BSLANet: a boundary and spatial localization-aware synergistic enhancement network for skin lesion segmentation

Chunbao Lu[1] , Guanxi Liu[1] , Hongyan Zhao[1], Weiman Xiao[1] and Tao Zhang[1(✉)]

[1] College of Software, Xinjiang University, Urumqi 830046, Xinjiang, China
xju_zhangtao@xju.edu.cn

Abstract. Accurate segmentation of skin lesions, a vital step in early skin cancer detection, is of utmost importance in improving patient survival rates. Although many deep learning-based methods have significantly progressed, challenges such as fuzzy lesion boundaries and complex spatial distribution remain. To resolve the above issues, this study proposes BSLANet, a novel boundary and spatial localization-aware synergistic enhancement network for skin lesion segmentation. To mitigate high-frequency information loss during downsampling, we introduce the Wavelet Attention Guided Downsampling Module (WAGDM). Furthermore, to enhance spatial understanding of complex lesions, we propose the Mixed Pooling Spatial Perception Attention (MPSPA). Lastly, to delineate finer-grained lesion boundaries and recalibrate lesion positions, we use the Differential Contrast Collaborative Feature Fusion Module (DCCFM) to improve semantic interaction in cross-layer feature fusion. Our experiments on four public skin lesion datasets—ISIC2016, ISIC2017, ISIC2018, and PH2—demonstrate that BSLANet surpasses current methods, achieving superior performance in skin lesion segmentation.

Keywords: Skin Lesion Segmentation, Downsampling, Feature Fusion.

1 Introduction

Manual evaluation for early skin lesion diagnosis is often subjective and inefficient[1]. To address this, medical image segmentation has evolved from Convolutional Neural Networks (CNNs) to hybrid models. Traditional CNNs suffer from limited receptive fields, making it difficult to model long-range dependencies. While Transformer-based hybrid models improve global modeling, they face high computational complexity. Recently, Mamba has shown significant advantages in long-sequence modeling[2].

However, existing Mamba-based models still face several challenges. First, downsampling often leads to the loss of high-frequency details[3]. Second, these models lack sufficient spatial awareness. Finally, semantic conflicts occur due to poor integration of deep and shallow features in the encoder-decoder structure. Therefore, the key to improving segmentation accuracy is to enhance high-

frequency feature capture, spatial perception, and cross-layer fusion while maintaining computational efficiency[4].

Based on the research background, this paper proposes a new boundary and spatial localization-aware synergistic enhancement algorithm for skin lesion segmentation, named BSLANet. The primary contributions of this work are summarized as follows:

- To reduce the loss of high-frequency details during downsampling, we propose the Wavelet Attention-Guided Downsampling Module (WAGDM). This module effectively enhances the representation of boundary features through a wavelet attention-based high-frequency compensation operation.
- To improve spatial feature capture and position awareness, we propose Mixed Pooling Spatial-Aware Attention (MPSPA) for the skip connection process. It generates spatial-aware feature maps via mixed pooling, thereby improving the model's ability to localize lesion coordinates.
- To address the semantic inconsistency caused by direct feature fusion across layers, we design a Difference Contrastive Synergistic Fusion Module (DCCFM). This module selectively fuses local and global differential information through a contrastive process guided by the current input features. This approach effectively reduces information conflicts and enhances semantic interaction, leading to more accurate boundary delineation and lesion re-localization.

2 Related Work

2.1 CNN and Transformer in Medical Image Segmentation

CNNs are the dominant method for medical image segmentation, with the U-Net[5]encoder-decoder architecture serving as the foundation. Variants such as Attention-Unet[6], ResUNet++[7], and DoubleU-Net[8] improve local feature modeling through attention mechanisms, residual connections, and multi-scale structures. Hybrid Transformer models (such as TransUNet[9] and TransFuse[10]) enhance global context modeling. Nevertheless, the computational complexity of self-attention grows quadratically with image resolution. This results in excessively high computational costs for high-resolution medical images[11-12].

