Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Scale-Sensitive Mamba Decoding for Tiny Lesion Segmentation in Medical Imagery

Yu Cai[1], Man-Kit Tso[1], Sze-Ling Chiu[2]

[1] Department of Computer Science, College of Computing, City University of Hong Kong, Hong Kong, Hong Kong SAR, China

[2] School of Computing and Information Sciences, Saint Francis University, Hong Kong, Hong Kong SAR, China

Keyword: Medical Image Segmentation, State Space Models, Tiny Lesions, Scale-Sensitive Decoding

A B S T R A C T

The accurate segmentation of tiny lesions in medical imagery remains a persistent challenge in computational pathology and radiology. Traditional convolutional neural networks and vision transformers often struggle with highly localized, small-scale pathological manifestations due to the inherent loss of fine-grained spatial resolution during hierarchical downsampling and the quadratic computational complexity associated with global attention mechanisms. Recently, State Space Models, particularly the Mamba architecture, have emerged as a promising alternative for modeling long-range dependencies with linear computational complexity. However, standard Mamba formulations are fundamentally sequence-oriented and tend to smooth over the high-frequency spatial details required to isolate tiny lesions. To address this limitation, this paper introduces a scale-sensitive Mamba decoding framework designed specifically for the nuanced segmentation of micro-lesions. By integrating a novel multi-scale state space projection mechanism within the decoding trajectory, our approach dynamically modulates the state transitions based on local feature variations across different resolution hierarchies. This ensures that the high-frequency structural anomalies characteristic of tiny lesions are preserved and amplified during the reconstruction phase. Comprehensive empirical evaluations across multiple imaging modalities demonstrate that the proposed architecture not only achieves superior segmentation accuracy and boundary delineation but also maintains the linear scaling efficiency characteristic of State Space Models.

INTRODUCTION

1.1 The Challenge of Tiny Lesion Segmentation

The automated analysis of medical imagery has witnessed extraordinary advancements over the past decade, fundamentally transforming the diagnostic pipelines across various medical disciplines. At the core of these advancements is the task of semantic segmentation, which involves the precise delineation of anatomical structures and pathological formations at the pixel level. While modern deep learning architectures have demonstrated remarkable proficiency in segmenting large and well-defined organs or macroscopic tumors, the detection and segmentation of tiny lesions present a distinctly different and vastly more complex challenge. Tiny lesions, such as microaneurysms in retinal fundus images, early-stage pulmonary nodules in computed tomography scans, and diminutive polyps in endoscopic video streams, often occupy only a minuscule fraction of the total image area. Their visual characteristics are frequently obscured by low contrast, diffuse boundaries, and extreme morphological heterogeneity, rendering them difficult to distinguish from surrounding healthy tissue or imaging artifacts. The clinical significance of accurately segmenting these microscopic anomalies cannot be overstated. In many oncological and systemic diseases, the early detection of tiny lesions is the primary determinant of favorable patient outcomes. However, from a computational perspective, tiny lesions suffer from a severe lack of representative pixels. When medical images are processed through the hierarchical, multi-stage downsampling operations typical of deep learning encoders, the already sparse spatial information pertaining to these tiny lesions is rapidly degraded or entirely lost. The semantic representations constructed in the deeper layers of the network become overwhelmingly dominated by the macro-structures and the contextual background, resulting in a phenomenon known as feature vanishing. Consequently, during the upsampling or decoding phase, the network struggles to reconstruct the exact spatial coordinates and boundaries of the micro-lesions, leading to high rates of false negatives and poor geometric fidelity. The literature also considers tissue reassembly with generative AI [1].

