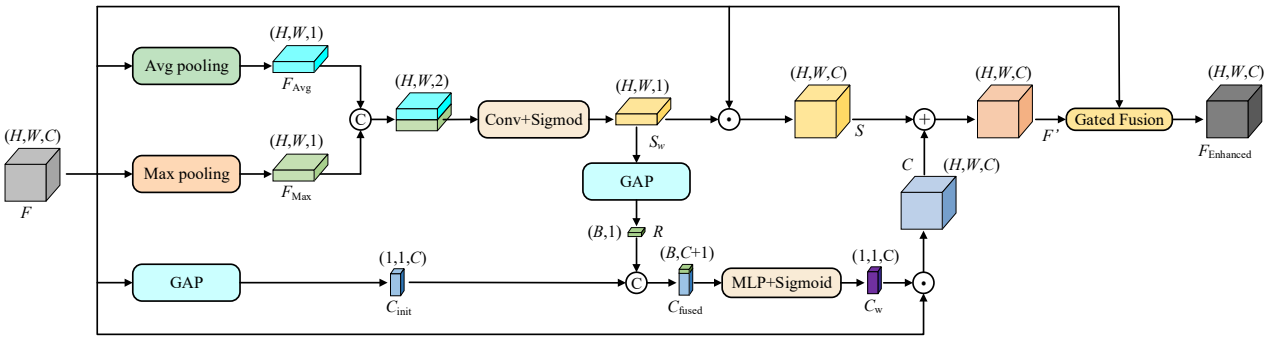


Figure 2. The architecture of MF-SCAB.

The role of the spatial attention branch is to model the distribution of importance across the spatial dimension, thereby highlighting lesion regions and suppressing background noise while generating spatial guidance vectors to provide global spatial prior information to the channel attention branch. Given the input skip-connected features F, in order to compress the channel dimension while preserving spatial structure, we first perform average pooling and max pooling operations on F along the channel dimension, as shown in Eqs (1) and (2), to obtain single-channel spatial statistical features.

$$F_{\text{Avg}} = \text{AvgPool}(F) \tag{1}$$

$$F_{\text{Max}} = \text{MaxPool}(F) \tag{2}$$

where F_{Avg} and F_{Max} denote the average and maximum responses of each channel at every spatial location, respectively, and AvgPool and MaxPool refer to channel-dimension pooling operations that extract global channel statistics and underpin the subsequent modeling and reconstruction of spatial weight features.

Subsequently, the two single-channel features are concatenated along the channel dimension, fed into a lightweight convolutional layer, and then passed through a Sigmoid activation function to generate the spatial weights S_{w}, as shown in Eq (3).

$$S_{\text{w}} = \sigma\left(\text{Conv}\left(\text{Concat}\left(\left[F_{\text{Avg}}, F_{\text{Max}}\right]\right)\right)\right) \tag{3}$$

where S_{w} denotes the spatial weight, σ represents the Sigmoid activation function, and Conv refers to the convolution operation that fuses the above pooled features to generate the final spatial weight.

To facilitate the transfer of spatial information to the attention channels, global average pooling is applied to the spatial weight map to extract a spatial guidance vector, as shown in Eq (4), which is used to represent global spatial salience.

$$R = \text{GAP}(S_{\text{w}}) \tag{4}$$

where R is the spatial guide vector and GAP stands for global average pooling.

The role of the channel attention branch is to model the distribution of importance across the channel dimension and to perform spatially guided channel adjustment via spatial guidance vectors.

First, global average pooling is applied to the input feature F, as shown in Eq (5), to extract global statistical information across the channel dimension.

$$C_{\text{init}} = \text{GAP}(F) \tag{5}$$

where C_{init} denotes the global feature of the initial channel, and each element corresponds to the global response of the input feature on the corresponding channel.

Subsequently, the spatial guidance vector R is fused with the initial channel features C_{init} to obtain the fused features C_{fused}, as shown in Eq (6), thereby enabling spatial priors to guide channel attention. The fused features C_{fused} are then fed into a multi-layer perceptron (MLP) with a Sigmoid activation function to generate the spatially guided and enhanced channel weights C_{w}, as shown in Eq (7). This operation injects spatial global importance into the channel dimension, enabling the channel attention to adaptively adjust weights based on spatial prior information. The MLP consists of two linear layers with ReLU activation functions, which are used to learn dependencies between channels.

$$C_{\text{fused}} = \text{Concat}(C_{\text{init}}, R) \tag{6}$$

$$C_{\text{w}} = \sigma\big(\text{MLP}(C_{\text{fused}})\big) \tag{7}$$

The role of the feedback fusion module is to achieve closed-loop feedback between spatial and channel information. The spatial weight S_{w} and channel weight C_{w} are each multiplied element-wise with the original feature F to obtain the spatially enhanced feature S and the channel-enhanced feature C, thereby enhancing the spatial feature response and channel feature representation of the lesion. Subsequently, the spatially enhanced feature S and the channel-enhanced feature C are fused through element-wise addition to obtain the fused feature F', thereby achieving complementarity between the enhanced information in the spatial and channel dimensions and completing the synergistic integration of the dual-attention features, as shown in Eq (8).

$$\begin{cases} S = F \odot S_{\text{w}} \\ C = F \odot C_{\text{w}} \\ F' = S + C \end{cases} \tag{8}$$

where $\odot$ denotes element-wise multiplication, bidirectional feedback enhancement is realized by highlighting lesion regions via spatial weights and emphasizing key channels through channel weights.

Subsequently, the enhanced feature F' and the original input feature F are fed into the gated fusion unit shown in Figure 3. The gating coefficient G is used to dynamically balance the original feature and the enhanced feature, thereby preventing the loss of detail caused by excessive feature enhancement, as shown in Eqs (9) and (10).

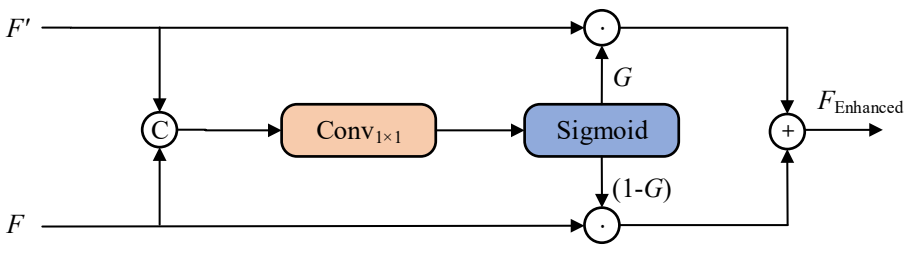


Figure 3. Gated fusion.

$$G = \sigma\left(\text{Conv}_{1\times 1}([F, F'])\right) \tag{9}$$

$$F_{\text{Enhanced}} = F \odot (1\text{-}G) + F' \odot G \tag{10}$$

In particular, the gate coefficient G is generated via a $\text{Conv}_{1\times1}$ and a sigmoid activation function and serves to dynamically adjust the relative contribution of the enhanced features. F_{Enhanced} represents the final output of MF-SCAB, which retains the global structure of the original features while enhancing the details and edge information of the lesion through a bidirectional feedback mechanism.