2.2 State Space Models

State Space Models (SSMs) achieve global long-range modeling with linear complexity. Mamba has been successfully applied to vision tasks. Models such as VM-Unet[13] and U-Mamba[14] combine SSMs with CNNs, showing outstanding performance in medical image segmentation. However, existing Mamba-based models still have limitations. They exhibit weak high-frequency detail capture, insufficient spatial awareness, and inadequate cross-layer feature fusion. These issues are particularly evident in the fine-grained segmentation of

skin lesions. To address this, we propose BSLANet. This model enhances detail preservation, spatial perception, and cross-layer fusion through wavelet attention-guided downsampling, mixed pooling spatial-aware attention, and a difference contrastive synergistic fusion module.

3 Method

The BSLANet framework comprises an encoder, a decoder, and core modules. The structure is shown in **Fig. 1** The network input is $x \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{3}}$. First, the Patch Embedding layer divides the input into $r \times r$ patches. These patches are mapped to $C = 96$ channels to generate the embedded feature $x^{'}$.

The encoder (**Fig. 1**(a)) contains four stages. Each stage uses $[2,2,2,2]$ VSS Blocks to extract features. The outputs are denoted as e_i $(i = 1,2,3,4)$, with channel counts of $[C, 2C, 4C, 8C]$, respectively. A Wavelet Attention-Guided Downsampling Module (WAGDM) is integrated at the end of the first three encoder stages. This module reduces dimensionality while preserving high-frequency details, which mitigates the detail loss common in traditional downsampling.

The decoder (**Fig. 1**(b)) also has four stages. The channel counts decrease to $[8C, 4C, 2C, C]$. Each layer is configured with $[2,2,2,1]$ VSS Blocks. The outputs are denoted as v_i, and the final layer output is d_i.

Regarding feature interaction, each Mixed Pooling Spatial-Aware Attention (MPSPA) module takes the encoder feature e_i as input to model spatial localization. The Difference Contrastive Synergistic Fusion Module (DCCFM) handles cross-layer feature fusion. For $DCCFM_1$, the inputs are e_1, the output of $MPSPA_1$, and v_4. In all other stages, the module directly fuses e_i and v_i.

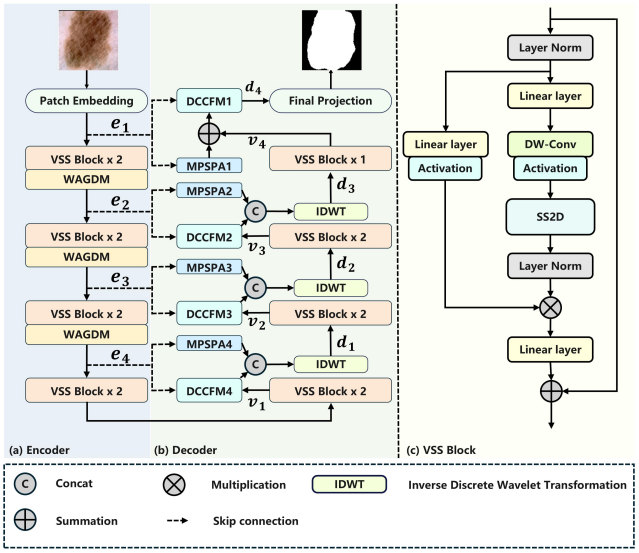

Fig. 1. The network architecture of BSLANet. BSLANet consists of (a) the encoder and (b) the decoder. (c) illustrates the VSS Block.

3.1 Wavelet Attention-Guided Downsampling Module (WAGDM)

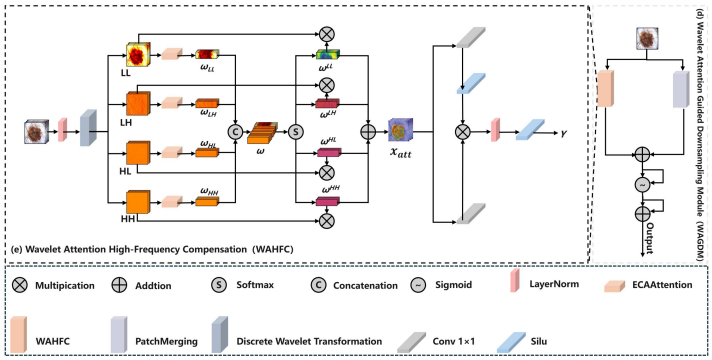

Fig. 2. (d) represents the architecture of WAGDM. (e) provides the details of the Wavelet Attention High-Frequency Compensation structure.

