An Improved Vision Mamba-Based Nested U-Net++ Architecture for Medical Image Segmentation

Rupak Chakraborty[1], Shashi Kant Gupta[2],
[1] Postdoctoral Researcher, LINCOLN UNIVERSITY COLLEGE, Malaysia; [2]Adjunct Professor, LINCOLN UNIVERSITY COLLEGE, Malaysia
Email ID pdf.rupak@lincoln.edu.my, rupak.jis@gmail.com

Abstract: Medical image segmentation plays an important role in computer-aided diagnosis by helping clinicians to identify anatomical structures and pathological regions accurately. Encoder-decoder approaches like UNet and UNet++ have achieved best performances due to their accurate extraction capabilities of feature and multi-scale fusion information. But these approaches often got stuck because of their lack of efficiency to understand global contextual information, inaccurate boundary localization and limited robustness when they are associated with complex medical images with varying contrast and noise levels. These advancements motivated to propose an improved Nested Vision Mamba UNet++ (VMUNet++) architecture that integrates the latest selective state-space (S6) modules where nested encoder-decoder techniques applied. The effectiveness of the proposed model has been tested on several publicly available benchmark datasets like ISIC2018, Synapse, and SegPC-2021. Experimental outcomes have been evaluated by some standard popular segmentation metrics like Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Boundary F1 Score (BF Score), Average Surface Distance (ASD), and Hausdorff Distance (HD95) and superiority of the proposed technique has been noted.
Keywords: SSM; VM-UNet++; Segmentation; BF Score; HD95.

Introduction

Image segmentation in medical field has made a strong impact in disease diagnosis, treatment planning, and clinical decision support. Automated segmentation of medical images helps medical professionals diagnose faster and more accurately. In recent years, transformer-based approaches achieved the greatest triumph in the related area. These approaches helped to enhance the segmentation accuracy as they are capable to capture both local and global image representations. The main drawback of Vision Transformer (ViT) was that learning local dependency information from the previous layers is not enough information. Several shortcomings of ViT-based models include demands for more data, deployment issues, fixed dimension constraints, poor locality-aware inductive bias, etc. Transformer-driven networks can model long-range global features, but they suffer from high computational costs and intensive resource requirements. To address these limitations, this work proposes an enhanced Vision Mamba-based nested encoder-decoder architecture that efficiently models long-range dependencies while preserving fine-grained spatial information through dense skip connections.

Furthermore, comprehensive experiments and ablation analyses are conducted to evaluate the effectiveness of the proposed framework.

Related work

This section will mainly survey the work done on medical image segmentation using Transformer-based and SSM-based models for accurate segmentation in recent years.

SSM and Transformer-based models for image segmentation

Transformer-based architectures effectively capture long-range dependencies but are computationally expensive and require large-scale training data, limiting their applicability in resource-constrained environments. To address these challenges, recent studies have shifted toward State Space Model (SSM)-based Vision Mamba (VM), which enables efficient bidirectional sequence modeling with lower computational complexity and improved scalability for high-resolution images [1]. Building on this concept, Ruan et al. [2] introduced VMUNet by integrating Visual State Space (VSS) blocks to enhance contextual feature extraction with fewer convolutional layers. Subsequently, Wu et al. [3] proposed H-VMUNet by extending the 2D Selective Scan (SS2D) mechanism for improved segmentation performance. Further advancements led to VMUNet++, which incorporates the nested skip connections of UNet++ to achieve more accurate feature fusion and decoder reconstruction [4]. Owing to its superior efficiency and segmentation capability, Vision Mamba-based architectures have emerged as a promising research direction, motivating the development of improved VMUNet++ models for medical image segmentation [5]. Additionally, Vision Mamba has recently demonstrated potential beyond computer vision, with emerging applications in speech recognition and speech enhancement.

Methodology

In previous section's literature review motivated to propose updated S6-based VM-UNet++ architecture for improved segmentation.

The main contribution of the proposed work has been summarized as follows:

1. Analyzed the Selective State Space Model (S6), which introduces selective information retention through learnable state-space parameters but remains limited to one-dimensional sequential data.
2. Then Selective Scan in Two Dimensions (SS2D) module has been incorporated to extend the applicability of S6 for image processing. SS2D has been considered as core computational component of the Vision Selective Scan (VSS) block.
3. Enhanced the VSS block by integrating an improved S6 model with SS2D, achieving better feature representation while maintaining the computational efficiency of the Mamba architecture.
4. The proposed VSS block has been incorporated into the VM-UNet++ to further enhance feature fusion and effectively preserve spatial information which result in reducing the

loss of detailed features. The proposed framework has been outlined in the below Figure 1.