3.3. *Wavelet-enhanced VSS block (W-VSSB)*

3.3.1. VSS block

As shown in Figure 4, the VSS module [21] first performs layer normalization on the input features, then divides them into two parallel branches: The first branch performs feature mapping via linear transformations and activation functions; the second branch sequentially passes through a linear layer, a deep separable convolution layer and an activation function, after which spatial context information is extracted by the two-dimensional selective scan unit SS2D, the operation of which is illustrated in Figure 5. The processing workflow of SS2D comprises three steps: scan expansion, S6 block feature extraction, and scan merging. After completing sequential processing along the four directions: Top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right, the results are merged and the feature dimensions are restored. The specific implementation of this module can be found in [20]. Upon obtaining two output streams, the normalized result from the second stream is fused element-wise with the features from the first stream. This is then integrated via a linear layer with the introduction of a residual connection, ultimately outputting the features processed by the module. This network uses SiLU as the default activation function.

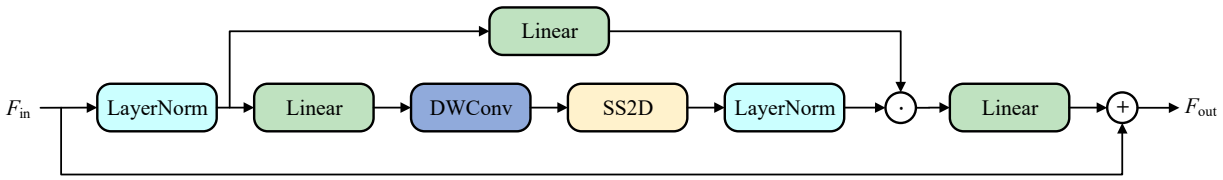


Figure 4. VSS block.

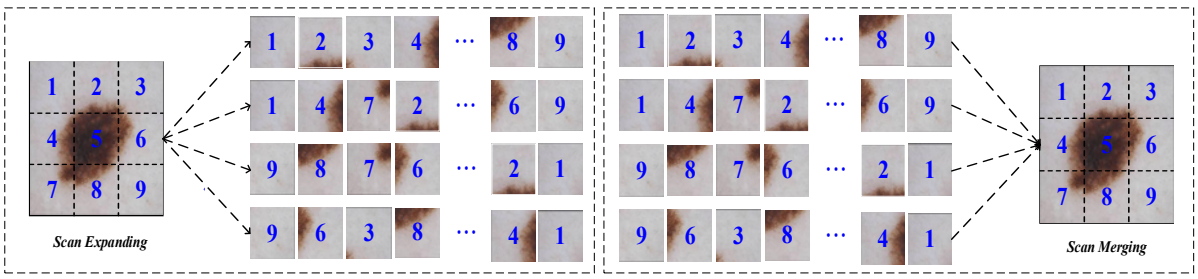


Figure 5. Image description for 2D-selective-scan.

3.3.2. Wavelet-enhanced VSS Block (W-VSSB)

The VSS Block has significant limitations in medical image segmentation tasks: Although its global modeling capability can capture large-scale structures, it is inadequate at modeling the fine-scale details commonly found in medical images, such as low-contrast lesion boundaries, minute anatomical structures, and imaging noise. This often results in blurred edges and loss of detail in segmentation results, making it difficult to meet the requirements for accurate segmentation of medical images with weak boundaries and high noise, such as those of skin lesions. To address these issues, we propose a W-VSSB, whose architecture is illustrated in Figure 6. Building upon the VSS architecture, this module introduces a DPI wavelet-enhanced branch and an adaptive gating fusion mechanism to achieve the synergistic optimization of global features and high-frequency details.

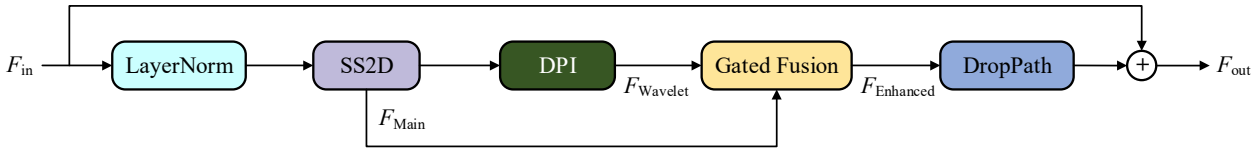
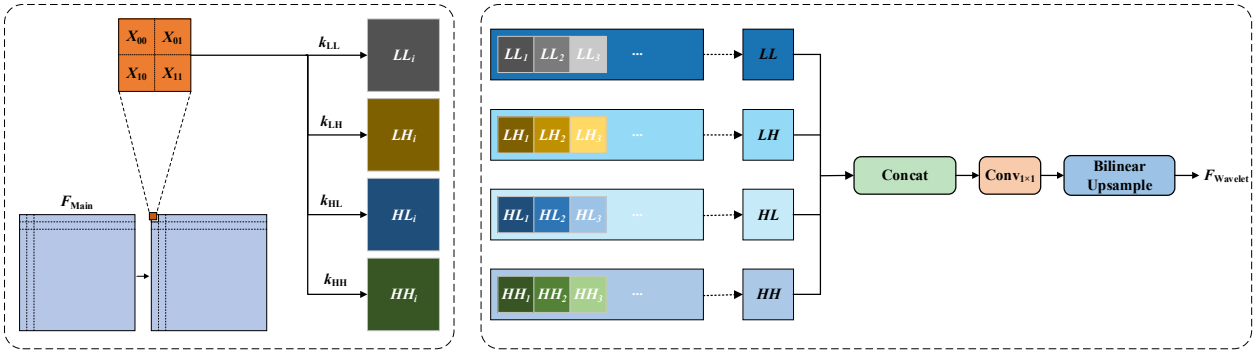


Figure 6. The architecture W-VSSB.



(a) (b)

Figure 7. DPI wavelet decomposition and enhancement. (a) DPI frequency-domain decomposition of a single feature unit, and (b) generation of wavelet-enhanced features.

In the W-VSSB, the input feature F_{in} is preprocessed via LayerNorm and fed into the SS2D module, which outputs the main feature F_{Main}; the main feature F_{Main} is then directly fed into the DPI wavelet transform module, as shown in Figure 7(a). The function of this module is to partition the input main feature F_{Main} into non-overlapping 2 × 2 local feature blocks with a step size of 2. For each 2 × 2 local block, four sets of fixed convolution kernels designed based on discrete smooth interpolation are used to perform weighted summation operations, yielding four corresponding wavelet coefficients: LL_i, LH_i, HL_i, and HH_i (where the subscript i denotes the i-th 2 × 2 block). The calculation method is shown in Eqs (11)–(15).

$$F_{Main} = \mathrm{SS2D}\big(\mathrm{LN}(F_{in})\big) \tag{11}$$

$$LL_i = \mathrm{Sum}\left(k_{LL}, Patch_i(F_{\mathrm{Main}})\right) \tag{12}$$

$$LH_i = \mathrm{Sum}\left(k_{LH}, Patch_i(F_{\mathrm{Main}})\right) \tag{13}$$

$$HL_i = \mathrm{Sum}\left(k_{HL}, Patch_i(F_{\mathrm{Main}})\right) \tag{14}$$

$$HH_i = \mathrm{Sum}\left(k_{HH}, Patch_i(F_{\mathrm{Main}})\right) \tag{15}$$

where $\mathrm{Sum}(\cdot,\cdot)$ represents the weighted sum of elements within a 2 × 2 local block, $Patch_i(\cdot)$ refers to the extraction of the i-th 2 × 2 local feature block from F, and k_{LL}, k_{LH}, k_{HL} and k_{HH} stand for four groups of fixed smoothing kernels corresponding to the low-frequency global component, horizontal high-frequency component, vertical high-frequency component, and diagonal high-frequency component, as defined in Eq (16).