To address the issue of high-frequency detail loss in traditional downsampling, we propose the Wavelet Attention-Guided Downsampling Module (WAGDM). The architecture of the WAGDM is shown in **Fig. 2**. This module uses a dual-branch synergy to preserve lesion boundaries and texture information.

The right branch uses Patch Merging to perform coarse-grained downsampling. The left branch integrates the Wavelet Attention High-Frequency Compensation (WAHFC) process. In this branch, the input x is decomposed into low-frequency (LL) and high-frequency components (LH, HL, HH) via the

Discrete Wavelet Transform (DWT). We then introduce ECA channel attention to capture dependencies between low- and high-frequency channels. This generates weights $\omega_{LL}, \omega_{LH}, \omega_{HL}$,and ω_{HH}.

These weights are normalized using Softmax and used to perform weighted fusion of the multi-band features, resulting in x_{att}. Finally, a 1×1 convolution and SiLU activation are applied to enhance feature representation. The outputs of both branches are added element-wise to achieve downsampling that balances resolution reduction with high-frequency information preservation. The core fusion and output formulas are:

$$x_{\text{att}} = \text{Softmax}(\omega_{LL}) \cdot LL + \text{Softmax}(\omega_{LH}) \cdot LH + \text{Softmax}(\omega_{HL}) \cdot HL + \text{Softmax}(\omega_{HH}) \cdot HH \quad (1)$$

$$Y = \text{Branch}_{\text{right}} + \text{SiLU}\big(\text{Conv}_{1\times1}(x_{\text{att}})\big) \quad (2)$$

This module effectively compensates for high-frequency loss. It provides more expressive feature inputs for subsequent skin lesion segmentation.

3.2 Mixed Pooling Spatial-Aware Attention Module (MPSPA)

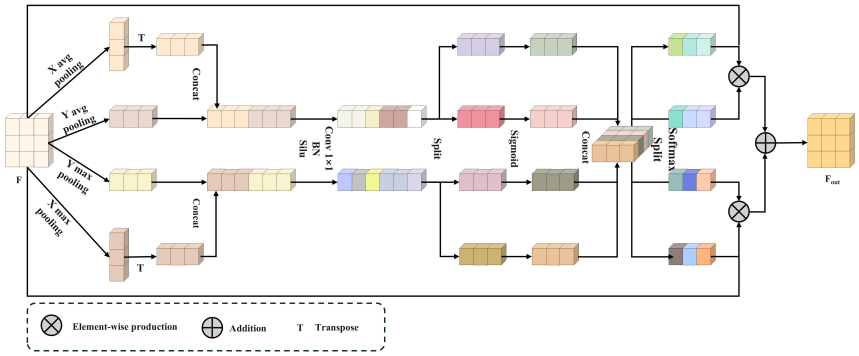

Fig. 3. Details of the proposed MPSPA.

We propose the Mixed Pooling Spatial-Aware Attention Module (MPSPA) to enhance spatial feature representation in skip connections .