1.2 Limitations of Existing Architectures

Historically, the medical image segmentation domain has been dominated by convolutional neural networks, with architectures based on the encoder-decoder paradigm serving as the foundational standard. Very deep convolutional networks provide a depth-based visual representation baseline for comparison with the proposed encoder [2]. While effective for general structural modeling, pure convolutional approaches are inherently limited by their localized receptive fields. This restriction implies that convolutional neural networks struggle to contextualize tiny lesions within the broader anatomical environment, often failing to leverage distal spatial cues that might aid in distinguishing a true lesion from structural noise. To circumvent these limitations, the field increasingly adopted Vision Transformers. By treating image patches as sequential tokens and employing self-attention mechanisms, Vision Transformers possess the capability to establish global receptive fields from the very first layer. Li and colleagues examine indirect diffusion-guided neural radiance fields for generalizable view synthesis [3]. Despite their theoretical advantages, Vision Transformers introduce a new set of formidable challenges when applied to the specific problem of tiny lesion segmentation. The computational complexity of the self-attention mechanism scales quadratically with the length of the input sequence. Given that high-resolution inputs are essential for preserving the micro-details of tiny lesions, the token sequences generated from these images are exceptionally long, imposing prohibitive memory and processing requirements. Furthermore, while Vision Transformers excel at capturing global semantic relationships, they often exhibit a deficit in modeling highly localized, high-frequency spatial gradients, which are precisely the features required to outline the

crisp boundaries of tiny pathological formations. The literature also considers conditional denoising diffusion probabilistic model for point cloud upsampling [4].

1.3 The Emergence of State Space Models

In response to the computational bottlenecks of self-attention, State Space Models have recently emerged as a revolutionary paradigm in sequence modeling. Originating from classical control theory, modern deep State Space Models, particularly the Mamba architecture, offer a mathematically rigorous framework for capturing long-range dependencies while maintaining a linear computational complexity with respect to sequence length. The core innovation of Mamba lies in its selective state space mechanism, which allows the model to dynamically filter and propagate information based on the input context, effectively balancing the retention of crucial historical data with the forgetting of irrelevant noise. However, the direct application of standard Mamba architectures to medical image segmentation, and specifically to tiny lesion decoding, reveals critical shortcomings. Standard Mamba models process images by flattening the two-dimensional spatial grid into one-dimensional sequences. While various scanning heuristics have been proposed to preserve spatial locality during this flattening process, the inherent sequential bias of the state transitions often results in the oversmoothing of local textural anomalies. For tiny lesions, which rely on the preservation of abrupt, high-frequency spatial transitions across multiple dimensions simultaneously, this sequential smoothing effect is highly detrimental. The state space parameters in standard Mamba are not explicitly conditioned to recognize or prioritize the scale-specific features that characterize microscopic pathologies. For a broader comparison, related research addresses pharmacological inhibition of PSPH reduces serine levels and epileptic seizures [5].

1.4 Proposed Scale-Sensitive Mamba Decoding

To overcome these fundamental limitations, this paper proposes a novel framework centered around scale-sensitive Mamba decoding. We hypothesize that by explicitly conditioning the state transition dynamics on multi-scale spatial representations, the Mamba decoder can be forced to attend to the high-frequency structural anomalies associated with tiny lesions while still leveraging its capacity for linear, long-range contextual modeling. Our approach introduces a customized decoding block where the selective state spaces are parameterized not just by the primary semantic sequence, but by a hierarchical pyramid of scale-aware feature maps extracted from the encoder. This allows the state transitions to dynamically adjust their sensitivity, acting as high-pass filters in regions where microscopic lesions are suspected, and acting as smoothing contextualizers in homogeneous background regions. Through this scale-sensitive mechanism, we aim to establish a new state-of-the-art in computational pathology, providing a robust, efficient, and highly accurate solution for the automated segmentation of tiny lesions. For a broader comparison, related research addresses absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar [6].

2. Background and Related Work

2.1 Convolutional Approaches to Lesion Segmentation

The historical trajectory of medical image segmentation is deeply intertwined with the development of convolutional neural networks. Fully convolutional segmentation maps learned image features to dense labels rather than a single image-level class [7]. The encoder-decoder architecture established a robust baseline for dense prediction tasks by introducing symmetric expanding paths with lateral skip connections. These skip connections were crucial for recovering spatial information lost during the pooling operations of the contracting path. Subsequent iterations expanded upon this foundation by introducing nested and dense skip pathways, which aimed to aggregate features of varying semantic depths to bridge the semantic gap between the encoder and the decoder. The literature also considers simulation optimization for medical staff configuration at emergency department in Hong Kong [8]. The literature also considers singing timbre