$$\begin{cases} k_{LL} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} \\ k_{LH} = \begin{bmatrix} 0.75 & -0.75 \\ 0.25 & -0.25 \end{bmatrix} \\ k_{HL} = \begin{bmatrix} 0.75 & 0.25 \\ -0.75 & -0.25 \end{bmatrix} \\ k_{HH} = \begin{bmatrix} 1.0 & -1.0 \\ -1.0 & 1.0 \end{bmatrix} \end{cases} \tag{16}$$

Once the wavelet coefficients for all local blocks in Figure 7(b) have been computed, they are rearranged and stitched together according to their original spatial positions to form four complete frequency subband feature maps. Subsequently, the four frequency subbands are concatenated in the channel dimension and adaptively fused using learnable weights to achieve adaptive fusion of multi-subband information. Finally, bilinear upsampling is applied to restore the feature to the same spatial dimensions as the principal feature, yielding the wavelet-enhanced feature F_{Wavelet}, as shown in Eqs (17)–(21).

$$LL = \mathrm{Concat}\left(\{LL_i\}_{i=1}^{N}\right) \tag{17}$$

$$LH = \mathrm{Concat}\left(\{LH_i\}_{i=1}^{N}\right) \tag{18}$$

$$HL = \mathrm{Concat}\left(\{HL_i\}_{i=1}^{N}\right) \tag{19}$$

$$HH = \mathrm{Concat}\left(\{HH_i\}_{i=1}^{N}\right) \tag{20}$$

$$F_{\mathrm{Wavelet}} = \mathrm{Upsample}\left(\sigma\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}(LL, LH, HL, HH)\right)\right)\right) \tag{21}$$

where N indicates the total number of 2 × 2 blocks, Concat denotes the operation of concatenating

discrete wavelet coefficients into a complete feature map, Upsample represents bilinear upsampling, and $\text{Conv}_{1\times1}$ is adopted to learn the importance weights of different sub-bands for adaptive filtering of edges, textures and noise.

To avoid the problems of excessive amplification of high-frequency noise and dilution of useful details during the wavelet enhancement process, W-VSSB introduces an adaptive gated fusion unit, as shown in Figure 8, to achieve a dynamic balance between the main features and the wavelet features. After concatenating the main feature F_{Main} and the wavelet-enhanced feature F_{Wavelet} in the channel dimension, these are fed into a $\text{Conv}_{1\times1}$ layer and a Sigmoid function to generate the adaptive gating coefficient G, thereby achieving adaptive weighting of the wavelet features; finally, the modulated wavelet features are fused with the main feature using a residual approach to obtain the enhanced feature F_{Enhanced}, as shown in Eqs (22)–(24).

$$G = \sigma\left(\text{Conv}_{1\times1}\left(\text{Concat}([F_{\text{Main}}, F_{\text{Wavelet}}])\right)\right) \tag{22}$$

$$F'_{\text{Wavelet}} = G \odot F_{\text{Wavelet}} \tag{23}$$

$$F_{\text{Enhanced}} = F_{\text{main}} + F'_{\text{Wavelet}} \tag{24}$$

where G is the adaptive gate coefficient and F'_{Wavelet} is the modulated wavelet feature.

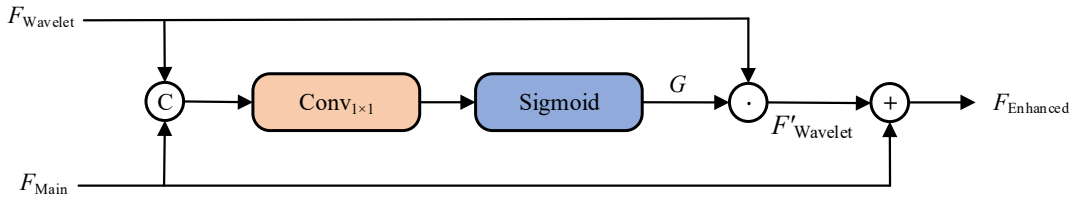


Figure 8. The gated fusion unit in W-VSSB.

After the enhanced feature F_{Enhanced} has been regularized via random path drop using DropPath, it is added to the residual of the original input feature F to obtain the output feature F_{out} of the W-VSSB, as shown in Eq (25).

$$F_{\text{out}} = F_{\text{in}} + \text{DropPath}(F_{\text{Enhanced}}) \tag{25}$$

In particular, DropPath(·) is used to mitigate model overfitting and stabilize the training process. Residual connections ensure effective gradient propagation, thereby preventing the vanishing gradient problem in deep neural networks.

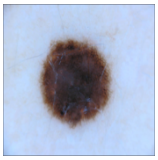
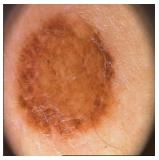
4. Experimental results and analysis

4.1. Datasets and experimental details

To comprehensively validate the segmentation accuracy and generalization performance of the proposed model, we selected three publicly available dermatological lesion datasets: ISIC2017 [36], ISIC2018 [37], and PH2 [38] for comparative experiments. Among these, the PH2 dataset comprises 400

high-quality medical images that have been meticulously annotated by specialist dermatologists, accompanied by manually generated segmentation masks, lesion classifications, and benign/malignant diagnostic results, thereby providing a stable and reliable benchmark for evaluation. The ISIC2017 and ISIC2018 datasets were publicly released by the International Skin Imaging Consortium, comprising 2150 and 2694 images, respectively. They cover various common skin lesion types, such as pigmented nevi and melanoma, and include real-world clinical noise such as uneven lighting and blurred lesion boundaries, thereby closely simulating actual diagnostic environments. The representative sample images and corresponding ground truth of each dataset are illustrated in Table 2. These three datasets complement one another in lesion morphological diversity and testing difficulty, which supports systematic evaluation of segmentation performance and robustness of models under complex scenarios.

Table 2. Representative sample images and corresponding ground truth from the three datasets.

Type dataset	ISIC2017	ISIC2018	PH2
Sample image			
ground truth			

All experiments in this study were conducted using Python 3.8 and the PyTorch 2.0.0 deep learning framework, with training and inference performed on an NVIDIA A10 graphics processing unit equipped with 32 GB of RAM. During data preprocessing, all input images were uniformly resized to a resolution of 256 × 256 pixels. Model training utilized the Adam optimizer, with an initial learning rate set to 0.001. The learning rate was dynamically adjusted using a cosine-annealing strategy, with a lower bound of 0.001. Regarding training parameters, the number of iterations was set to 300 for each dataset, with a batch size of 32. To guarantee a robust and unbiased performance evaluation, a 5-fold cross-validation strategy was implemented across all three datasets. In each fold, the dataset was initially partitioned into 5 independent subsets. One subset (20%) was reserved strictly as the unseen test set, while the remaining 4 subsets (80%) were further split into training and validation sets at a ratio of 7:1. Consequently, the final ratio for training, validation, and testing was maintained at 7:1:2, with all subsequent quantitative empirical results reported as (mean ± standard deviation) across the 5 folds. The BceDice hybrid loss function was selected as the loss function, as shown in Eqs (26)–(28).

