Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 2, Pages: 18-25, Year:2025

Article

Multi-Scale Dual-Stream CNN–Mamba Hybrid Framework for Accurate Nuclei Segmentation in H&E Images

Nimra Bukhari[1], Shabir Hussain[1,5*], Hafiz Muhammad Anwar Shahzada[2], Rafique Haider[2], Hassan Munir[1], Imran Ali Mudassar[1], Yang Yu[3], Akmal Khan[4]

[1] Department of Computer Science, National College of Business Administration and Economics, Rahim Yar khan, 64200, Pakistan
[2] Department of Computer Science, Khawaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, Punjab, Pakistan
[3] School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China
[4] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[5] Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

* Correspondence: Shabir Hussain (e-mail shabir.nicaas@gmail.com; shabir.hussain@sz.tsinghua.edu.cn)

Submitted: 12-08-2025**, Revised:** 13-11-2025, **Accepted:** 28-11-2025

Abstract

Accurate nuclei segmentation in H&E-stained histopathology images is critical for quantitative tissue analysis and downstream computational pathology tasks. We propose MS-DS Mamba-Seg, a multi-scale dual-stream framework that integrates CNN-based local feature extraction with Mamba-based global context modeling to enhance nuclei segmentation. The Local Stream captures fine-grained morphological details such as nuclear boundaries and texture, while the Global Stream models long-range spatial dependencies across tissue regions. A novel Fusion Gate adaptively combines these complementary features, producing high-resolution and morphologically precise segmentation masks. Evaluated on the NuInsSeg dataset, MS-DS Mamba-Seg achieves a Dice score of 95.41%, outperforming baseline methods including Attention U-Net, UNETR, and SegMamba, while maintaining structural consistency and boundary accuracy. These results demonstrate the framework's effectiveness in leveraging complementary local and global features for robust nuclei segmentation.

Keywords: Nuclei Segmentation, Histopathology, Multi-scale Dual-Stream, CNN-Mamba, Vision Mamba, NuInsSeg Dataset,

1. Introduction

Histopathology image analysis plays a crucial role in tissue-specific molecular processes [1], modern cancer diagnostics [2], tumor grading, tissue characterization, and precision therapy guidance. Central to these tasks is the accurate segmentation of cell nuclei, which enables downstream analysis such as nuclear morphology assessment, spatial pattern mining, and cellular phenotyping. However, nuclei segmentation [3] in H&E images remains challenging due to the presence of overlapping nuclei, heterogeneous staining, variable morphology, and complex tissue architecture. Analysis of H&E images has advanced significantly, providing state-of-the-art

 https://doi.org/10.57041/eb60k036

solutions for segmentation [4]. Through the application of deep learning methods research has progressed from standard fully convolutional networks to more sophisticated and carefully crafted architectures aimed at tasks such as classification [6], segmentation and detection [7]. Artificial intelligence [8] multiprocessing approach based on transformer for nuclei segmentation. Traditional deep learning [9] techniques struggle to generalize across such variability, motivating the general implementation of deep learning-based segmentation approaches [10,11]. CNN architectures such as U-Net and UNet++ achieve strong boundary sensitivity but lack global awareness due to their limited receptive field. Vision Transformer (ViT)-based models capture long-range relationships but exhibit quadratic complexity, high memory consumption, and instability on high-resolution pathology inputs. While YOLO-based instance segmentation models offer lightning-fast inference, they often fall short when it comes to producing precise pixel-level masks for overlapping nuclei. In comparison, MASKYLO Yolo-Mask RCNN based model [12] stood out by delivering high-accuracy H&E histology segmentation, showcasing the power of hybrid deep learning approaches in capturing fine-grained, context-aware features. In this work, we introduce MS-DS Mamba-Seg, a dual-stream hybrid architecture that merges CNN and Mamba encoders. The model features a Local Stream for fine texture and edge representation, a Global Stream for sequence-aware structural understanding, and a Fusion Gate for adaptive integration. Our contributions are as follows:

Model's contribution

- A novel dual-stream architecture combining convolutional locality and Mamba-based global modeling.
- A Fusion Gate module that adaptively merges spatial-channel features from both streams A Fusion Gate module that adaptively merges spatial-channel features from both streams
- A lightweight multi-scale decoder designed for high-resolution segmentation.
- Extensive evaluations demonstrating state-of-the-art performance on NuInsSeg with superior efficiency.

2. Related Work

2.1. Local Feature Learning with Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have long been the cornerstone of medical image segmentation, particularly in 3D volumes [13]. Their strength lies in their ability to capture fine-grained local features. SegResNet [14] introduces a variational autoencoder (VAE) as a reconstruction branch, improving feature extraction and enhancing robustness. Despite their effectiveness, CNN-based architectures [15] have notable limitations. Their inherent design focuses on local receptive fields, making it challenging to capture long-range contextual dependencies.