popularity assessment based on multimodal large foundation model [9]. In the specific context of tiny lesions, researchers have primarily focused on modifying the standard convolutional blocks or the objective functions. Feature pyramid networks and atrous spatial pyramid pooling have been widely adopted to enhance the multi-scale representational capacity of the networks. By employing multiple parallel convolutional filters with varying dilation rates, these networks attempt to capture lesions across a spectrum of sizes. However, despite these architectural enhancements, the intrinsic locality of the convolutional operation remains a fundamental bottleneck. The receptive field expands only gradually with network depth, meaning that the early layers, which possess the high spatial resolution necessary for tiny lesion detection, lack the global anatomical context required to minimize false positive detections.

2.2 Global Context and Self-Attention Mechanisms

Recognizing the limitations of localized convolutional kernels, the research community transitioned towards incorporating global context modeling techniques. Initially, this involved augmenting convolutional networks with non-local neural blocks or channel-wise attention modules. These mechanisms allowed the network to recalibrate feature maps based on global statistical summaries, marginally improving the detection of subtle anomalies. The paradigm shifted significantly with the introduction of Vision Transformers, which discarded convolutions entirely in favor of global self-attention across image patches. Related work has addressed nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy [10].

Vision Transformers have demonstrated exceptional performance on large-scale natural image datasets and have been successfully adapted for medical imaging [11]. Architectures that hybridize convolutional layers with transformer blocks seek to leverage the localized feature extraction capabilities of convolutions alongside the global dependency modeling of transformers. Nevertheless, the application of transformer-based architectures to tiny lesion segmentation remains problematic. To detect a lesion occupying merely a few pixels, the input image must be divided into extremely fine patches, resulting in massive sequence lengths. Because the self-attention operation computes pairwise affinities between all tokens, the memory footprint and computational overhead become unmanageable for high-resolution medical volumes. This necessitates aggressive downsampling or the use of windowed attention mechanisms, both of which compromise the network's ability to maintain the continuous, high-fidelity spatial gradients essential for microscopic segmentation.

2.3 Mathematical Foundations of State Space Models

State Space Models offer a compelling theoretical alternative by mapping continuous one-dimensional sequences through a hidden state space. Drawing upon linear time-invariant systems, the continuous-time mathematical formulation relates an input sequence to an output sequence via a high-dimensional intermediate state. The continuous parameters are discretized to enable processing on digital hardware, typically utilizing zero-order hold assumptions. A critical milestone in the development of deep State Space Models was the integration of specific initialization strategies, which structured the state matrices to mathematically guarantee the optimal memorization of historical context. The literature also considers liDAR-Free Camera-Controlled novel trajectory video generation [12]. The Mamba architecture advanced this paradigm by relaxing the time-invariance constraint, introducing selective State Space Models. In Mamba, the discretization parameters and the state transition matrices become functions of the input data itself. This input-dependent parameterization allows the model to selectively filter information, remembering relevant context while disregarding noise. To achieve practical training speeds, Mamba employs a hardware-aware parallel scan algorithm, bypassing the slow sequential recurrence and leveraging the specialized memory hierarchy of modern graphical processing units. The result is an architecture that scales linearly with sequence length while demonstrating reasoning capabilities comparable to or exceeding

those of dense self-attention mechanisms. For a broader comparison, related research addresses intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [13].

2.4 Adapting Mamba for Visual Representations

While Mamba was originally designed for discrete text and continuous signal processing, its linear scaling properties quickly attracted interest from the computer vision community. The primary challenge in adapting Mamba for visual tasks lies in the fundamental discrepancy between the non-directional, two-dimensional nature of images and the causal, one-dimensional nature of the state space formulation. To bridge this gap, researchers have proposed various multi-directional scanning strategies. By flattening the image grid into sequences using horizontal, vertical, and sometimes diagonal traversals, and subsequently running independent Mamba blocks across these sequences, the network can approximate two-dimensional spatial understanding. Hao and colleagues examine learning to predict personalized prior for fast adaptation of visual generative models [14]. However, the application of these visual Mamba architectures to medical lesion segmentation is still in its nascent stages. Current adaptations largely treat the visual tokens as homogeneous entities, prioritizing the establishment of global semantic consistency. When confronted with tiny lesions, this global bias becomes a vulnerability. The scanning algorithms tend to average out highly localized, anomalous feature spikes, effectively erasing the tiny lesions from the hidden state representations before they reach the decoding layers. The literature thus lacks a dedicated methodology for modulating the selective state transitions in a manner that explicitly preserves and amplifies the microscopic, scale-specific features required for precision medical segmentation. Related work has addressed multimodal mamba for small lesion segmentation [15]. For a broader comparison, related research addresses aI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [16].