Fig. 3 illustrates the MPSPA module. Given an input feature $F \in \mathbb{R}^{\mathbb{C}\times\mathbb{H}\times\mathbb{W}}$, the module first performs adaptive average and max pooling along both horizontal and vertical directions. This generates four direction-aware features: $F_H^{\text{avg}}, F_W^{\text{avg}}, F_H^{\text{max}}, and F_W^{\text{max}}$. After dimension transposition and spatial concatenation, these features pass through a shared 1×1 convolution (including BN and SiLU activation) to compress channels and enhance non-linear expression. The resulting features are then decoupled into directional components. Finally, Sigmoid and Softmax functions are applied to generate refined spatial attention weights. The core fusion formula is:

$$F_{\text{out}} = F \odot \text{Softmax}\left(\sigma\left(\text{Conv}_{1\times1}\left(\left[F_H^{\text{avg}}; F_W^{\text{avg}}; F_H^{\text{max}}; F_W^{\text{max}}\right]\right)\right)\right) \quad (3)$$

By combining multi-directional mixed pooling with an attention

mechanism, this module adaptively weighs feature contributions. This effectively strengthens the spatial discriminability of lesion regions and improves segmentation accuracy.

3.3 Difference Contrastive Synergistic Fusion Module

To address semantic conflicts and detail loss caused by the direct fusion of deep and shallow features in skin lesion segmentation, we propose the Difference Contrastive Synergistic Fusion Module (DCCFM). **Fig. 4** presents the overall architecture of the DCCFM module. This module strengthens cross-layer correlations through differentiated feature processing and a dynamic interaction mechanism.

Given shallow features X and deep features Y, the module first uses Dynamic Average/Max Pooling Selectors (DAPS/DMPS) to generate local details (la), global context (ga), local saliency (lm), and global saliency (gm). A Feature-Aware Module (FAM/CLS) captures differences through feature subtraction to modulate these into la^*, lm^*, and fine-grained global features ga^* and gm^*. Subsequently, 1×1 convolutions generate interaction features X_a and Y_b.

The module then performs element-wise subtraction and multiplication on X_a and Y_b. These results are concatenated and processed by a 1×1 convolution to obtain the contrastive correlation feature F. Finally, F is concatenated with X and Y. After Channel Shuffling (CS), a 7×7 convolution and a Sigmoid function generate the weight ω. The output F_{out} is obtained through weighted fusion followed by a final 1×1 convolution. The core fusion process is as follows:

$$F_{\text{out}} = f^{1\times1}\big(X \odot \omega + Y \odot (1-\omega)\big) \tag{4}$$

where the weight ω is calculated as:

$$\omega = \sigma(f^{7\times7}(CS(Concat(f^{1\times1}(Concat(X_a - Y_b, X_a \odot Y_b)), X, Y))) \tag{5}$$

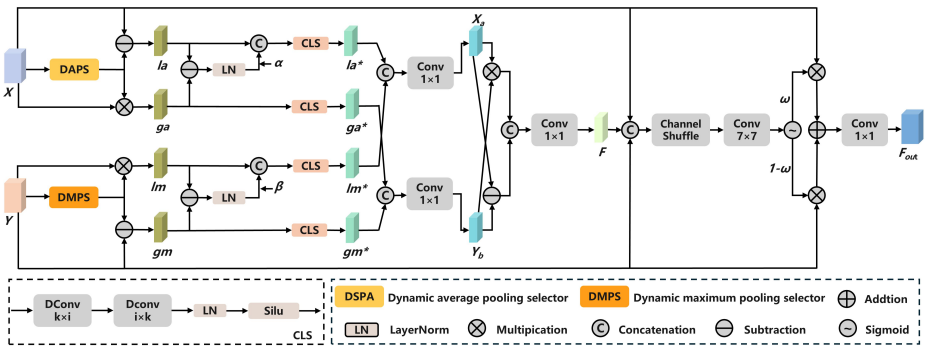

Fig. 4. Details of the proposed DCCFM.

4 Experiments

4.1 Datasets

The ISIC 2016, 2017, and 2018 datasets contain 1,279, 2,150, and 2,694 dermoscopy images with masks, respectively. Following a 7:3 split, the training sets consist of 900, 1,500, and 1,886 images, while the testing sets contain 379, 650, and 808 images. For the PH2 dataset, a typical 8:2 partition was applied to the 200 images, resulting in 160 training instances and 40 testing instances.