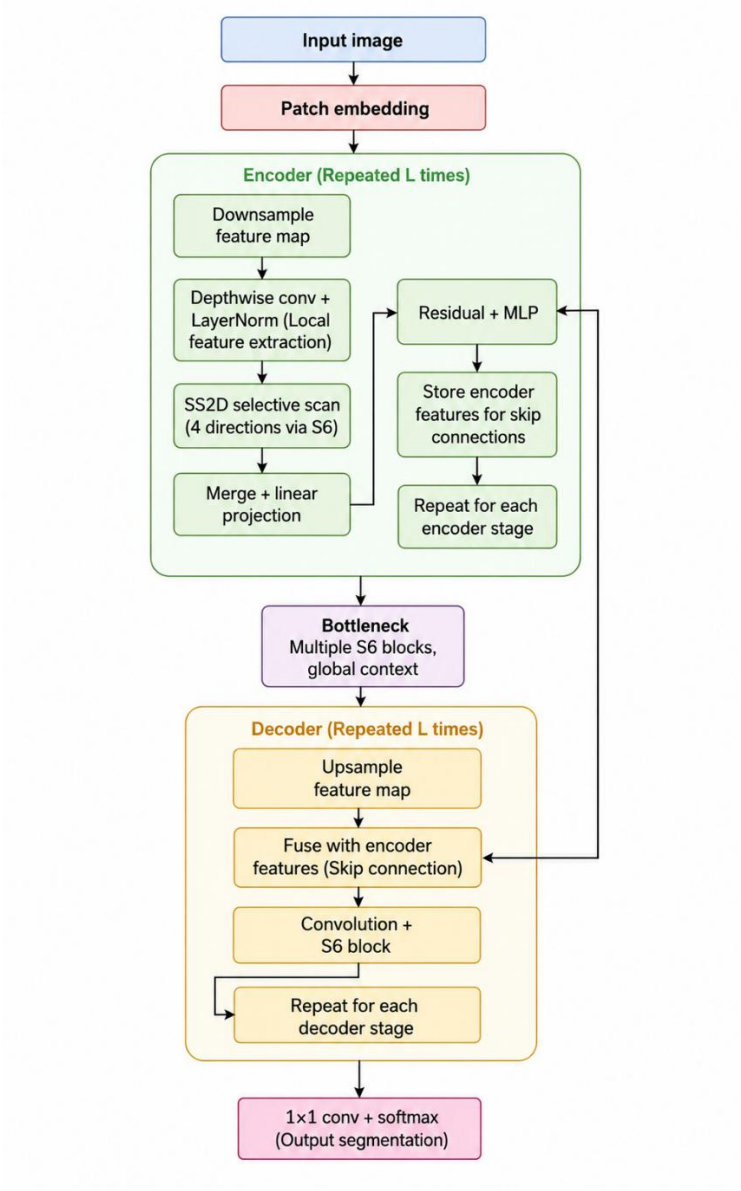

Figure 1 Work flow diagram of the proposed model

Data availability

- ISIC2018 dataset is available in **https://challenge.isic-archive.com/data/#2018.**
- Synapse dataset is available on **https://www.kaggle.com/datasets/dogcdt/synapse.**
- segPC-2021 dataset is available on **https://www.kaggle.com/datasets/sbilab/segpc2021dataset.**