$$L_{\text{Bce}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{26}$$

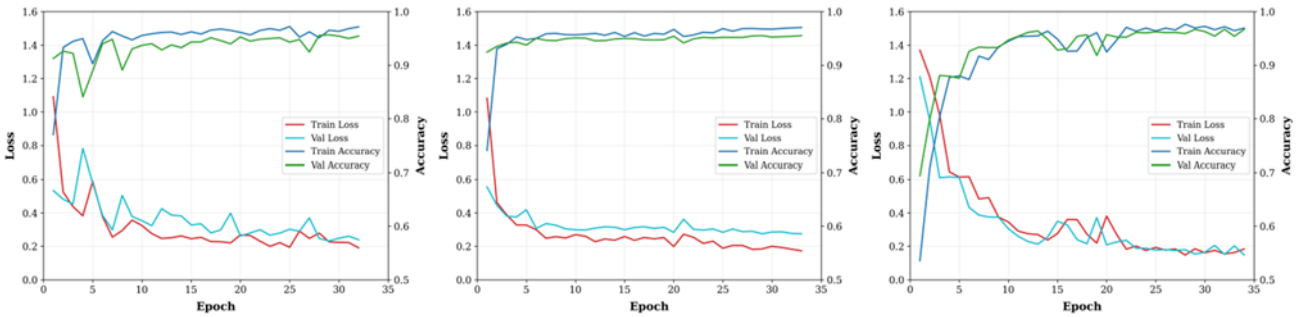
$$L_{\text{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{27}$$

$$L_{\text{BceDice}} = \lambda_1 L_{\text{Bce}} + \lambda_2 L_{\text{Dice}} \tag{28}$$

where N denotes the total number of samples, y_i and p_i represent the true label and the predicted value,

respectively, $|X|$ and $|Y|$ stand for the set of true labels and the set of predicted values, respectively, and λ_1 and λ_2 denote the weights of the two loss functions, of which are set to 1 by default in this paper.

Figure 9 plots the training loss, validation loss, training accuracy, and validation accuracy curves of WMSM-Net for ISIC2017, ISIC2018, and PH2 datasets to illustrate the model's convergence behavior during training. It can be seen that train loss and val loss continuously decline, while accuracy steadily rises for all three datasets. The gap between training and validation metrics remains small throughout the training process, which indicates that WMSM-Net has stable convergence and no obvious overfitting.



(a) (b) (c)

Figure 9. Training curves of WMSM-Net for three dermoscopic datasets. (a) ISIC2017 Datasets Training, (b) ISIC2018 Datasets Training, and (c) PH2 Datasets Training.

4.2. Evaluation metrics

We employ five evaluation metrics to quantify the segmentation performance of the model, including the mean Intersection-over-Union (*mIoU*), Dice Similarity Coefficient (*DSC*), Accuracy (*Acc*), Specificity (*SP*), and Sensitivity (*SE*). Among these, *mIoU* is used to calculate the mean of the ratio of the intersection and union of the predicted and ground truth (GT) regions; *DSC* is used to assess the similarity between the predicted segmentation results and the ground truth annotations; *Acc* represents the percentage of correctly classified results across all samples; *SP* measures the proportion of true negative samples that are correctly identified; and *SE* measures the proportion of true positive samples that are correctly detected. ASSD quantifies the boundary alignment error between predicted lesion contours and ground truth contours, where smaller values indicate higher consistency of segmentation boundaries. The calculation methods for these evaluation metrics are shown in Eqs (29)–(35).

$$IoU = \frac{TP}{TP + FP + FN} \tag{29}$$

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i \tag{30}$$

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{31}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{32}$$

$$SP = \frac{TN}{TN + FP} \tag{33}$$

$$SE = \frac{TP}{TP + FN} \tag{34}$$

$$ASSD = \frac{d(S_{pred} \rightarrow S_{gt}) + d(S_{gt} \rightarrow S_{pred})}{2} \tag{35}$$

where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative. S_{pred} denotes the surface voxel set of the predicted segmentation contour, and S_{gt} denotes the surface voxel set of the ground truth contour. $d(p,S)$ denotes the minimal Euclidean distance from point p to the surface S.

4.3. Comparison with state-of-the-art models

4.3.1. Performance evaluation for the ISIC 2017 dataset

Table 3 presents the quantitative evaluation results of various comparison methods for the ISIC2017 dataset. The experimental data show that the proposed WMSM-Net achieves the optimal results on core metrics, including mIoU, DSC, Acc, SE, and ASSD, reaching 83.51%, 91.01%, 96.59%, 89.67%, and 1.02, respectively. Compared with the Mamba-based benchmark model VM-UNet, WMSM-Net improves mIoU, DSC, Acc, and SE by 1.08%, 0.65%, 0.23%, and 0.83%, respectively, while reducing ASSD from 1.34 to 1.02. Such comprehensive performance gains mainly benefit from the designed W-VSSB module, which extracts high-frequency lesion details via wavelet frequency-domain enhancement to fix the weak feature capture defects of vanilla state-space models, delivering stronger feature discrimination for the noise-rich ISIC2017 dataset.

Table 3. Comparative experimental results for the ISIC2017 dataset.

Model	mIoU (%)	DSC (%)	Acc (%)	SP (%)	SE (%)	ASSD
UNet [6]	76.10 ± 2.96	86.40 ± 1.92	95.00 ± 0.71	97.98 ± 0.23	82.57 ± 2.93	2.70 ± 0.70
UNetV2 [9]	80.13 ± 0.89	88.97 ± 0.55	95.74 ± 0.30	97.29 ± 0.75	89.23 ± 2.48	1.72 ± 0.12
TransUNet [13]	80.27 ± 1.88	89.05 ± 1.16	95.83 ± 0.39	97.61 ± 0.18	88.33 ± 1.50	1.94 ± 0.38
ASP-VMUNet [29]	81.84 ± 1.66	90.01 ± 1.01	96.25 ± 0.52	98.34 ± 0.48	87.47 ± 1.62	1.33 ± 0.26
SkinMamba [30]	81.35 ± 1.14	89.71 ± 0.89	96.25 ± 0.20	97.97 ± 0.33	88.62 ± 1.41	1.56 ± 0.16
TransFuse [15]	79.88 ± 0.43	88.82 ± 0.26	95.82 ± 0.20	98.10 ± 0.55	86.24 ± 2.03	1.59 ± 0.17
MALUNet [10]	80.00 ± 1.38	88.88 ± 0.86	95.80 ± 0.45	97.90 ± 0.31	87.03 ± 1.66	1.52 ± 0.21
VM-UNet [25]	82.43 ± 1.74	90.36 ± 1.05	96.36 ± 0.32	98.15 ± 0.57	88.84 ± 1.94	1.34 ± 0.39
WMSM-Net	83.51 ± 1.09	91.01 ± 0.65	96.59 ± 0.35	98.22 ± 0.67	89.67 ± 1.40	1.02 ± 0.19

The visualization results shown in Figure 10 further confirm the superiority of the algorithm proposed in this paper. When dealing with complex lesion textures and blurred boundaries, the segmentation masks generated by WMSM-Net demonstrate significantly superior edge consistency

compared to other mainstream algorithms. A comparison reveals that WMSM-Net is capable of effectively suppressing artefact noise in the images and accurately capturing minute extensions of lesions, indicating that the MF-SCAB module has successfully enhanced key semantic information during the feature recalibration process.

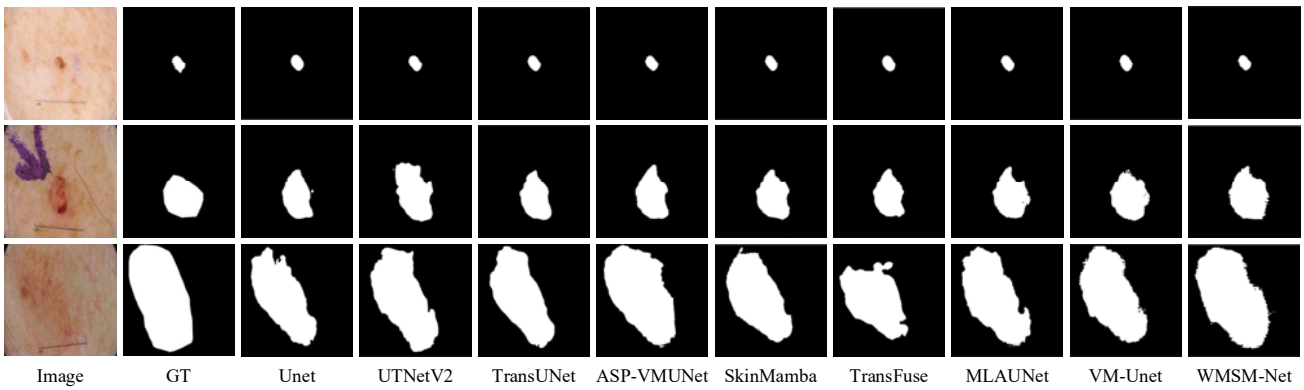


Figure 10. Visualization of model segmentation results for the ISIC2017 dataset.