4.2 Evaluation Metrics

The five criteria utilized for segmentation assessment include Dice Similarity Coefficient (DSC), mean Intersection over Union (mIoU), Accuracy (Acc), Specificity (Spe), and Sensitivity (Sen). Their exact formulas are provided below:

$$\mathrm{DSC} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$\mathrm{mIoU} = \frac{TP}{TP + FP + FN} \tag{7}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\mathrm{Spe} = \frac{TN}{TN + FP} \tag{9}$$

$$\mathrm{Sen} = \frac{TP}{TP + FN} \tag{10}$$

4.3 Ablation Experiments

We conducted ablation experiments to verify the effectiveness of WAGDM, MPSPA, and DCCFM. Compared to the baseline, these three modules individually improved mIoU by 0.85%, 0.56%, and 1.13%, and Dice by 0.52%, 0.34%, and 0.69%, respectively. Combining WAGDM and DCCFM resulted in a further increase of 1.43% in mIoU and 0.87% in Dice. When all three modules were integrated, the model achieved an mIoU of 82.27% and a Dice score of 90.27%. These final results represent improvements of 1.78% and 1.08% over the baseline, respectively(see **Table. 2**).

Visual heatmaps(see **Fig. 6**) demonstrate that WAGDM enhances boundary responses, while MPSPA focuses on the lesion centers. DCCFM effectively suppresses background noise. The synergy of these three modules produces more precise responses that are highly consistent with the ground truth.

4.4 Results and Visual Analysis

Table. 1 Comparative results of diverse approaches on several datasets.

Dataset	Method	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
ISIC2016	Unet[5]	84.54	91.33	94.63	94.72	92.01
	UNet++[15]	84.39	91.27	94.40	94.66	92.03
	UTNetV2[16]	84.58	91.35	94.57	94.65	92.08
	nnU-Net[17]	85.26	92.57	94.52	95.76	93.00
	Attention-Unet[6]	84.78	91.51	94.78	95.13	92.20
	Swin-Unet[18]	84.94	91.89	94.90	95.32	92.88
	HC-Mamba[19]	86.95	92.87	95.01	95.87	93.23
	VMUnet[13]	86.80	92.81	94.90	95.83	93.21
	H-vmunet[20]	85.87	92.76	94.81	95.90	93.17
	VM-UNet-V2[21]	84.63	91.39	94.68	94.81	92.10
	SkinMamba[22]	87.28	93.01	95.40	96.36	93.25
	Ours	**88.66**	**93.99**	**95.73**	**96.95**	**93.52**
ISIC2017	Unet[5]	78.11	87.71	96.07	97.57	86.74
	UNet++[15]	77.50	88.59	96.26	97.71	88.45
	UTNetV2[16]	78.39	87.88	96.09	96.34	89.31
	nnU-Net[17]	79.43	88.53	96.16	97.82	88.82
	Attention-Unet[6]	78.45	87.75	96.27	97.60	86.68
	Swin-Unet[18]	79.97	88.86	96.21	97.30	88.72
	HC-Mamba[19]	79.49	88.57	96.19	97.83	88.03
	VMUnet[13]	79.75	88.73	96.30	97.95	86.93
	H-vmunet[20]	79.26	88.15	96.10	97.42	88.90
	VM-UNet-V2[21]	79.10	88.33	96.13	97.88	87.42
	SkinMamba[22]	80.11	88.82	96.33	97.97	88.17
	Ours	**81.59**	**89.86**	**96.62**	**98.09**	**89.32**
ISIC2018	Unet[5]	77.26	87.17	93.94	96.95	84.57
	UNet++[15]	78.12	87.72	94.02	96.06	87.69
	UTNetV2[16]	79.15	88.36	94.32	96.16	88.59
	nnU-Net[17]	80.25	89.04	94.73	96.53	87.94
	Attention-Unet[6]	79.80	88.76	94.56	96.6	88.10
	Swin-Unet[18]	79.91	88.83	94.45	95.68	90.04
	HC-Mamba[19]	80.33	89.09	94.64	96.17	87.87
	VMUnet[13]	80.49	89.19	94.70	96.57	89.61
	H-vmunet[20]	80.40	89.13	94.61	95.85	89.76
	VM-UNet-V2[21]	79.32	88.46	94.48	96.87	86.95
	SkinMamba[22]	81.05	89.53	94.84	96.20	90.19
	Ours	**82.27**	**90.27**	**95.27**	**96.90**	**90.61**
PH2	Unet[5]	84.67	88.54	93.38	90.57	90.01
	UNet++[15]	83.45	88.28	93.26	90.51	89.74
	UTNetV2[16]	85.29	88.75	93.50	90.65	90.15
	nnU-Net[17]	86.16	89.08	93.82	91.19	90.76
	Attention-Unet[6]	85.85	88.87	93.67	90.88	90.23
	Swin-Unet[18]	87.09	90.62	94.12	91.47	91.28
	HC-Mamba[19]	88.87	92.55	94.72	94.26	92.83
	VMUnet[13]	89.57	92.84	94.86	94.69	93.53
	H-vmunet[20]	88.23	91.29	94.58	93.95	92.34
	VM-UNet-V2[21]	87.68	91.19	94.37	92.68	91.79
	SkinMamba[22]	90.24	93.53	95.14	95.07	94.25
	Ours	**92.58**	**96.14**	**96.11**	**96.45**	**95.78**