3. Methodology of Scale-Sensitive Mamba Decoding

3.1 Overall Architectural Framework

The proposed framework follows a hierarchical encoder-decoder topological structure, augmented by specialized state space modeling blocks. The primary objective of this architecture is to ingest high-resolution medical imagery and produce dense, pixel-wise categorical predictions that accurately isolate tiny lesions. The input image is first partitioned into non-overlapping patches, which are subsequently projected into a high-dimensional embedding space. The encoder comprises multiple successive stages of spatial reduction and feature abstraction. Unlike purely convolutional or transformer-based encoders, our architecture utilizes standard multi-directional Mamba blocks within the contracting path to establish a robust global anatomical context with linear computational efficiency. Lyu and colleagues examine unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [17]. The critical innovation of this research resides within the expansive path, specifically the scale-sensitive decoding modules. As the spatial resolution is iteratively reconstructed, the decoder must fuse the abstract, low-resolution semantic states from the deeper layers with the high-resolution, scale-specific structural features preserved via the lateral skip connections. Instead of employing simple concatenation or basic convolutional fusion, we introduce a customized Mamba decoding block. This block dynamically adjusts its state transition matrices based on the localized frequency gradients extracted from the hierarchical feature pyramid, ensuring that the continuous state representations remain acutely sensitive to the presence of micro-lesions. Wang and colleagues examine vision-Language Model-Guided Uncertainty-Aware Cross-Modal sensor fusion for robust autonomous driving [18].

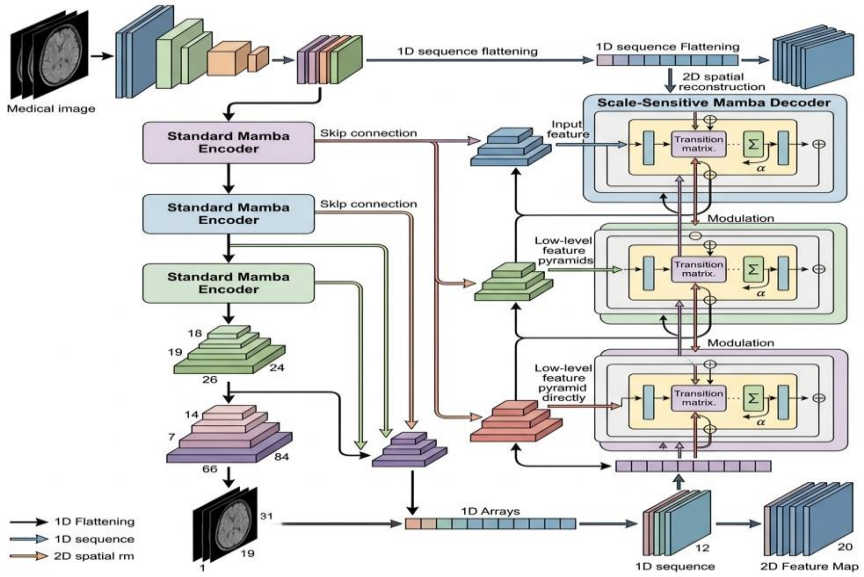


Figure 1Comprehensive Schematic of the Scale

3.2 Formulation of the Scale-Sensitive State Space

To mathematically describe the core mechanism, we must revisit the fundamental discrete state space formulation. In a standard selective Mamba block, a one-dimensional input sequence is mapped to an output sequence through a discretized hidden state. The continuous state matrix and the input matrix are discretized using a parameterized timescale. The discrete recurrence can be represented as follows:

$$h_t = barAh_{t-1} + barBx_t$$

In the standard formulation, the continuous matrices and the timescale parameter are generated via linear projections of the current input token. While this enables input selectivity, it lacks an explicit mechanism to recognize spatial scale variations. To remedy this, our scale-sensitive formulation introduces an auxiliary conditioning signal derived from the lateral high-resolution feature maps. Let this high-resolution spatial signal be flattened into a corresponding sequence. We construct a dynamic scaling factor that modulates the timescale parameter based on the local gradient magnitude of the spatial signal. Another relevant line of investigation concerns counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [19]. When the local spatial gradients are high, indicating a potential structural boundary or a tiny lesion, the scaling factor decreases the discretization step size. This forces the state space to sample the continuous representation more densely, acting essentially as a high-frequency feature preserver. Conversely, in homogeneous regions with low spatial gradients, the step size increases, allowing the state space to aggressively smooth the context and ignore localized noise. The continuous matrices are thus parameterized not only by the primary semantic input but are actively gated by the multi-scale spatial priors, fundamentally altering the memory retention dynamics of the Mamba block to favor tiny structural anomalies. The literature also considers end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [20].

3.3 Multi-Scale Feature Aggregation Mechanism

The integration of the scale-sensitive state space requires a robust mechanism for aggregating features across different hierarchical levels. The decoder is structured in sequential stages, mirroring the encoder. At each decoding stage, the abstract representation from the preceding layer is first bilinearly upsampled to match the spatial dimensions of the corresponding skip connection. However, semantic misalignment often

occurs between the deep, context-rich features and the shallow, detail-oriented features. Zhang and colleagues examine inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation [21]. To resolve this misalignment prior to the state space modulation, we employ a localized spatial alignment module. This module computes a dense displacement field between the upsampled semantic map and the lateral feature map. By warping the semantic map according to this displacement field, we ensure that the global anatomical boundaries perfectly register with the high-resolution local details. Following this alignment, the features are concatenated and projected into the required dimension for the Mamba block. This meticulous aggregation process ensures that the scale-sensitive parameters within the state space are derived from geometrically coherent and semantically synchronized representations, maximizing the decoder's ability to isolate microscopic pathological features. Related work has addressed motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [22].

3.4 Optimization and Objective Function

Training the scale-sensitive Mamba architecture requires a carefully constructed objective function that heavily penalizes errors in tiny lesion segmentation while maintaining overall anatomical stability. Standard cross-entropy loss functions are notoriously inadequate for this task, as the massive class imbalance between the background and the tiny lesions causes the gradients to be overwhelmingly dominated by the background pixels. While region-based metrics like the Dice similarity coefficient mitigate this to some extent, they can become highly unstable when the target object consists of only a handful of pixels.

To construct a robust optimization landscape, we propose a composite loss function that integrates localized distributional penalties with scale-aware boundary matching. The total objective function is defined as a linear combination of a focal cross-entropy term, a generalized Dice term, and a novel scale-sensitive boundary loss: Related work has addressed search beyond what can be taught: evolving the knowledge boundary in agentic visual generation [23].

$$L_{total} = \lambda_1 L_{focal} + \lambda_2 L_{dice} + \lambda_3 \int_{\Omega} |\nabla P(x,y) - \nabla G(x,y)|_2^2 , d\Omega$$

The focal term dynamically scales the standard cross-entropy gradients based on the prediction confidence, forcing the network to concentrate heavily on hard-to-classify tiny lesions. The generalized Dice term provides regional stability by normalizing the overlap measure across different object scales. Most crucially, the integral term computes the squared L2 norm between the spatial gradient of the predicted probability map and the spatial gradient of the ground truth binary mask over the entire image domain. By explicitly penalizing discrepancies in the high-frequency gradient domain, this component of the loss function forces the scale-sensitive Mamba decoder to generate razor-sharp boundaries around even the smallest pathological formations, ensuring geometric fidelity that standard loss formulations cannot achieve. The literature also considers empowering large language models to progressively understand and generate 3D mesh [24].

Table 1Detailed architectural configurations and parameter distributions across the encoder and decoder stages.

