ACCURATE 3D MEDICAL IMAGE SEGMENTATION WITH MAMBAS

Luca Lumetti, Vittorio Pipoli*, Kevin Marchesini*, Elisa Ficarra, Costantino Grana, Federico Bolelli*

University of Modena and Reggio Emilia, Italy

{name.surname}@unimore.it

ABSTRACT

CNNs and Transformer-based architectures are recently dominating the field of 3D medical segmentation. While CNNs face limitations in the local receptive field, Transformers require significant memory and data, making them less suitable for analyzing large 3D medical volumes. Consequently, fully convolutional network models like U-Net are still leading the 3D segmentation scenario. Although efforts have been made to reduce the Transformers computational complexity, such optimized models still struggle with content-based reasoning. This paper examines Mamba, a Recurrent Neural Network (RNN) based on State Space Models (SSMs), which achieves linear complexity and has outperformed Transformers in long-sequence tasks. Specifically, we assess Mamba's performance in 3D medical segmentation using three widely recognized and commonly employed datasets and propose architectural enhancements to improve its segmentation effectiveness by mitigating the primary shortcomings of existing Mamba-based solutions.

***Index Terms*—** 3D Segmentation, Mamba, Medical Imaging, RNN

1. INTRODUCTION

Automatic algorithms for the segmentation of anatomical structures are widely adopted in medical image analysis to aid medical practice and provide support for surgical planning [1]. Among the existing architectures, Convolutional Neural Networks (CNNs) [2] have certainly dominated the scene, with U-Net [3] emerging as a particularly effective model due to its U-shaped encoder-decoder structure with skip connections. Following the success of U-Net, numerous variants, including Res-U-Net [4], Dense-U-Net [5], V-Net [6], 3D U-Net, and nnU-Net [7], have introduced enhancements, each aiming to improve segmentation quality through structural adjustments. Yet, CNNs inherently struggle with capturing global dependencies due to the localized nature of convolutional operations.

To address these limitations, recent studies have aimed to integrate Transformer attention mechanisms [8] into U-Net architectures [9–11], enhancing models' ability to capture both local and global features. Variants such as TransUNet [12], UNETR [13], and Swin-UNETR [14] have achieved performance gains by integrating multi-head attention layers; however, these adaptations come with increased computational costs, especially in large 3D volumes, due to the quadratic complexity of standard attention mechanisms. In response, methods like window-based and axial-based attention have been proposed, and new linear attention mechanisms [15–17] have emerged, though they still fall short for long-context modeling in high-dimensional data.

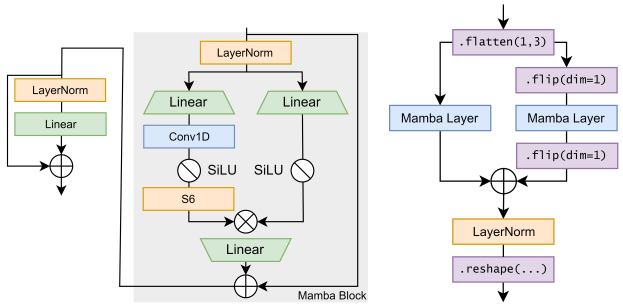
More recently, a State-Space Model (SSM)-based architecture, Mamba [18], has shown remarkable promise in tasks requiring long-context reasoning, such as NLP and genomics, effectively handling inputs of up to a million tokens with linear-time complexity. Mamba has achieved state-of-the-art performance on various tasks compared to models like GPT-J-6B [19] and Pythia [20], establishing itself as a strong alternative to Transformers for long-sequence processing.

Given their effectiveness and versatility, Mamba-based architectures have been rapidly adapted to various domains, including Computer Vision [21]. Several researchers have devoted significant efforts to adapting the Mamba architecture for both 2D and 3D segmentation, demonstrating promising results [22–26]. Initial adaptations, such as UMambaEnc and UMambaBot [27], incorporate Mamba layers into U-Net architectures, either in the encoder or in the bottleneck.