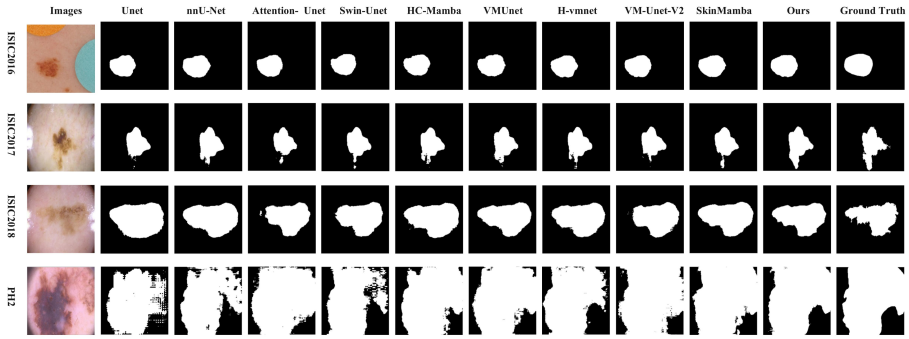

Fig. 5. Visual results across four datasets.

Table. 1 show that BSLANet outperforms mainstream models like UNet and SkinMamba across four datasets. These performance gains mainly stem from three core modules. WAGDM restores boundary details lost during downsampling through high-frequency compensation. MPSPA uses mixed pooling to enhance spatial awareness, achieving precise lesion localization. Finally, DCCFM resolves semantic conflicts in cross-layer fusion through differentiated processing. On the ISIC 2018 and PH2 datasets, the model achieves mIoU scores of 82.27% and 92.58%, respectively.

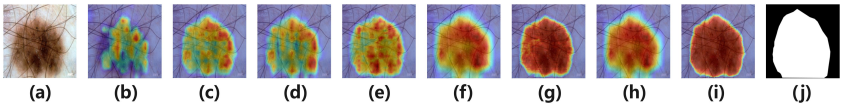

Fig. 6. Heatmap visualization of the ablation study for BSLANet. (a) Input image; (b) Baseline model; (c) + WAGDM; (d) + MPSPA; (e) + DCCFM; (f) + WAGDM + MPSPA; (g) + WAGDM + DCCFM; (h) + MPSPA + DCCFM; (i)+ WAGDM + MPSPA + DCCFM (Ours); (j) Ground Truth (GT).

BSLANet demonstrates high fidelity to the ground truth in the visual results (see **Fig. 5**). By using WAGDM to preserve boundary details, MPSPA to enhance spatial localization, and DCCFM to mitigate semantic conflicts, our model delivers superior performance.