4.3.2. Performance evaluation for the ISIC 2018 dataset

Table 4 presents the quantitative comparison results for the ISIC2018 dataset with richer lesion morphological diversity. WMSM-Net achieves optimal mIoU (83.61%), DSC (91.08%), Acc (96.19%), SP (98.10%), and minimal ASSD (1.22). Compared with the Mamba baseline VM-UNet, our model raises mIoU by 0.90%, DSC by 0.54%, Acc by 0.14%, and SP by 0.11%, while lowering the boundary error ASSD from 1.24 to 1.22. The elevated specificity demonstrates that the MF-SCAB spatial-channel feedback module effectively suppresses false segmentation on normal skin tissue and eliminates semantic confusion, enabling the model to deliver precise lesion segmentation on samples with complex morphologies.

Table 4. Comparative experimental results for the ISIC2018 dataset.

Model	mIoU (%)	DSC (%)	Acc (%)	SP (%)	SE (%)	ASSD
UNet [6]	76.94 ± 1.26	86.97 ± 0.80	94.58 ± 0.33	97.58 ± 0.49	83.71 ± 2.59	2.29 ± 0.13
UNetV2 [9]	80.66 ± 0.74	89.29 ± 0.45	95.44 ± 0.27	97.50 ± 0.63	87.98 ± 1.84	1.52 ± 0.14
TransUNet [13]	81.59 ± 0.60	89.86 ± 0.37	95.73 ± 0.46	97.32 ± 0.68	89.72 ± 1.00	1.40 ± 0.16
ASP-VMUNet [29]	82.33 ± 1.13	90.31 ± 0.68	95.84 ± 0.29	97.53 ± 0.49	**89.74 ± 1.13**	1.30 ± 0.16
SkinMamba [30]	81.08 ± 0.50	89.55 ± 0.51	95.62 ± 0.22	97.82 ± 0.26	87.56 ± 1.06	1.67 ± 0.11
TransFuse [15]	80.32 ± 0.94	89.09 ± 0.58	95.30 ± 0.32	97.13 ± 0.65	88.70 ± 1.48	1.58 ± 0.12
MALUNet [10]	79.40 ± 0.68	88.52 ± 0.42	95.13 ± 0.13	97.37 ± 0.48	87.00 ± 1.40	1.55 ± 0.17
VM-UNet [25]	82.71 ± 0.47	90.54 ± 0.28	96.05 ± 0.07	97.99 ± 0.28	88.90 ± 0.95	1.24 ± 0.17
WMSM-Net	**83.61 ± 0.45**	**91.08 ± 0.26**	**96.19 ± 0.13**	**98.10 ± 0.28**	89.30 ± 1.16	**1.22 ± 0.12**

The visual comparison results in Figure 11 demonstrate that WMSM-Net exhibits exceptional contour recovery capabilities when processing lesions with highly irregular shapes. Even in the presence of significant scale variations within the lesion region, our method can produce segmentation

results with high geometric integrity through the synergistic effects of multi-scale frequency-domain enhancement and long-range context modelling. This demonstrates the robust generalization capabilities of the model on complex large-scale datasets.

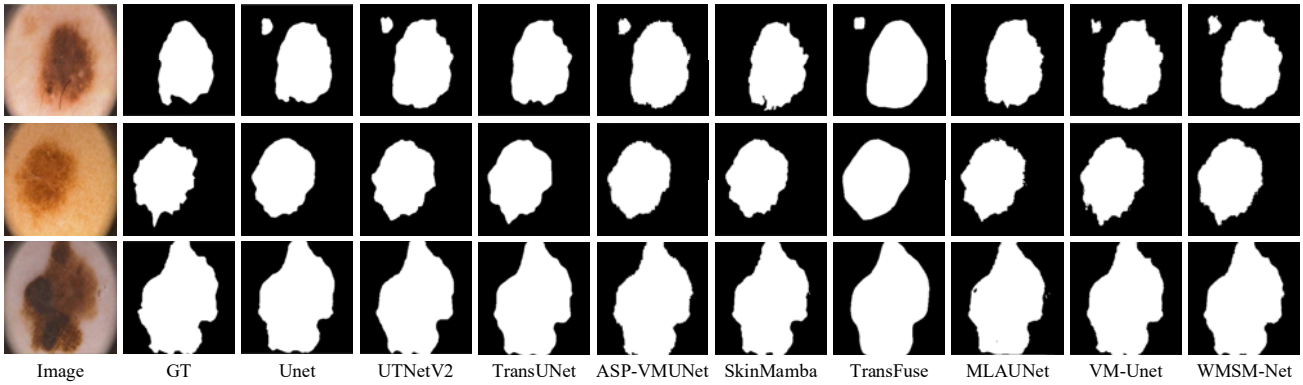


Figure 11. Visualization of model segmentation results for the ISIC2018 dataset.

4.3.3. Performance evaluation for the PH2 dataset

Table 5 lists the quantitative comparison results for the PH2 dermoscopic benchmark dataset. WMSM-Net outperforms all competitors and achieves the optimal mIoU (89.69%), DSC (94.56%), Acc (96.50%), SP (97.06%), and minimal ASSD of 0.54. Compared with VM-UNet, the Mamba baseline model, our method lifts mIoU by 0.88%, DSC by 0.49%, Acc by 0.36%, and SP by 0.53%, while reducing the boundary error ASSD from 0.64 to 0.54. The overall stable performance gains demonstrate that the combination of W-VSSB and MF-SCAB strengthens feature extraction and multi-scale calibration, delivering stronger feature representation and generalization robustness even on high-quality simple dermoscopic images.

Table 5. Comparative experimental results for the PH2 dataset.

Model	mIoU (%)	DSC (%)	Acc (%)	SP (%)	SE (%)	ASSD
UNet [6]	81.36 ± 3.06	89.70 ± 1.85	93.03 ± 1.77	94.70 ± 2.39	89.88 ± 2.96	1.65 ± 0.87
UNetV2 [9]	84.91 ± 3.15	91.81 ± 1.85	94.66 ± 1.39	95.13 ± 2.74	93.73 ± 1.50	1.10 ± 0.82
TransUNet [13]	87.56 ± 1.64	93.36 ± 0.93	95.56 ± 0.55	95.93 ± 1.28	95.27 ± 1.67	0.79 ± 0.45
ASP-VMUNet [29]	85.01 ± 3.30	91.88 ± 1.78	95.20 ± 0.49	96.45 ± 1.21	92.22 ± 1.66	1.06 ± 0.46
SkinMamba [30]	85.67 ± 3.28	92.24 ± 2.08	94.68 ± 1.15	94.92 ± 1.84	94.40 ± 1.84	1.02 ± 0.81
TransFuse [15]	84.71 ± 1.43	91.72 ± 0.84	94.67 ± 0.61	96.09 ± 1.48	91.82 ± 2.62	0.99 ± 0.54
MALUNet [10]	85.23 ± 1.46	92.02 ± 0.86	94.78 ± 0.25	95.59 ± 0.96	93.09 ± 3.03	1.15 ± 0.29
VM-UNet [25]	88.81 ± 0.01	94.07 ± 0.61	96.14 ± 0.38	96.53 ± 1.33	**95.46 ± 2.58**	0.64 ± 0.55
WMSM-Net	**89.69 ± 1.91**	**94.56 ± 1.06**	**96.50 ± 0.50**	**97.06 ± 0.80**	95.40 ± 2.15	**0.54 ± 0.39**