Despite the efficient segmentations provided by these configurations, we identify a significant drawback inherent in these methodologies that arises from the recurrent nature of Mamba, which we denote as *initial hidden state problem*. Specifically, when Mamba processes a sequence, it lacks context for the early elements. In contrast, by the time it processes the final ones, it has observed nearly the entire sequence. Consequently, when a 3D volume is unrolled and processed by a Mamba layer, the capability to process the initial portions is generally weaker compared to that of the later portions of the volume.

Paper Contribution. This paper explores further improvements to Mamba-based architectures for 3D segmentation and addresses the initial hidden state problem by proposing bidirectional and multidirectional Mambas. Experimental evaluations on datasets like MSD BrainTumour, Synapse

*Equal contribution. Authors are allowed to list their names first in their respective CVs.



(a) (b)

Fig. 1: (a) depicts our (unidirectional) Mamba Layer (in gray the original Mamba block), while (b) represents our bidirectional 3D Mamba layer, which includes (a).

Multi-organ, and ACDC demonstrate the potential of these Mamba-enhanced models for high-dimensional segmentation tasks. Our code is publicly available to encourage further development and ensure the reproducibility of the results.[1]

2. METHOD

State-Space Models (SSMs) are mathematical frameworks employed to represent dynamic systems by mapping inputs to latent states and outputs. To efficiently handle long sequences, the Mamba architecture leverages structured SSMs, integrating the HiPPO theory [28] to enhance memory retention and utilizing a selection mechanism to filter relevant information. This approach addresses the limitations inherent in traditional SSMs, particularly when applied within neural networks. Mamba's efficient implementation combines these elements with linear projections and convolutions, making it suitable for tasks that require complex sequence modeling. For a comprehensive understanding, the readers are referred to the original Mamba publication [18].

Segmenting Volumes with Mamba. One of the significant issues in medical image segmentation is patch extraction and down-sampling, which hinders voxel-wise details and contextual information to make the training process computationally feasible. Unlike Vision Transformer, which faces quadratic self-attention costs and requires patch extraction to reduce input size, Mamba allows linear-time sequence modeling of the input, preventing any sampling. However, like Transformers, Mamba processes only one-dimensional sequences. Thus, applying it to two-dimensional images or three-dimensional volumes requires flattening pixels or voxels into a one-dimensional sequence.

Motivated by these considerations and by taking inspiration from the ViT architecture [29], our first proposal integrates a wrapped version of the Mamba block into a U-Net-like U-shaped architecture. The wrapper, named Mamba

[1] https://github.com/LucaLumetti/TamingMambas

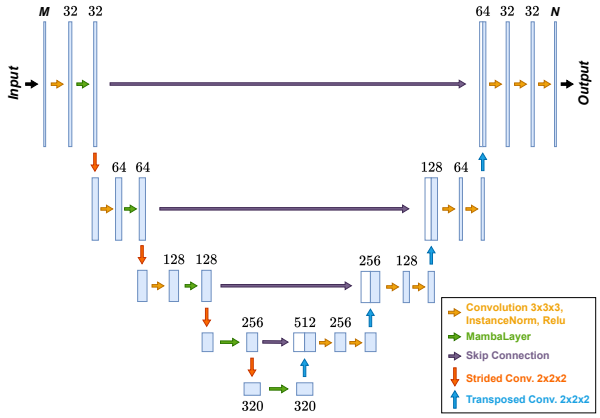


Fig. 2: U-Net-like architecture integrating our proposed Mamba layers. By properly selecting the Mamba layers (green arrows), *Unidirectional*, *Bidirectional*, and *Multidirectional* are obtained.

layer, consists of an additional LayerNorm and an MLP head followed by a skip connection and takes a 3D volume with dimensions `(B, H, W, D, C)` flattened along the spatial dimensions as input.

More specifically, we incorporate a single unidirectional Mamba layer (Fig. 1a) before each pooling convolution and the bottleneck of U-Net. This strategic placement aims to improve overall contextual understanding, addressing the typical limitations of convolutions in capturing global context while also minimizing the increase in parameters. We will refer to this model as *Unidirectional* (Fig. 2).