Table. 2 Ablation results on the ISIC2018 dataset

Network	WAGDM	MPSPA	DCCFM	mIoU(%)	Dice(%)
Baseline				80.49	89.19
+WAGDM	√			81.34	89.71
+MPSPA		√		81.05	89.53
+DCCFM			√	81.62	89.88
+WAGDM+MPSPA	√	√		81.45	89.77
+WAGDM+DCCFM	√		√	81.92	90.06
+MPSPA+DCCFM		√	√	81.73	89.96
+WAGDM+MPSPA+DCCFM	√	√	√	**82.27**	**90.27**

5 Summary

This paper proposes BSLANet for skin lesion segmentation. Our approach reduces high-frequency detail loss during downsampling through WAGDM. It also enhances spatial localization awareness using MPSPA and mitigates semantic conflicts in cross-layer features via DCCFM. Experiments on several ISIC and PH2 datasets demonstrate that our method outperforms multiple State-of-the-Art and Mamba-based models. In the future, we will further optimize segmentation accuracy and extend this approach to other medical imaging tasks.

Acknowledgments. This work was supported by Finance Science and Technology Project of Xinjiang Uyghur Autonomous Region under Grant 2023B01029-1 and Xinjiang Key Laboratory of Intelligent Computing and Smart Applications, School of Software, Xinjiang University, Urumqi, China.

References

1. Roky A H, Islam M M, Ahasan A M F, et al. Overview of skin cancer types and prevalence rates across continents[J]. Cancer pathogenesis and therapy, 2025, 3(02): 89-100.
2. Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces[J]. arXiv preprint arXiv:2111.00396, 2021.
3. Xu G, Liao W, Zhang X, et al. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation[J]. Pattern recognition, 2023, 143: 109819.
4. Peng J, Lv K, Wang G, et al. MLSA-YOLO: a multi-level feature fusion and scale-adaptive framework for small object detection: J. Peng et al[J]. The Journal of Supercomputing, 2025, 81(4): 528.
5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234-241.
6. Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.
7. Kaur A, Singh Y, Chinagundi B. ResUNet++: a comprehensive improved UNet++ framework for volumetric semantic segmentation of brain tumor MR images[J]. Evolving Systems, 2024, 15(4): 1567-1585.
8. Jha D, Riegler M A, Johansen D, et al. Doubleu-net: A deep convolutional neural network for medical image.
9. Chen J, Mei J, Li X, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers[J]. Medical Image Analysis, 2024, 97: 103280.
10. Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2021: 14-24.
11. Liao W, Zhu Y, Wang X, et al. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation[J]. arXiv preprint arXiv:2403.05246, 2024.
12. Jiang Y, Li Z, Chen X, et al. Mlla-unet: Mamba-like linear attention in an efficient u-shape

model for medical image segmentation[J]. arXiv preprint arXiv:2410.23738, 2024.
13. Ruan J, Li J, Xiang S. Vm-unet: Vision mamba unet for medical image segmentation[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024.
14. Ma J, Li F, Wang B. U-mamba: Enhancing long-range dependency for biomedical image segmentation[J]. arXiv preprint arXiv:2401.04722, 2024.
15. Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//International workshop on deep learning in medical image analysis. Cham: Springer International Publishing, 2018: 3-11.
16. Gao Y, Zhou M, Liu D, et al. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark[J]. arXiv preprint arXiv:2203.00131, 2022.
17. Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.
18. Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205-218.
19. Xu J. Hc-mamba: Vision mamba with hybrid convolutional techniques for medical image segmentation[J]. arXiv preprint arXiv:2405.05007, 2024.
20. Wu R, Liu Y, Liang P, et al. H-vmunet: High-order vision mamba unet for medical image segmentation[J]. Neurocomputing, 2025, 624: 129447.
21. Zhang M, Yu Y, Jin S, et al. VM-UNET-V2: rethinking vision mamba UNet for medical image segmentation[C]//International symposium on bioinformatics research and applications. Singapore: Springer Nature Singapore, 2024: 335-346.
22. Zou S, Zhang M, Fan B, et al. Skinmamba: A precision skin lesion segmentation architecture with cross-scale global state modeling and frequency boundary guidance[J]. arXiv preprint arXiv:2409.10890, 2024.