2.2. Global Context Modeling with Transformers

Transformers [16] initially developed for NLP, have been adapted for 3D medical image segmentation due to their ability to model long-range dependencies via self-attention. By capturing global interactions among voxels, they often outperform CNNs on volumetric data.

2.3. Efficient Long- Range Modeling with Mamba and Hybrid Network

AttmNet [17] for example, integrates multi-scale convolution, self-attention, and Mamba layers within a U-shaped architecture to enhance lesion segmentation. SegEO Mamba [18] addressed this by introducing PE-Upsampling for integrating encoder features during upsampling and BMSC Attention to enhance skip-connection fusion. This decoder-centric enhancement significantly improved multi-organ and brain tumor segmentation, achieving a 90.97% Dice score on BraTS2023 and demonstrating robust generalizability.

3. Methodology

3.1. Proposed Framework

We introduce MS-DS Mamba-Seg, a multi-scale dual-stream hybrid architecture designed to enhance nuclei segmentation in H&E-stained histopathology images. The framework integrates the strengths of convolutional processing and long-range Selective State Space Modeling (Mamba) within a unified design. Unlike conventional U-Net derivatives or transformer-based models that rely on a single feature extraction pathway, MS-DS Mamba-Seg employs two complementary parallel encoders. The Local Stream, implemented using a CNN encoder, focuses on fine-grained morphological cues, enabling precise extraction of cellular boundaries and texture characteristics that are critical in H&E tissue interpretation. In contrast, the Global Stream, based on a Vision Mamba encoder, captures extended spatial dependencies and broader tissue architecture that surpass the intrinsic receptive field limitations of convolutions. Features from both streams are fused via a Fusion Gate and refined by the decoder to produce high-resolution segmentation masks.

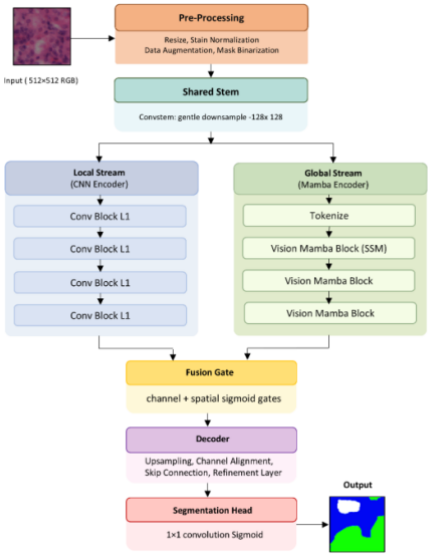


Figure A: Overview of the proposed *MS-DS Mamba-Seg Framework.* A dual CNN–Mamba stream extracts local boundaries and global tissue context for precise nuclei segmentation masks.

3.2. Dataset and Pre Processing

Experiments were conducted on the **NuInsSeg** dataset [19], containing H&E histopathology patches with pixel-level masks. Images were resized to 512×512×3, and **Macenko's stain normalization** was applied. Data augmentation included flips, rotations, color jitter, Gaussian blur, and sharpening, while masks were binarized to distinguish nuclei from the background.

3.3. Architecture Details

A) Shared Convolutional Stem

The stem module performs initial low-level feature extraction while applying a gentle spatial downsampling to preserve fine tissue morphology. It consists of a 3×3 convolution followed by a GELU activation and a 2× reduction in spatial resolution. This operation produces a shared feature map of size $128 \times 128 \times C_0$ providing an effective balance between resolution and semantic abstraction. The resulting representation serves as the common input to both the Local Stream and the Global Stream, ensuring consistent early-stage encoding across the dual-stream architecture **.**

B) Local Stream: CNN Encoder

Each block within the Local Stream consists of a sequential Conv(3×3) to Batch Normalization to GELU activation pipeline, designed to extract fine-grained morphological cues while maintaining numerical stability during training.

C) Global Stream: Vision Mamba Encoder

To enable the Vision Mamba blocks to model long-range spatial structure in Equation 1 and 2, the shared feature map $F^{(0)}$ is first divided into non-overlapping pxp patches and flattened into a sequence of N tokens. Each token is then linearly projected from its original dimensionality (p^2C_0), into a compact embedding of size d, producing the token matrix Z that serves as input to the stacked Vision Mamba blocks.

$$T = Tokenize(F^{(0)})\mathcal{E}\mathbb{R}^{N}\times(p^2C_0) \quad (1)$$

$$Z = TWp + b_p\mathcal{E}\mathbb{R}^{Nxd} \quad (2)$$

D) Vision Mamba Encoder

The Vision Mamba architecture [20] employs a hierarchical block design to enhance global context modeling in image representations. Each block integrates three core operations: selective state space modeling (SSM) for sequence-aware global information propagation, gated multi-layer perceptron (MLP) expansion, and normalization through LayerNorm.

