IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 April 2025, accepted 29 April 2025, date of publication 15 May 2025, date of current version 28 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3570461

RESEARCH ARTICLE

Taming Mambas for 3D Medical Image Segmentation

LUCA LUMETTI[1,*], VITTORIO PIPOLI[1,2,*], KEVIN MARCHESINI[1,*], ELISA FICARRA[1], COSTANTINO GRANA[1], (Member, IEEE), AND FEDERICO BOLELLI[1], (Associate Member, IEEE)

[1]Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, 41125 Modena, Italy
[2]Department of Computer Science, University of Pisa, 56126 Pisa, Italy

Corresponding author: Federico Bolelli (federico.bolelli@unimore.it)

This work was supported in part by the University of Modena and Reggio Emilia and Fondazione di Modena, through FAR 2023 and Fondo di Ateneo per la Ricerca Dipartimentale (FARD)-2024 funds (Fondo di Ateneo per la Ricerca), by Italian Ministry of Research, under the Complementary Actions to National Recovery and Resilience Plan (NRRP) "Fit4MedRob—Fit for Medical Robotics" under Grant PNC0000007; and in part by the Progetti di Rilevante Interesse Nazionale (PRIN) 2022 "AIDA: explAinable multImodal Deep learning for personAlized oncology" under Grant 20228MZFAA.

*Luca Lumetti, Vittorio Pipoli, and Kevin Marchesini contributed equally to this work.

ABSTRACT Recently, the field of 3D medical segmentation has been dominated by deep learning models employing Convolutional Neural Networks (CNNs) and Transformer-based architectures, each with its distinctive strengths and limitations. CNNs are constrained by a local receptive field, whereas Transformers are hindered by their substantial memory requirements as well as their data hunger, making them not ideal for processing 3D medical volumes at a fine-grained level. For these reasons, fully convolutional neural networks, as nnU-Net, still dominate the scene when segmenting medical structures in large 3D medical volumes. Despite numerous advancements toward developing Transformer variants with subquadratic time and memory complexity, these models still fall short in content-based reasoning. A recent breakthrough is Mamba, a Recurrent Neural Network (RNN) based on State-Space Models (SSMs), outperforming Transformers in many long-context tasks (million-length sequences) on famous natural language processing and genomic benchmarks while keeping a linear complexity. In this paper, we evaluate the effectiveness of Mamba-based architectures in comparison to state-of-the-art convolutional and Transformer-based models for 3D medical image segmentation across three well-established datasets: Synapse Abdomen, MSD BrainTumor, and ACDC. Additionally, we address the primary limitations of existing Mamba-based architectures by proposing alternative architectural designs, hence improving segmentation performances. The source code is publicly available to ensure reproducibility and facilitate further research: https://github.com/LucaLumetti/TamingMambas

INDEX TERMS Medical imaging, 3D segmentation, Mamba, U-Net, transformers, RNNs.

I. INTRODUCTION

Image segmentation is crucial in the analysis of medical images, typically serving as the preliminary step for examining anatomical structures and surgical planning [1], [2], [3], [4]. During recent years, Convolutional Neural Networks (CNN) [5] and, in particular, U-shaped Fully Convolutional Neural Networks (FCNN) have garnered widespread acceptance within the research community [6]. Their success can be attributed to their local receptive field, which allows them to capture substantial contextual information while maintaining relatively low GPU memory consumption. Additionally, their ability to achieve competitive performance with limited training data has contributed to their widespread adoption. Despite their effectiveness, after the outbreak of Vision Transformer [7], FCNN has been replaced by hybrid architectures, made up of both Convolutional and Multi-Head Attention layers [8]. This architectural modification is designed to alleviate the limitations imposed by the local receptive field of FCNN by leveraging the

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

Multi-Head Attention mechanism of Transformers, which, in contrast, offers remarkable capabilities for modeling long-range contextual information. Several attempts have been made in the literature to integrate transformer-based architectures into the classic U-Net [9], [10], [11], [12], [13], [14], [15]. Even if these methods led to improvement in performance, they came at the cost of the quadratic memory footprint of the attention mechanism, which, alongside their data hungriness, makes these approaches not ideal when applied to 3D volumes. Under conditions of data scarcity, it is common for transformer-based architectures to underperform FCNN.

In this regard, the latest research put a lot of effort into the reduction of the computational cost of the transformer architecture, proposing subquadratic attention mechanism [16], [17], linear attention mechanism [18], [19], [20], and gating [21]. However, most of them fall short when it comes to context modeling, in particular when the context length is considerably high. Recently, the field of sequence modeling has been greatly influenced by an innovative architecture based on the State-Space Model (SSM) [22] known as Mamba [23]. Notably, as opposed to previous SSMs, Mamba incorporates an innovative selection mechanism that dynamically filters input data in an input-dependent manner, thereby excluding irrelevant information while retaining relevant information indefinitely. Moreover, Mamba mitigates a common challenge in SSM–the exponential scaling of gradients–by leveraging the principles of the HiPPO theory [24]. As a consequence, Mamba has shown state-of-the-art capabilities in several Natural Language Processing (NLP) and genomic tasks outperforming Transformer, improving the modeling of big context up to the order of a million tokens, making it a suitable candidate to efficiently process also 3D volumes, where the number of tokens reaches the same order of magnitude.