The visual analysis is shown in Figure 12. When handling low-contrast boundary scenarios specific to the PH2 dataset, WMSM-Net is able to accurately localize lesion contours, avoiding the over-segmentation commonly observed in traditional convolutional models. Particularly in complex samples with hair-covered lesions, output boundaries of the model closely match GT annotations,

owing to the advantages of the wavelet transform in extracting edge features. The experimental results consistently demonstrate that WMSM-Net achieves an optimal balance between segmentation accuracy and robustness, making it highly valuable for clinical applications.

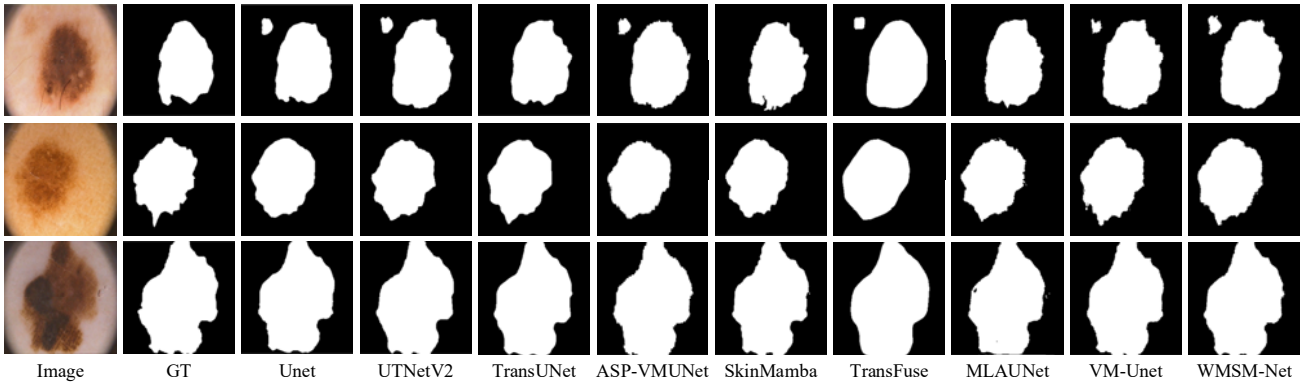


Figure 12. Visualization of model segmentation results for the PH2 dataset.

4.4. Ablation experiment

To validate the effectiveness of the two customized modules embedded in the proposed WMSM-Net and quantify the individual and joint contributions of each component to skin lesion segmentation, we carried out comprehensive ablation experiments for three public dermoscopy datasets (ISIC2017, ISIC2018, and PH2). We adopted the classic VMUNet as the baseline network and incrementally integrated the proposed W-VSSB and MF-SCAB modules to build four comparative model variants: vanilla VMUNet, VMUNet+W-VSSB, VMUNet+MF-SCAB, and the complete WMSM-Net with both modules. Tables 6 and 7 summarize the quantitative metrics, including mIoU, DSC, Accuracy, Specificity, Sensitivity, and ASSD across all test sets.

Table 6. Ablation experiments for the ISIC2017 and ISIC2018 datasets.

Dataset	Baseline	Modules		Metrics					
		W-VSSB	MF-SCAB	mIoU (%)	DSC (%)	Acc (%)	SP (%)	SE (%)	ASSD
ISIC2017	√			82.43 ± 1.74	90.36 ± 1.05	96.36 ± 0.32	98.15 ± 0.57	88.84 ± 1.94	1.34 ± 0.39
	√	√		83.38 ± 1.31	90.93 ± 0.78	96.63 ± 0.25	**98.62 ± 0.17**	88.23 ± 0.82	1.08 ± 0.19
	√		√	83.01 ± 1.58	90.70 ± 1.55	96.43 ± 0.45	98.05 ± 0.62	89.70 ± 1.46	1.30 ± 0.58
	√	√	√	**83.51 ± 1.09**	**91.01 ± 0.65**	**96.59 ± 0.35**	98.22 ± 0.67	**89.67 ± 1.40**	**1.02 ± 0.19**
ISIC2018	√			82.71 ± 0.47	90.54 ± 0.28	96.05 ± 0.07	97.99 ± 0.28	88.90 ± 0.95	1.24 ± 0.17
	√	√		83.55 ± 0.99	**91.09 ± 0.51**	96.02 ± 0.53	97.70 ± 0.54	**90.03 ± 1.41**	1.15 ± 0.04
	√		√	83.15 ± 0.79	90.80 ± 0.47	96.10 ± 0.30	98.04 ± 0.34	89.05 ± 0.51	1.19 ± 0.13
	√	√	√	**83.61 ± 0.45**	91.08 ± 0.26	**96.19 ± 0.13**	**98.10 ± 0.28**	89.30 ± 1.16	**1.22 ± 0.12**

From the unified experimental trend observed for ISIC2017, ISIC2018, and PH2, each module brings consistent performance gains when added separately. The designed W-VSSB introduces DPI wavelet decomposition to excavate high-frequency lesion textures and weak boundary information, which prominently reduces the ASSD boundary error of the baseline model for all datasets.

Nevertheless, we observe a slight transient drop in Sensitivity when only W-VSSB is activated, since frequency enhancement suppresses partial responses of tiny lesions while boosting background discrimination (higher Specificity). The standalone MF-SCAB mutual feedback attention bridge optimizes multi-scale skip feature transmission via closed-loop spatial-channel calibration, which notably lifts the lesion recall rate (Sensitivity) and alleviates background interference.

When combining W-VSSB and MF-SCAB simultaneously to form the full WMSM-Net, all datasets achieve the optimal comprehensive performance. Compared with the original VMUNet baseline, the complete model attains the maximum mIoU, DSC, and Accuracy, along with the minimal ASSD boundary distance error for ISIC2017, ISIC2018, and PH2. The complementary property of the two modules balances lesion recall and background identification perfectly: W-VSSB restores fine-grained edge details lost during downsampling, while MF-SCAB dynamically highlights lesion salient regions to offset the minor sensitivity decline caused by wavelet enhancement. The stable performance improvement across three datasets verifies that the collaborative design of W-VSSB and MF-SCAB effectively strengthens multi-scale lesion perception, weak boundary segmentation capacity and model robustness under diverse dermatoscopic imaging conditions.

Table 7. Ablation experiments for the PH2 datasets.