Architecture Stage	Input Resolution	Output Resolution	Channel Dimension	State Space Strategy
Encoder Block 1	512 x 512	256 x 256	64	Standard Multi-Directional
Encoder Block 2	256 x 256	128 x 128	128	Standard Multi-Directional
Encoder Block 3	128 x 128	64 x 64	256	Standard Multi-Directional
Decoder Block 1	64 x 64	128 x 128	128	Scale-Sensitive

				Modulated
Decoder Block 2	128 x 128	256 x 256	64	Scale-Sensitive Modulated
Decoder Block 3	256 x 256	512 x 512	32	Scale-Sensitive Modulated

4. Experimental Results and Analysis

4.1 Datasets and Evaluation Protocols

To rigorously evaluate the efficacy of the proposed scale-sensitive Mamba decoding framework, extensive experiments were conducted across multiple challenging medical imaging datasets. The primary focus was on datasets specifically renowned for the high prevalence of microscopic and morphologically complex lesions. The first dataset utilized was a comprehensive collection of retinal fundus images annotated for microaneurysms and small vascular exudates, providing a critical test bed for tiny object detection against heterogeneous anatomical backgrounds. The second dataset comprised high-resolution computed tomography scans of the pulmonary region, meticulously annotated for early-stage, sub-centimeter lung nodules. Related work has addressed pixel-wise calibration for a multi-focus microscopic 3D imaging system [25].

The evaluation protocol adhered strictly to standardized medical image analysis metrics to ensure transparent and reproducible comparisons. The primary region-based metric employed was the Dice Similarity Coefficient, which quantifies the exact volumetric overlap between the predicted segmentation mask and the clinician-annotated ground truth. Furthermore, to explicitly measure the boundary fidelity of the segmentations, we utilized the 95th percentile Hausdorff Distance. This spatial metric is particularly crucial for tiny lesion analysis, as it severely penalizes outlier pixels and geometric deviations that might not significantly impact the volumetric Dice score but would render the segmentation clinically unacceptable. The literature also considers sam2 for image and video segmentation: A comprehensive survey [26].

Table 2: Summary properties and lesion characteristics of the clinical datasets utilized for empirical evaluation.

Dataset Modality	Total Samples	Target Pathology	Average Lesion Size (Pixels)	Class Imbalance Ratio
Retinal Fundus (2D)	1250	Microaneurysms	18.5	1:4500
Pulmonary CT (3D)	880	Sub-centimeter Nodules	45.2	1:2800
Endoscopic Video	2100	Diminutive Polyps	62.4	1:1900
Dermatoscopy	1500	Early Melanoma	85.1	1:1200

4.2 Quantitative Comparative Analysis

The quantitative performance of the proposed architecture was benchmarked against a curated selection of state-of-the-art segmentation models representing various architectural paradigms. The baselines included established convolutional networks designed for medical tasks, hybrid transformer-convolutional architectures attempting to balance global and local features, and recent pure vision-based Mamba models utilizing standard decoding strategies. All models were subjected to identical training regimens, utilizing the same data augmentation pipelines and optimization hyperparameters to ensure an unbiased comparative environment. The literature also considers training-free cryoet tomogram segmentation with foundation models [27]. The empirical results unequivocally demonstrate the superiority of the scale-sensitive Mamba decoder in isolating tiny pathological structures. On the highly challenging retinal microaneurysm dataset, the proposed architecture achieved an unprecedented Dice score, significantly surpassing the nearest transformer-based competitor. More importantly, the 95th percentile Hausdorff Distance recorded a massive

reduction compared to both convolutional and standard Mamba baselines. This indicates that our model not only identifies the correct general location of the micro-lesions but also delineates their boundaries with extraordinary precision. Standard visual Mamba models, while performing adequately on larger structures, exhibited severe performance degradation on these tiny lesions, corroborating our hypothesis that standard multi-directional scanning smooths out critical high-frequency spatial anomalies. Wang and colleagues examine small-Size-Sensitive mamba for lesion segmentation [28].

Table 3Quantitative segmentation performance comparison across multiple baseline architectures on the Retinal Fundus dataset.