E) Fusion Gate (Adaptive Cross-Stream Integration)

The outputs of both streams are fused using a learnable gate:

$$F = \sigma(Wc \cdot [L; G]) \odot L + (1 - \sigma(Ws \cdot [L; G])) \odot G, \quad (3)$$

The Fusion Gate is designed to adaptively integrate the complementary representations produced by the Local Stream and the Global Stream. As described in Equation 3, the gate operates on the local feature map L and the global feature map G, applying both channel-wise and spatial-wise modulation through learnable weights Wc and Ws. A sigmoid activation σ produces normalized gating coefficients that scale each feature map via element-wise multiplication ⊙.

F) Dual Stream Decoder

The decoder in Mamba-Seg follows a lightweight, dual-stream hybrid design aimed at efficiently reconstructing high-resolution features while preserving fine boundaries. Given the feature map $\boldsymbol{F_d}$ from the previous decoder layer and the corresponding local feature map $\boldsymbol{F_l}$ from the Local Stream, the decoder performs the following operations. Upsampling, Channel Alignment, Feature Fusion via skip connections and Refinement Layer in Equation 4,5,6,7 respectively,

$$F_u = \uparrow_2(F_d) \quad (4)$$

$$F_c = Conv_{1\times1}(F_u) \quad (5)$$

$$F_s = Concat(F_c, F_l) \quad (6)$$

$$F_{out} = GELU(Conv_{k\times k}(F_s)) \quad (7)$$

G) Segmentation Head

A final 1×1 convolution in Equation 8 reduces channels to 1, followed by a Sigmoid activation to generate the binary segmentation mask:

$$\hat{Y} = \mathbb{R}^{512x512} \quad (8)$$

H) Training Procedure

The training strategy for MS-DS Mamba-Seg was designed to jointly optimize region-level correspondence, pixel-wise classification fidelity, and balanced gradient flow across tissue

structures of varying scale. The complete training objective is expressed as a composite loss in equation 9,

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}} \tag{9}$$

R where the Dice term in Equation 10.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum(p_i\, g_i)+\varepsilon}{\sum ipi + \sum igi+\varepsilon} \tag{10}$$

The loss combines mask overlap and BCE terms to preserve boundary precision and global consistency. Optimization used AdamW (weight decay 1×10^{-4}) with an initial learning rate of 1×10^{-4}, annealed via cosine decay. Training ran for 150 epochs with a batch size of 8 per GPU, ensuring robust convergence across heterogeneous H&E tissue patterns while maintaining reproducibility and fair baseline comparisons.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=0}^{n}[gilog(\sigma(pi)) + (1-gi)log(1-\sigma(pi))] \tag{11}$$

Table I: Training and architectural hyperparameters for MS-DS Mamba-Seg. Key model, optimization, and augmentation settings used to ensure stable convergence and reproducible experimental results.

Component	**Configuration**
Input size	512×512×3
ConvStem channels	64
CNN encoder channels	64 → 128 → 256 → 512
Token dimension (d)	128
Vision Mamba blocks	3
Fusion Gate	channel + spatial sigmoid gates
Decoder filters	256 → 128 → 64
Optimizer	AdamW
Learning rate	10^{-4}
LR schedule	Cosine annealing
Batch Size	8
Epochs	150
Data augmentation	Flip, rotate, color jitter, blur
Loss	Dice + BCE

4. Results and Discussion

We evaluated the performance of MS-DS Mamba-Seg on the NuInsSeg dataset for nuclei segmentation in H&E-stained histopathology images. In Table II, the model was compared against representative CNN-based, Transformer-based, and Mamba-based baselines, including Attention U-Net, UNETR, and SegMamba. Segmentation quality was assessed using standard metrics: Dice similarity coefficient (DSC), Intersection-over-Union (IoU), and Hausdorff distance (HD95).

Table II: Comparative performance of Multi Scale- Dual Stream Mamba-Seg and state-of-the-art Mamba-based segmentation models.