Dataset	Baseline	Modules		Metrics					
		W-VSSB	MF-SCAB	mIoU (%)	DSC (%)	Acc (%)	SP (%)	SE (%)	ASSD
PH2	√			88.81 ± 0.01	94.07 ± 0.61	96.14 ± 0.38	96.53 ± 1.33	**95.46 ± 2.58**	0.64 ± 0.55
	√	√		89.30 ± 1.92	94.19 ± 1.07	96.07 ± 0.46	96.65 ± 1.39	95.03 ± 1.56	0.54 ± 0.74
	√		√	89.16 ± 1.61	94.28 ± 0.90	96.05 ± 0.38	97.06 ± 1.46	94.31 ± 1.58	0.56 ± 0.55
	√	√	√	**89.69 ± 1.91**	**94.56 ± 1.06**	**96.50 ± 0.50**	**97.06 ± 0.80**	95.40 ± 2.15	**0.54 ± 0.39**

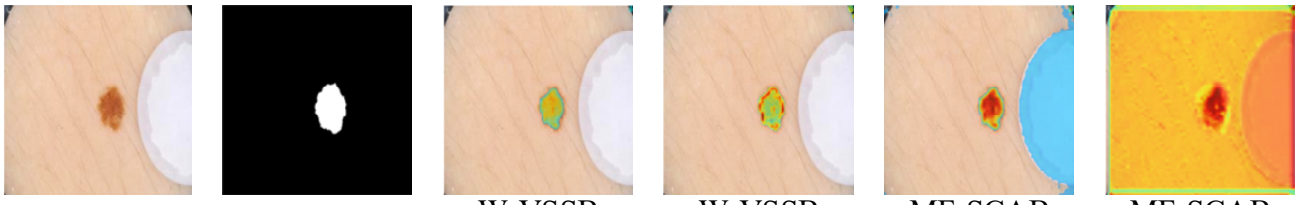


Figure 13. Visualization of feature maps.

Furthermore, to intuitively validate the feature enhancement and noise suppression performance of the proposed W-VSSB and MF-SCAB modules, we present intermediate channel-averaged feature heatmap visualizations in Figure 13, paired with the original dermoscopic image and lesion ground truth mask for direct comparison. The input feature map of W-VSSB shows a relatively compact lesion response with blurry edge contours. After wavelet-domain high-frequency enhancement, the W-VSSB module enriches the fine texture information on lesion boundaries and amplifies subtle pathological features inside the lesion, making the outline of the target lesion more distinct against the background. For the MF-SCAB module, its input feature heatmap contains obvious irrelevant activation on the right side of the image caused by the circular white artifact, which belongs to invalid background interference. Benefiting from spatial-channel mutual-feedback attention, MF-SCAB greatly

suppresses these out-of-lesion redundant feature activations, compresses the overall feature response range, and forces the model to concentrate feature weights strictly on the valid lesion area that matches the contour of the ground truth mask. These qualitative visualization results verify that the two collaboratively designed modules are capable of excavating fine-grained lesion boundary characteristics while eliminating complex irrelevant background distractions, enabling the network to learn more discriminative feature representations and facilitating precise skin lesion segmentation.

4.5. Model complexity and computational efficiency analysis

To comprehensively evaluate the practical deployment feasibility and architectural efficiency of the proposed WMSM-Net, we conducted a rigorous quantitative analysis of model complexity, as shown in Table 8. Specifically, the total number of network parameters (Params) and Floating Point Operations (FLOPs) were measured as the primary metrics for computational overhead.

Table 8. Model complexity.

Category	Model	Params (M)	FLOPs
CNN-based	UNet [6]	31.04	8.433
	MALUNet [10]	0.18	0.083
Transformer-based	TransUNet [13]	105.32	56.660
	TransFuse [15]	26.20	11.219
	UNetV2 [9]	25.02	5.399
Mamba-based	SkinMamba [30]	24.67	4.070
	ASP-VMUNet [29]	26.14	4.250
	VM-UNet [25]	27.43	4.112
Ours	WMSM-Net (MF-SCAB)	27.83	4.114
	WMSM-Net (W-VSSB)	34.88	4.379
	WMSM-Net (Full)	35.28	4.381

Table 8 presents a detailed comparison against alternative competitive state-of-the-art models as well as our internal sub-module variants. As indicated, while CNN-based light-weight architectures such as MALUNet sustain minimal computational footprints, their segmentation capacities under clinical noise are drastically restricted. Conversely, in comparison with heavy Transformer-based networks (e.g., TransUNet and TransFuse), our complete WMSM-Net secures higher segmentation accuracy across three benchmark datasets while maintaining a more compact parameter scale and reduced FLOPs. It is worth noting that compared to VM-UNet, our network exhibits a higher computational overhead in terms of both parameters and FLOPs. However, this marginal increment in complexity yields a substantial boost in boundary delineation and overall segmentation precision, thereby demonstrating a highly favorable performance-overhead trade-off for clinical scenarios where diagnostic accuracy is paramount.

Furthermore, the structural ablation detailed at the bottom of Table 8 explicitly clarifies that the step-by-step integration of the Wavelet-enhanced Visual State Space Block (W-VSSB) and the Mutual-Feedback Spatial-Channel Attention Bridge (MF-SCAB) introduces a strictly limited and acceptable computational increment over the baseline. This trade-off paradigm strongly verifies that the

substantial accuracy improvements yield directly from our collaborative frequency-spatial architectural innovation, rather than an unjustified brute-force scaling of model capacity.

5. Conclusions

In this paper, we introduce WMSM-Net, a visual state space network designed to remedy missing edge features and insufficient long-range dependencies in skin lesion segmentation. WMSM-Net innovatively integrates DPI wavelet transform into the Mamba architecture; its W-VSSB module enables multi-scale frequency decoupling and edge extraction, scaling up the model's sensitivity to weak boundaries and minute lesions. Furthermore, a spatial-channel mutual feedback block (MF-SCAB) is engineered to tackle scale variations and feature distortion by establishing bidirectional paths that isolate key pathological regions from background artifacts. Comparative experiments for ISIC 2017, ISIC 2018, and PH2 datasets verify that WMSM-Net yields superior segmentation accuracy and boundary reconstruction over mainstream baselines.

Despite the superior performance of WMSM-Net in capturing high-frequency edge details and long-range dependencies, several limitations warrant discussion. First, regarding computational efficiency, the dual-pathway architecture involving wavelet decomposition and bidirectional mutual-feedback mechanisms increases the computational overhead compared to purely lightweight Mamba-based models. Consequently, deploying WMSM-Net on resource-constrained embedded clinical hardware requires further optimization, such as model quantization or pruning. Second, the frequency-domain enhancement via W-VSSB relies on wavelet coefficients, which can occasionally be sensitive to severe artifacts (e.g., surgical markings, skin gels, or bubbles) that are absent in the clean training data. This sensitivity may lead to isolated false-positive segmentations in complex clinical environments, suggesting that further robustness training against such artifacts is necessary.