Results Analysis

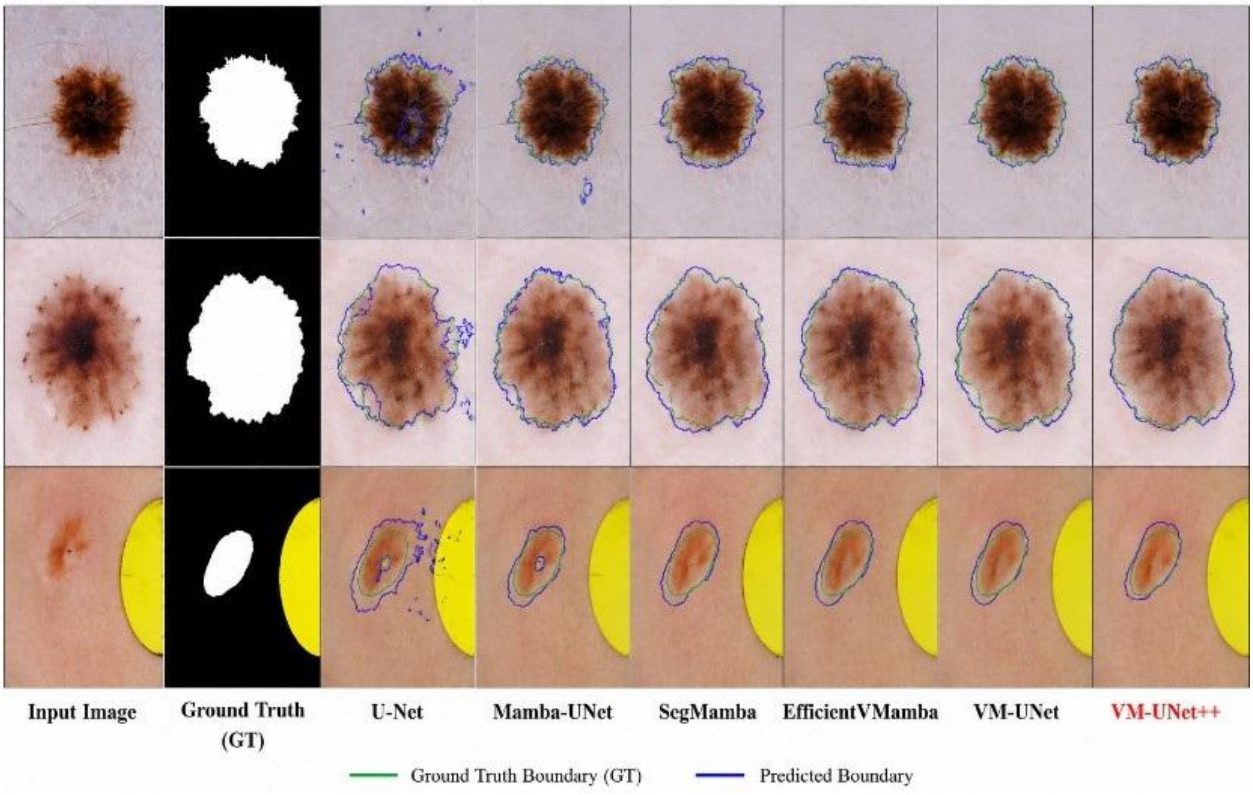

Figure 2 Comparative outcomes of some recent Mamba-based approaches on ISIC 2018

The outcome of the proposed VM-UNet++ along with some recent state-of-art Mamba-based approaches have been shown in the Figure 2. The boundary delineation parameters like Boundary F1 Score (BF Score), The Average Surface Distance (ASD), Boundary IoU and HD95 have been evaluated for the above outcomes and noted in the Table 1.

Table 1 Results of different Mamba-based approaches on ISIC2018 Dataset

Model	BF Score	ASD	BIoU	HD95
UNet	92.06	4.84	74.05	23.90
Mamba-UNet	94.86	3.91	77.19	22.48
SegMamba	95.72	3.69	78.98	21.22
EfficientVMamba	96.06	3.17	81.76	20.66
VMUNet	97.21	2.68	82.65	19.78
VMUNet++	97.69	2.21	83.29	18.69

Conclusion

This paper reviewed the evolution of medical image segmentation methods from UNet++ to Vision Mamba (VM)-based architectures. The proposed approach focuses the limitations of transformer-based models in terms of computational complexity and resource requirements. To address these challenges, the Vision Mamba-based VM-UNet++ architecture employing the 2D Selective Scan (SS2D) mechanism has been presented as an efficient alternative for accurate

segmentation. Experimental evaluation on the ISIC 2018 dataset demonstrated promising quantitative and qualitative results. Future work will focus on validating the proposed model on additional benchmark datasets, including Synapse, SegPC-2021, to further assess its robustness and generalization capability.

References

1. C. Qin, P. Tu, X. Chen, and J. Troccaz, "A novel registration-based algorithm for prostate segmentation via the combination of ssm and cnn", *Medical Physics,* vol. 49, pp. 5268-5282, 2022.
   https://doi.org/10.1002/mp.15698
2. J.Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation", *ACM Transactions on Multimedia Computing, Communications, and Applications,* 2025.
   https://doi.org/10.1145/3767748
3. R. Wu, Y. Liu, P. Liang, and Q. Chang, "H-vmunet: High-order vision mamba unet for medical image segmentation", *Neurocomputing,* vol. 624, pp.129447, 2025.
   https://doi.org/10.1016/j.neucom.2025.129447
4. W. Tang, Z. Wu, W. Wang, Y. Pan, and W. Gan, "Vm-unet++ research on crack image segmentation based on improved vm-unet", *Scientific Reports,* vol. 15, pp. 8938, 2025
   https://doi.org/10.1038/s41598-025-92994-7
5. Z. Sui, H. Liu, and J. He, "Vm-unet++: Vision mamba with adaptive feature reweighting and multiscale fusion for medical image segmentation", *In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), (IEEE, 2024)*, pp. 816-820, 2024.
   10.1109/ICICML63543.2024.10957962
6. G. Cao et al., "Long-range Sequential Modeling Mamba For 3D Abdominal Multi-Organ Segmentation," *2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan*, pp. 431-436, 2024
   https://doi.org/10.1109/ICIIBMS62405.2024.10792780
7. X. Pei, T. Huang, and C. Xu, "EfficientVMamba: Atrous selective scan for lightweight visual Mamba," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 6443–6451, 2025.
   https://doi.org/10.1609/aaai.v39i6.32690