Dealing with Multi-directionality. In contrast to the self-attention mechanism of Transformers, where each token can gather information from every other token in the sequence, Mamba restricts each token to infer only information from the current state. This results in an approximation of the past tokens only and makes it sensitive to the sequence order. This means that when Mamba is employed for image segmentation tasks, the very first pixels (or voxels) in the sequence do not have any context awareness.

Taking this into consideration, our second proposal consists of integrating two Mamba layers into a unified module called the bidirectional 3D Mamba layer. This module flattens the spatial dimensions and manages the sequence bidirectionally by feeding one of the two layers with the sequence in the backward direction. Subsequently, the output of the latter layer is reversed to its original order and then summed with the output of the former. Finally, the sum is normalized and reshaped back into the original 3D shape (Fig. 1b).

Integrating U-Net architecture with such a bidirectional Mamba layer in place of its unidirectional version results in the model that we call *Bidirectional*. This variant aims to overcome the limitations of unidirectional context awareness by allowing each token to access information from both past

Table 1: Left: 5-fold cross-validation results on ACDC, Synapse Abdomen, and BrainTumor datasets. Our proposals are marked with *. The best results are in **bold** while the second best are underlined. Right: Computational comparison on the Synapse dataset. Our proposals are marked with *. The parameters are expressed in millions [M] and VRAM in gigabytes [GB]. Training and inference times, expressed in hours [h] and seconds [s], respectively, are obtained on an Nvidia A100 with 80GB of memory. All competitor models were trained for 1000 epochs, as recommended by most of their original papers, while our method achieved convergence in only 300 epochs. Inference times are the average across all test volumes.

	Model	ACDC	Brain Tumor		Synapse Abdomen						
		DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	Params	GFLOPs	VRAM	Training	Inference
CNNs	nnU-Net [7]	91.42	4.53	85.74	10.91	86.21	30.64	410.11	7.65	9.20	21.80
	nnU-Net ResEnc [7]	90.84	4.12	85.60	7.70	86.61	57.50	502.49	10.00	10.00	22.20
	MedNeXt-M-K3 [30]	91.64	6.35	85.27	18.99	85.70	32.65	248.03	15.32	67.60	153.60
	MedNeXt-M-K5 [30]	90.70	6.67	84.79	17.30	86.00	34.75	308.01	18.85	218.30	416.90
Transformers	TransU-Net [12]	89.75	13.18	64.14	32.27	77.24	96.07	88.91	16.25	26.50	73.90
	CoTr [31]	90.90	9.96	68.21	9.35	84.67	50.12	369.22	8.10	18.60	41.40
	UNETR [13]	88.72	9.04	70.92	19.15	78.10	92.49	75.76	15.29	15.40	39.50
	Swin-UNETR [14]	91.36	6.77	84.07	11.02	83.64	62.83	384.20	13.91	22.00	38.70
	nnFormer [9]	91.87	4.05	86.34	11.14	86.56	150.50	213.41	9.73	8.20	20.60
Mamba	UMambaBot [27]	90.44	**3.80**	86.35	7.35	86.88	41.95	156.32	13.55	22.00	54.20
	UMambaEnc [27]	90.07	4.17	86.16	7.83	87.82	42.85	231.18	26.42	37.90	89.30
	Ours (*Unidirectional*)*	91.33	3.82	86.66	7.91	87.48	61.49	480.90	25.61	12.70	99.60
	Ours (*Bidirectional*)*	91.50	3.85	85.75	5.99	88.29	64.75	494.17	27.31	16.50	134.10
	Ours (*Multidirectional*)*	**92.04**	3.84	**86.70**	**5.98**	**88.93**	68.46	527.56	36.92	18.20	149.00

and future positions within the sequence, ensuring a more comprehensive context at every spatial position.