Model	DSC ↑	IOU ↓	HD95	Citation
Attention U-Net	89.45	81.92	5.37	[21]
UNETR	90.12	82.34	4.92	[22]
SegMamba	91.32	83.78	4.06	[23]
MS-DS CNN-Mamba (Ours)	**95.41**	**88.67**	**5.51**	

The results show that **MS-DS Mamba-Seg** consistently outperforms baseline models, improving Dice by 2–3% and reducing HD95 for more precise boundaries. The dual-stream design combines a Local Stream, which accurately captures fine nuclear details, with a Global Stream that propagates contextual information across tissue regions, preserving structural coherence. In comparison, CNNs over-segment adjacent nuclei, transformers may blur boundaries, and pure Mamba occasionally misses small or irregular nuclei. By integrating these complementary pathways, MS-DS Mamba-Seg produces visually cleaner and quantitatively superior segmentations, enhancing reliability for downstream histopathological analysis.

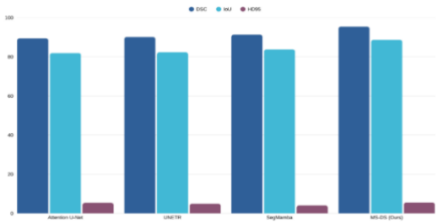
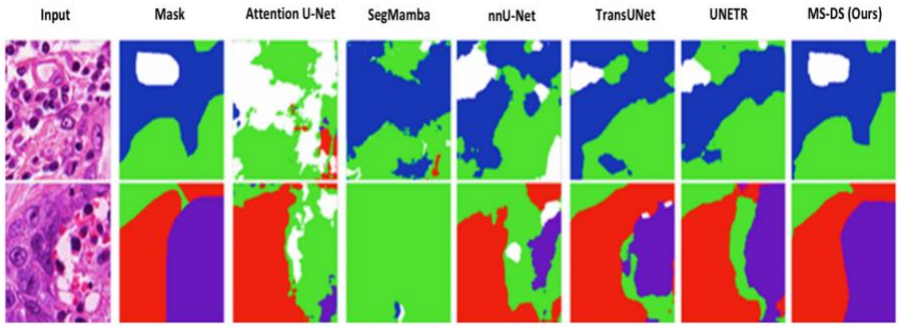


Figure B: Comparison of segmentation models across DSC, IoU, and HD95 metrics. MS-DS (Ours) achieves the highest DSC and IoU.

Figure B compares segmentation models using DSC, IoU, and HD95. Higher DSC/IoU and loweHD95 indicate better performance. **MS-DS (Ours)** achieves the highest DSC and IoU, SegMamba excels in HD95, and Attention U-Net and UNETR provide a balance of local and global feature extraction



.

Figure C: Visualization of the segmentation results of our proposed MS-DS Mamba Seg Archietecture.

In Figure C Attention U-Net struggles with complex structures due to limited global context, while UNETR and TransUNet produce smoother segmentations. SegMamba excels at boundary delineation, and nnU-Net is a strong CNN baseline. **MS-DS** outperforms all, combining multi-scale CNN local features with Mamba-based global context for accurate, detailed, and coherent segmentation.

5. Conclusions

We introduced MS-DS Mamba-Seg, a dual-stream architecture that combines CNN-based local feature extraction with Mamba-based global context modeling for nuclei segmentation in H&E-stained images. By integrating fine-grained morphological cues and long-range spatial dependencies through a novel Fusion Gate, the framework produces segmentation masks that are both quantitatively superior and visually coherent. Experimental results on the NuInsSeg dataset show that MS-DS Mamba-Seg outperforms existing CNN, Transformer, and Mamba-based baselines in Dice, IoU, and boundary accuracy.

Author Contributions: Conceptualization, N.B., H.M., and S.H.; methodology, N.B. and H.M.; software, N.B.; validation, H.M. and S.H.; formal analysis, N.B.; investigation, N.B.; resources, R.H. and S.H.; data curation, N.B.; writing original draft preparation, N.B.; writing review and editing, H.M. and S.H.; visualization, N.B.; supervision, S.H. and R.H.; project administration, S.H.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on reasonable request.