Based on these limitations, in future work, we will focus on three directions. First, we will lightweigh the dual-path architecture via quantization, pruning, and distillation for deployment on low-power clinical hardware. Second, artifact-specific data augmentation and an auxiliary suppression branch will be integrated to enhance robustness against clinical noise and eliminate false positives. Third, we will validate our framework on multi-device, diverse dermatological datasets to improve its clinical generalization.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province under Grant No. 2024J01821, the High-Level Cultivation Project of Minnan Normal University under Grant No. MSGJB2024009, and the Principal Foundation of Minnan Normal University under Grant No. KJ18010.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. C. Dayananda, N. Yamanakkanavar, T. Nguyen, B. Lee, AMCC-Net: An asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images, *Eng. Appl. Artif. Intell.*, **122** (2023), 106154. https://doi.org/10.1016/j.engappai.2023.106154
2. S. Qiu, C. Li, Y. Feng, S. Zuo, H. Liang, A. Xu, GFANet: Gated fusion attention network for skin lesion segmentation, *Comput. Biol. Med.*, **155** (2023), 106462. https://doi.org/10.1016/j.compbiomed.2022.106462
3. C. Dong, D. Dai, Y. Zhang, C. Zhang, Z. Li, S. Xu, Learning from dermoscopic images in association with clinical metadata for skin lesion segmentation and classification, *Comput. Biol. Med.*, **152** (2023), 106321. https://doi.org/10.1016/j.compbiomed.2022.106321
4. H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, Z. Wen, FAT-Net: Feature adaptive transformers for automated skin lesion segmentation, *Med. Image Anal.*, **76** (2022), 102327. https://doi.org/10.1016/j.media.2021.102327
5. M. Karri, C. S. R. Annavarapu, U. R. Acharya, Skin lesion segmentation using two-phase cross-domain transfer learning framework, *Comput. Methods Programs Biomed.*, **231** (2023), 107408. https://doi.org/10.1016/j.cmpb.2023.107408
6. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, Cham, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
7. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in *International Workshop on Deep Learning in Medical Image Analysis*, Springer International Publishing, Cham, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
8. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods*, **18** (2021), 203–211. https://doi.org/10.1038/s41592-020-01008-z
9. Y. Peng, D. Z. Chen, M. Sonka, U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation, in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, IEEE, (2025), 1–5. https://doi.org/10.1109/ISBI60581.2025.10980742
10. J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, Malunet: A multi-attention and light-weight unet for skin lesion segmentation, in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, (2022), 1150–1156. https://doi.org/10.1109/BIBM55620.2022.9995040
11. L. Yuan, J. Song, Y. Fan, FM-Unet: Biomedical image segmentation based on feedback mechanism Unet, *Math. Biosci. Eng.*, **20** (2023), 12039–12055. https://doi.org/10.3934/mbe.2023535
12. L. Yuan, J. Song, Y. Fan, MCNMF-Unet: A mixture Conv-MLP network with multi-scale features fusion Unet for medical image segmentation, *PeerJ Comput. Sci.*, **10** (2024), 1798. https://doi.org/10.7717/peerj-cs.1798
13. J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, et al., TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers, *Med. Image Anal.*, **97** (2024), 103280. https://doi.org/10.1016/j.media.2024.103280

14. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in *European Conference on Computer Vision*, Springer Nature Switzerland, Cham, (2022), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
15. Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, Cham, (2021), 14–24. https://doi.org/10.1007/978-3-030-87193-2_2
16. Y. Fan, J. Song, L. Yuan, Y. Jia, HCT-Unet: Multi-target medical image segmentation via a hybrid CNN-transformer Unet incorporating multi-axis gated multi-layer perceptron, *Visual Comput.*, **41** (2025), 3457–3472. https://doi.org/10.1007/s00371-024-03612-y
17. Y. Fan, J. Song, Y. Lu, X. Fu, X. Huang, L. Yuan, DPUSegDiff: A dual-path U-Net segmentation diffusion model for medical image segmentation, *Electron. Res. Arch.*, **33** (2025), 2947–2971. https://doi.org/10.3934/era.2025129
18. A. L. Sharma, K. Sharma, P. Ghosal, Skin lesion segmentation: A systematic review of computational techniques, tools, and future directions, *Comput. Biol. Med.*, **196** (2025), 110842. https://doi.org/10.1016/j.compbiomed.2025.110842
19. A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, preprint, arXiv:2312.00752.
20. Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, et al., Vmamba: Visual state space model, *Adv. Neural Inf. Process. Syst.*, **37** (2024), 103031–103063. https://doi.org/10.52202/079017-3273
21. Z. Wang, J. Q. Zheng, Y. Zhang, G. Cui, L. Li, Mamba-unet: Unet-like pure visual mamba for medical image segmentation, preprint, arXiv:2402.05079.
22. J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, preprint, arXiv:2401.04722.
23. H. Chen, Q. Liu, Z. Fu, L. Liu, UM-Mamba: An efficient U-network with medical visual state space for medical image segmentation, *Comput. Vision Image Understanding*, **259** (2025), 104436. https://doi.org/10.1016/j.cviu.2025.104436
24. C. Yu, Z. Fu, Z. Zhang, C. Chen, MMU-Net: An efficient medical image segmentation model combining multi-scale feature information, *Biomed. Signal Process. Control*, **112** (2026), 108265. https://doi.org/10.1016/j.bspc.2025.108265
25. J. Ruan, J. Li, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, *ACM Trans. Multimedia Comput. Commun. Appl.*, **2024** (2024). https://doi.org/10.1145/3767748
26. R. Arora, B. Raman, K. Nayyar, R. Awasthi, Automated skin lesion segmentation using attention-based deep convolutional neural network, *Biomed. Signal Process. Control*, **65** (2021), 102358. https://doi.org/10.1016/j.bspc.2020.102358
27. K. Hu, J. Lu, D. Lee, D. Xiong, Z. Chen, AS-Net: Attention synergy network for skin lesion segmentation, *Expert Syst. Appl.*, **201** (2022), 117112. https://doi.org/10.1016/j.eswa.2022.117112
28. Y. Sun, Y. Dai, Q. Zhang, Y. Wang, S. Xu, C. Lian, MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation, *Pattern Recogn.*, **139** (2023), 109524. https://doi.org/10.1016/j.patcog.2023.109524
29. M. Bao, S. Lyu, Z. Xu, Q. Zhao, C. Zeng, W. Bai, et al., ASP-VMUNet: Atrous shifted parallel vision mamba U-Net for skin lesion segmentation, preprint, arXiv:2503.19427.

30. S. Zou, M. Zhang, B. Fan, Z. Zhou, X. Zou, Skinmamba: A precision skin lesion segmentation architecture with cross-scale global state modeling and frequency boundary guidance, preprint, arXiv:2409.10890.
31. T. N. Q. Nguyen, Q. H. Ho, D. T. Nguyen, H. M. Q. Le, V. T. Pham, T. T. Tran, MambaU-lite: A lightweight model based on Mamba and integrated channel-spatial attention for skin lesion segmentation, in *The International Conference on Intelligent Systems & Networks*, Springer Nature Singapore, Singapore, (2025), 49–58. https://doi.org/10.1007/978-981-95-1746-6_6
32. H. Liu, Y. Fu, S. Zhang, J. Liu, Y. Wang, G. Wang, et al., GCHA-Net: Global context and hybrid attention network for automatic liver segmentation, *Comput. Biol. Med.*, **152** (2023), 106352. https://doi.org/10.1016/j.compbiomed.2022.106352
33. H. Liu, W. Sun, Y. Fu, S. Zhang, J. Jin, J. Fang, et al., Lifting wavelet transform-guided network with histogram attention for liver segmentation in CT scans, *Inf. Fusion*, **170** (2026), 104153. https://doi.org/10.1016/j.inffus.2026.104153
34. J. Fang, S. Qi, Y. Fu, S. Zhang, J. Jin, D. Wang, et al., Semi-supervised dual-teacher comparative learning with bidirectional balanced copy-paste for medical image segmentation, *Biomed. Signal Process. Control*, **119** (2026), 109985. https://doi.org/10.1016/j.bspc.2026.109985
35. H. Liu, J. Yang, Y. Fu, S. Zhang, D. Wang, J. Fang, et al., Adaptive modality-channel-search Hadamard multimodal fusion network for multi-phase liver tumor segmentation, *Biomed. Signal Process.*, **124** (2026), 110695. https://doi.org/10.1016/j.bspc.2026.110695
36. N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, (2018), 168–172. https://doi.org/10.1109/ISBI.2018.8363547
37. N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC), preprint, arXiv:1902.03368.
38. T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, J. Rozeira, PH^2-A dermoscopic image database for research and benchmarking, in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, (2013), 5437–5440. https://doi.org/10.1109/EMBC.2013.6610779

AIMS Press