However, in 3D segmentation, spatial orientation spans over three axes. By applying the bidirectional Mamba layer, we are limiting the context integration across multiple directions, which is essential for each voxel to use spatial information in all orientations. As an example, if we were to consider only a single flattened sequence, such as `(H, W, D)`, the distance between the first token at index `(0, 0, 0)` and the token at index `(0, 0, 1)` would be $H * W$ instead of 1. Typically, the values of H and W are in the order of 10^2, resulting in a total distance of 10^4. A better solution would instead be to process all the six possible permutations of the three spatial dimensions `(H, W, D)` of a 3D volume, resulting in a total of 12 sequences when accounting for both forward and backward directions. To meet memory constraints, only four of the possible directions have been considered in our experiments, i.e., `(H, W, D)`, `(H, D, W)`, `(W, H, D)`, and `(D, W, H)`. Incorporating multiple directions maintains linear complexity while enhancing spatial awareness. To aggregate the output sequences of all the modules involved, we stack each sequence on a new axis and compute the mean value across it. This module replaces the bidirectional 3D Mamba layer, resulting in a new architecture we refer to as *Multidirectional*.

Implementation Details. All of our models have been trained for 300 epochs using RAdam, a learning rate of 0.0003, and a linear learning rate scheduler. Mamba blocks have been initialized as proposed in the original Mamba publication [18]. Training has been performed on an A100 Nvidia GPU using CUDA 11.8 and PyTorch 2.1.2.

3. EXPERIMENTAL RESULTS

Datasets. Following the literature on medical image segmentation [9, 13, 31], experiments have been carried out on three different well-known datasets: MSD BrainTumour [32], Synapse Multi-organ [33], and ACDC [34]. The first includes 484 MRI images, which have been split for training and testing according to [13]. The second dataset, Synapse, includes 3,779 axial contrast-enhanced abdominal CT images from 30 scans. Following the split proposed by [12], our experiments employ 18 cases for training and 12 for testing. Finally, ACDC comprises 100 heart MRI scans, again split for training and testing according to [12].

Metrics & Baselines. In our experiments, we include both HD95 and Dice score metrics, commonly employed in medical image segmentation tasks.

Comparison has been performed on recently proposed methods for medical image segmentation, considering CNN, Transformers, and Mamba-based architectures. Our analysis includes the original nnU-Net [7] configuration making use of the vanilla U-Net architecture (nnU-Net), and its variations based on U-Net with residual connections in the encoder (nnU-Net ResEnc). Furthermore, the Transformer-inspired CNN modification based on ConvNeXt blocks, MedNeXt [30], has been considered in its two variations K3, and K5. For what concerns Tranformer-based architectures, we compare our proposals with TransU-Net [12], CoTr [31], a hybrid architecture combining convolutional and Transformer modules, UNETR [13], Swin-UNETR [14], and the recently published nnFormer [9]. Finally, we include UMamba [27] in its two variations UMambaBot and UMambaEnc.

In our experiments, a standardized scheme for hyperparameter configuration has been adopted. Whenever available, the capabilities of the self-configuration method are employed. Otherwise, we opted for the default configuration (if any) or the one closest to the respective dataset, reducing the learning rate until convergence. Models are trained from scratch without any pre-training data, except for the TransU-Net model, which is pre-trained as recommended in its original paper [12]. The nnU-Net five-fold cross-validation schema has always been employed.

Results. As shown in Tab. 1, our proposed models consistently outperform all competing approaches, demonstrating superior overall performance across all the considered datasets and metrics. Among the models evaluated, our proposed *Multidirectional* consistently outperforms all the others across most of the experimental settings. Notably, excluding nnFormer, our Mamba-based architectures achieve improvements of more than 3 Dice coefficient points compared to the best performing Transformer-based architectures and up to 1 Dice point over nnU-Net.

Remarkably, our proposals surpass existing Mamba-based models in the literature, demonstrating that addressing the initial hidden state problem by employing multidirectional Mamba layers enhances segmentation performances.

Finally, a comprehensive computational comparison is reported considering the number of parameters, GFLOPs, and GPU memory on the Synapse dataset. Our proposed models have a higher number of parameters compared to CNN approaches, while they are comparable to or often have fewer parameters than Transformer-based models. More specifically, the number of parameters of our models ($\sim$60M) are, on average, the double with respect to nnU-Net ($\sim$31M), comparable to those of nnU-Net ResEnc ($\sim$57M), and much lower than those of Transformer-based models (from $\sim$95M of TransU-Net and UNETR, up to 150M of nnFormer).