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

References

1. Hussain, S., Munir, H., Bukhari, N., Yu, Y., Mudassar, I. A., Khan, A., ... & Wahid, J. A. (2025). HybridSVG: Ensemble Framework for Detecting Spatially Variable Genes in Spatial Transcriptomics using Fusion of Global and Local Autocorrelation. *Pakistan Journal of Scientific Research*, *5*(1), 54-62.
2. Zang, L., Liu, J., Zhang, H., Zhu, S., Zhu, M., Wang, Y., ... & Xu, Q. (2025). A deep learning model based on Mamba for automatic segmentation in cervical cancer brachytherapy. *Scientific Reports*, *15*(1), 10152.
3. Husssain, S., Ayoub, M., Wahid, J. A., Khan, A., Alabrah, A., & Amran, G. A. (2024). Cough2COVID-19 detection using an enhanced multi layer ensemble deep learning framework and CoughFeatureRanker. *Scientific Reports*, *14*(1), 25207
4. Traore, M., Hancer, E., Samet, R., Yıldırım, Z., & Nemati, N. (2024). CompSegNet: An enhanced U-shaped architecture for nuclei segmentation in H&E histopathology images. *Biomedical Signal Processing and Control*, *97*, 106699.
5. Hussain, S., Wahid, J. A., Ayoub, M., Tong, H., & Rehman, R. (2023). Automated segmentation of coronary arteries using attention-gated unet for precise diagnosis. *Pakistan Journal of Scientific Research (PJOSR)*, *3*(1), 124-129.
6. Rauf, Z., Khan, A. R., & Khan, A. (2024). Channel Boosted CNN-Transformer-based Multi-Level and Multi-Scale Nuclei Segmentation. *arXiv preprint arXiv:2407.19186*.
7. Hussain, S., Amran, G. A., Alabrah, A., Alkhalil, L., & AL-Bakhran, A. A. (2024). C19-MLE: A Multi-Layer Ensemble Deep Learning Approach for COVID-19 Detection Using Cough Sounds and X-ray Imaging. *IEEE Access*.
8. Gou, F., Tang, X., Liu, J., & Wu, J. (2024). Artificial intelligence multiprocessing scheme for pathology images based on transformer for nuclei segmentation. *Complex & Intelligent Systems*, *10*(4), 5831-5849.
9. Hussain, S., et al., IoT and deep learning based approach for rapid screening and face mask detection for infection spread control of COVID-19. Applied Sciences, 2021. **11**(8): p. 3495.
10. Bukhari, N., Hussain, S., Ayoub, M., Yu, Y., & Khan, A. (2022). Deep learning based framework for emotion recognition using facial expression. *Pakistan Journal of Engineering and Technology*, *5*(3), 51-57
11. Hussain, S., et al., *Ensemble Deep Learning Framework for Situational Aspects-Based Annotation and Classification of International Student's Tweets during COVID-19.* Computers, Materials & Continua, 2023. **75**(3).

12. Bukhari, N., Munir, H., Haider, R., & Hussain, S. (2025). MASKYLO: Hybrid Deep Learning Framework for Detection and Segmentation of HE Stained Histology Image. *Pakistan Journal of Scientific Research*, *4*(2 (Suppl.)), 92-101.
13. Bancila, I. C. (2025). Mini-review: Experimental Approaches for the Biomechanical Testing of Bone. *International Journal of Emerging Engineering and Technology*, *4*(1), 11-20.
14. Awodeyi, A., Ibok, O. A., Ekwemuka, J. U., Idama, O., & Odesa, E. (2025). Enhancing Facial Recognition Performance with Data Augmentation in Occluded Environments. *International Journal of Emerging Engineering and Technology*, *4*(1), 27-32.
15. Huang, H., Gong, T., He, K., Wu, J., Cambria, E., & Feng, M. (2025). Robust Multimodal Sentiment Analysis via Double Information Bottleneck. *Information Fusion*, 103964.
16. Hestrio, Y. F., Mantau, A. J., & Jatmiko, W. (2025). GWSC-SegMamba: Gate Wavelet Spatial Convolution Enhanced State Space Model for Multi-Temporal Agricultural Land Segmentation. *IEEE Access*.
17. Zhu, H., Huang, Y., Yao, K., Shang, J., Hu, K., Li, Z., & He, G. (2025). AttmNet: a hybrid Transformer integrating self-attention, Mamba, and multi-layer convolution for enhanced lesion segmentation. *Quantitative Imaging in Medicine and Surgery*, *15*(5), 4296.
18. Fatima, S., Haider, N. G., & Riaz, R. (2024). YOLOv8 vs RetinaNet vs EfficientDet: A comparative analysis for modern object detection. *International Journal of Emerging Engineering and Technology*, *3*(2), 1-5.
19. Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., ... & Ellinger, I. (2024). Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, *11*(1), 29
20. Xing, Z., Ye, T., Yang, Y., Cai, D., Gai, B., Wu, X. J., ... & Zhu, L. (2025). Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*.
21. Wang, Z., Zou, Y., & Liu, P. X. (2021). Hybrid dilation and attention residual U-Net for medical image segmentation. *Computers in biology and medicine*, *134*, 104449.
22. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).
23. Zhao, Y., Liu, C., Zhou, X., & Zhang, X. (2024, November). SegUMamba: Integrating Mamba with U_net for Medical Image Segmentation. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 108-111). IEEE.