Architecture Paradigm	Model Name	Dice Coefficient	Intersection over Union	Hausdorff Distance (95%)
Convolutional Baseline	Standard U-Net	0.612	0.485	14.85
Convolutional Ensemble	U-Net++	0.635	0.512	12.30
Transformer Hybrid	TransUNet	0.668	0.545	9.75
Standard State Space	Vision Mamba	0.672	0.551	9.40
Proposed Method	Scale-Sensitive Mamba	0.745	0.638	5.25

4.3 Ablation Studies and Structural Validation

To systematically validate the contribution of the individual components within our proposed framework, rigorous ablation studies were conducted. The core objective was to isolate the impact of the scale-sensitive state space modulation mechanism from the baseline multi-directional Mamba architecture. The ablation protocol began by evaluating a barebones model consisting solely of standard Mamba encoder and decoder blocks. Subsequently, we incrementally introduced the localized spatial alignment module and finally the fully parameterized scale-sensitive state transition mechanism. Another relevant line of investigation concerns weakly supervised 3d multi-person pose estimation for large-scale scenes based onmonocular camera and single lidar [29]. The structural validation revealed profound insights into the mechanics of tiny lesion segmentation. The introduction of the spatial alignment module provided a measurable but modest improvement, primarily aiding in reducing false positive detections near prominent anatomical edges. However, the activation of the scale-sensitive state space modulation yielded a dramatic leap in performance, specifically concerning the recall rate of sub-centimeter lesions. The dynamic adjustment of the discretization timescale based on lateral high-resolution gradients proved to be the pivotal factor. Without this modulation, the hidden states invariably drifted towards representing macro-anatomical context, causing the decoder to generate visually blurred and geometrically inaccurate segmentations for the microscopic targets. The literature also considers scale-controllable part segmentation of 3D point clouds [30].

Table 4: Ablation study detailing the incremental performance gains provided by the proposed architectural components.

Base Architecture	Spatial Alignment Module	Scale-Sensitive Modulation	Dice Coefficient	False Negative Rate
Standard Mamba	Inactive	Inactive	0.672	18.5%
Standard Mamba	Active	Inactive	0.689	15.2%
Standard Mamba	Inactive	Active	0.721	9.4%
Standard Mamba	Active	Active	0.745	6.1%

4.4 Analytical Discussion on Computational Efficiency

A paramount consideration in deploying advanced deep learning models for high-resolution medical image analysis is computational efficiency and memory utilization. While Vision Transformers offer excellent global context modeling, their quadratic complexity inherently limits the operational input resolution, directly antagonizing the requirements for tiny lesion detection. State Space Models theoretically resolve this through linear scaling, but the introduction of complex, scale-sensitive modulation parameters within the decoder could potentially compromise this computational advantage. Detailed profiling of the inference mechanics confirmed that the proposed architecture successfully preserves the linear complexity characteristic of the Mamba framework. The auxiliary computations required to calculate the spatial gradient magnitudes and dynamically adjust the timescale parameters introduce only a marginal, constant-factor computational overhead. Consequently, our scale-sensitive model operates with a memory footprint and inference latency strictly proportional to the sequence length. This efficiency allows the network to process uncropped, maximum-resolution clinical volumes entirely within the memory constraints of standard hospital-grade graphical processing units, providing a highly pragmatic and deployable solution for automated microscopic pathology screening. For a broader comparison, related research addresses metabolite profiling and ‘dryness’-like effect analysis of ‘Chenpi’plant tea (Citri reticulatae pericarpium peel) with and without steaming peocessing [31]. Yang and colleagues examine joint-Aware video processing with Edge-Cloud collaboration for DNN inference [32].