4. CONCLUSION

This paper aims to assess the efficacy of the Mamba State Space Model for 3D medical image segmentation, comparing it with advanced convolutional and Transformer-based architectures. In addition, we propose alternative designs for Mamba architectures to address their key limitations. Specifically, we integrate Mamba at various stages within the standard U-Net framework, using both single-directional, bi-directional, and multi-directional implementations. The overall framework blends convolutions and state-space models, leveraging the former to encode precise spatial information while addressing the latter to model long-range voxel-level interactions. Mamba offers a dual advantage, providing a global context alongside voxel-wise precision, the former absent in traditional convolutional layers due to limited receptive fields and the latter absent in Transformers due to their computational complexity.

Our experimental results highlight the substantial improvement in HD95 and DSC metrics on three well-known datasets compared to nnU-Net and different Transformer-based networks. We showcase Mamba versatility by adapting it from its original use in text generation and large language models to achieve state-of-the-art results in a completely different task. This adaptability highlights Mamba potential beyond its initial design, demonstrating its efficacy on image encoding and segmentation.

5. ACKNOWLEDGMENTS

This project has received funding from the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2024 funds (Fondo di Ateneo per la Ricerca), and from the European Union's Horizon 2020 research and innovation programme under GA No. 965193.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [32–34]. Ethical approval was not required, as confirmed by the license attached with the open-access data.

7. REFERENCES

[1] Asgari Taghanaki et al., "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, 2021.

[2] Lecun Yann et al., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 1998.

[3] Ronneberger Olaf et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[4] Foivos I. Diakogiannis et al., "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, 2020.

[5] Sijing Cai et al., "Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 6, 2020.

[6] Milletari Fausto et al., "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016.

[7] Isensee Fabian et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, 2021.

[8] Vaswani Ashish et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] Hong-Yu Zhou et al., "nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer," *IEEE Transactions on Image Processing*, 2023.

[10] Luca Lumetti et al., "Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal," *IEEE Access*, 2024.

[11] Luca Lumetti et al., "Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs," in *International Conference on Pattern Recognition*, 2024.

[12] Jieneng Chen et al., "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, 2024.

[13] Ali Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[14] Ali Hatamizadeh et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *International MICCAI Brainlesion Workshop*, 2022.

[15] Sinong Wang et al., "Linformer: Self-Attention with Linear Complexity," *ArXiv*, vol. abs/2006.04768, 2020.

[16] Krzysztof Marcin Choromanski et al., "Rethinking Attention with Performers," in *International Conference on Learning Representations*, 2021.

[17] Iz Beltagy et al., "Longformer: The Long-Document Transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[18] Albert Gu et al., "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[19] Ben Wang et al., "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," `https://github.com/kingoflolz/mesh-transformer-jax`, 2021.

[20] Stella Biderman et al., "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling," in *International Conference on Machine Learning*, 2023.

[21] Yue Liu et al., "VMamba: Visual State Space Model," *arXiv preprint arXiv:2401.10166*, 2024.

[22] Haifan Gong et al., "nnMamba: 3D Biomedical Image Segmentation, Classification and Landmark Detection with State Space Model," *arXiv preprint arXiv:2402.03526*, 2024.

[23] Jiacheng Ruan et al., "VM-UNet: Vision Mamba UNet for Medical Image Segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[24] Zhaohu Xing et al., "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[25] Hanwei Zhang et al., "A Survey on Visual Mamba," *Applied Sciences*, vol. 14, no. 13, 2024.

[26] Ziyang Wang et al., "Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation," *arXiv preprint arXiv:2402.05079*, 2024.

[27] Jun Ma et al., "U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[28] Albert Gu et al., "HiPPO: Recurrent Memory with Optimal Polynomial Projections," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[29] Alexey Dosovitskiy and other, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] Saikat Roy et al., "MedNeXt: Transformer-driven Scaling of ConvNets for Medical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[31] Yutong Xie et al., "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2021.

[32] Michela Antonelli et al., "The Medical Segmentation Decathlon," *Nature Communications*, vol. 13, no. 1, 2022.

[33] Bennett Landman et al., "Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge," in *Multi-Atlas Labeling Beyond Cranial Vault — Workshop Challenge*, 2015, vol. 5.

[34] Olivier Bernard et al., "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, 2018.