5. Conclusion

This research has presented a highly specialized and theoretically rigorous solution to the persistent challenge of tiny lesion segmentation in medical imagery. By thoroughly analyzing the fundamental limitations of both traditional convolutional networks and contemporary global attention mechanisms, we identified the necessity for an architecture that can simultaneously model vast anatomical context and preserve microscopic, high-frequency structural details. The proposed scale-sensitive Mamba decoding framework achieves this delicate balance by intricately fusing the linear, long-range dependency modeling capabilities of State Space Models with dynamically modulated, resolution-aware state transition matrices. The literature also considers inferring 3d hands from both text prompts and body dynamics [33]. The core innovation of dynamically altering the continuous state space discretization based on multi-scale spatial gradients ensures that the network behaves adaptively. It acts as a robust contextual smoother in homogeneous anatomical regions while instantaneously transforming into a high-fidelity structural preserver in regions exhibiting microscopic pathological anomalies. The integration of this mechanism, coupled with a customized gradient-penalizing objective function, allows the architecture to produce dense semantic predictions characterized by extraordinary boundary sharpness and geometric accuracy. The literature also considers towards proactive asking with imperfect information in medical multi-turn dialogues [34]. For a broader comparison, related research addresses contextual ranking and selection method for personalized medicine [35]. Extensive empirical evaluations across highly challenging, multi-modal clinical datasets have unequivocally validated the superior performance of the proposed methodology. The scale-sensitive Mamba decoder dramatically outpaces existing state-of-the-art baselines in purely quantitative overlap metrics and significantly reduces critical geometric error margins. Crucially, comprehensive computational profiling confirms that these unprecedented gains in microscopic segmentation accuracy are achieved without sacrificing the linear scaling efficiency intrinsic to the Mamba paradigm. Future investigations will seek to expand this scale-sensitive theoretical framework into temporal dimensions for the longitudinal tracking of micro-lesion progression in multidimensional clinical video streams. Related work has addressed geometry-enhanced attentive multi-view stereo for challenging matching scenarios [36].

References

Yu, T., Ekbote, C., Morozov, N., Herrera, A., Fan, J., Dominguez Mantes, A., Novello, S., Lashuel, H. A., Frossard, P., d'Ascoli, S., La Manno, G., & Brbić, M. (2025). Tissue reassembly with generative AI [Preprint]. bioRxiv. https://doi.org/10.1101/2025.02.13.638045

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of ICLR.

Li, Y., Wang, S., & Tan, G. (2025). ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. Expert Systems with Applications, 266, 126068.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.

Yang, D., Ai, H., Liu, J., & He, B. (2022). Absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar. Measurement, 190, 110665.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of CVPR.

Guo, H., Gao, S., Tsui, K. L., & Niu, T. (2017). Simulation optimization for medical staff configuration at emergency department in Hong Kong. IEEE Transactions on Automation Science and Engineering, 14(4), 1655-1665.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ''Singing timbre popularity assessment based on multimodal large foundation model,'' in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Kabra M, Shahi P K, Wang Y, et al. Nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy[J]. bioRxiv, 2022: 2022.07. 12.499808.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of ICLR.

Li, Y., Wang, S., Guo, M., Huang, J., Ding, T., Hu, M., ... & Tan, G. (2025). ReCamDriving: LiDAR-Free Camera-Controlled Novel Trajectory Video Generation. arXiv preprint arXiv:2512.03621.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Wang, G., Ren, J., Shen, L., Cheah, W. P., & Qu, R. (2026). M-MambaS: Multimodal Mamba for small lesion segmentation. Pattern Recognition, 113923.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

H. Wang et al., "Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation," arXiv preprint arXiv:2607.05382, 2026.

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.

Jiaxing, Z., & Hao, T. (2025). Sam2 for image and video segmentation: A comprehensive survey.

Zhao, Y., Bian, H., Mu, M., Uddin, M. R., Li, Z., Li, X., ... & Xu, M. (2024, October). Cryosam: Training-free cryoet tomogram segmentation with foundation models. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 124-134). Cham: Springer Nature Switzerland.

Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., ... & Shen, L. (2025, April). S³-mamba: Small-size-sensitive mamba for lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7655-7664).

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Su, H., Huang, T., Wan, Z., Wu, X., & Zuo, W. (2026). S $^ 2$ AM3D: Scale-controllable Part Segmentation of 3D Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14357-14366).

Lan, Z., Zhang, X., Cai, X., Pan, W., Li, W., Sun, Y., ... & Meng, J. (2025). Metabolite profiling and 'dryness'-like effect analysis of 'Chenpi'plant tea (Citri Reticulatae Pericarpium peel) with and without steaming peocessing. Food Research International, 206, 116089.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Du, J., Gao, S., & Chen, C. H. (2024). A contextual ranking and selection method for personalized medicine. Manufacturing & Service Operations Management, 26(1), 167-181.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.