Thus, Mamba provides a potential solution to the challenges posed by FCNNs and Vision Transformers, offering an architecture that models a larger receptive field than FCNNs while maintaining linear complexity. This results in a more resource-efficient design relative to transformer-based models, particularly in long-context reasoning tasks, in which Mamba-based architectures consistently outperform Transformers. Consequently, Mamba represents a practical compromise that combines extensive contextual modeling with efficient resource utilization, making it well-suited for advancing 3D medical image segmentation.

In summary, this paper aims to investigate the effectiveness of Mamba for 3D image segmentation by comparing state-of-the-art Mamba-based architectures with convolutional and Transformer-based segmentation models. Additionally, we seek to address the primary limitations of current Mamba-based architectures by proposing various strategies for integrating Mamba within a U-Net-based architecture. Specifically, we examine the impact of modeling directionality on one or more axes and explore the use of Mamba as a selective copying mechanism in skip connections. To perform our experimental evaluation, we employ three different well-known datasets, MSD BrainTumor [25], Synapse Multi-organ [4], and ACDC [26].

II. RELATED WORK

Convolutional Neural Networks (CNNs) [5] have been the dominant solution for both 2D and 3D medical image segmentation for years. Among these, U-Net [6], characterized by its U-shaped symmetric encoder-decoder structure with skip connections, represents an effective architecture that subsequent models have continued to adopt until the present day. Following U-Net, several variants have been introduced, including Res-U-Net [27], Dense-U-Net [28], V-Net [29], 3D U-Net and its state-of-the-art ecosystem nnU-Net [30], each proposing enhancements to the original framework. Despite their advancements, CNNs inherently face limitations in capturing global patterns due to the locality of the convolutional operator. In response, significant research efforts have been directed towards integrating the attention mechanisms of Transformers [8] with U-Net-based architectures. This integration aims to leverage both local and global dependencies, as evidenced by models such as MedFormer [9], TransUNet [10], Swin-UNet [11], UNETR [12], and Swin-UNETR [13]. However, the attention mechanism's quadratic complexity forces the imposition of constraints, such as window-based or axial-based attention, to mitigate computational demands. While various studies have attempted to reduce this complexity [18], [19], [20], [31], none have matched the performance of traditional attention mechanisms in long-context modeling.

Recent developments have introduced a novel architecture, Mamba [23], predicated on state-space modeling [22], [32], which promises capabilities for long-context content-based reasoning with linear-time complexity. Mamba has demonstrated superior performance over state-of-the-art transformer models, such as Pythia-6.9B [33], GPT-J-6B [34], OPT-6.7B [35], Hyena [36], in tasks requiring long-context content-based reasoning, such as natural language processing and genomic analyses, with inputs of up to a million-length scales.

Due to their effectiveness and versatility, Mamba-based architectures have been rapidly adapted to various domains, including Computer Vision [37]. In addition, given that segmenting 3D volumes can be seen as processing sequences composed of millions of voxels, several researchers have devoted significant efforts to adapting the Mamba architecture for both 2D and 3D segmentation, yielding promising results [38], [39], [40], [41], [42], [43]. Among the various contributions, UMamba remains one of the most significant in the field, given the model's ability to adapt effectively to new datasets without the need for extensive hyperparameter tuning. In particular, in [44], the authors propose two architectures, namely UMamba Enc and UMamba Bot, both inheriting their core structure from U-Net and harnessing Mamba-based layers. The former integrates the Mamba

layers in the encoder part of the architecture, while the latter integrates a single Mamba layer in the bottleneck. Despite their effectiveness, the authors did not focus on the *directionality problem* that derives from employing a recurrent network to extract patterns from data that has more than one spatial dimension. Indeed, once a 3D volume is flattened into a sequence, each voxel is assigned a position within the sequence. This results in the model being able to analyze the latter elements of the sequence by leveraging information from the preceding part. However, it lacks contextual information when processing the initial part of the sequence.

The aforementioned advancements in the field motivated us to devise Mamba architectures for 3D image segmentation, paying attention to the directionality.

III. METHOD

In this section, all the theoretical concepts related to the vanilla Mamba architecture (a set of stacked Mamba blocks) are introduced. Then, we thoroughly explain how Mamba blocks can be employed to extract patterns from 3D volumes and illustrate approaches to integrate such blocks into a U-Net architecture for 3D medical imaging segmentation.

A. PRELIMINARIES

A State-Space Model (SSM) is a mathematical representation of a dynamic system that maps a 1D input $x(t) \in \mathbb{R}$ to a ND latent state $h(t) \in \mathbb{R}^N$ before projecting it to a 1D output signal $y(t) \in \mathbb{R}$. This system uses $A \in \mathbb{R}^{N\times N}$ as the evolution parameter, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ as the projection parameters:

$$h'(t) = Ah(t) + Bx(t) \tag{1}$$
$$y(t) = Ch(t) \tag{2}$$

Together, the previous equations aim to predict the state of a system from observed data. Since the input is expected to be continuous, the main representation of the SSM is a continuous-time representation.

To employ previous equations in a real-world scenario, and more specifically into a neural network, a discretization of the variable t is required and can be achieved by introducing a step-size parameter Δ and a discretization rule, which in this case is the *zero-order hold*:

$$\overline{A} = \exp(\Delta A) \tag{3}$$
$$\overline{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B \tag{4}$$

This leads to the following discrete state-space model that can be computed in a recurrent fashion:

$$h_{t+1} = \overline{A}h_t + \overline{B}x_t \tag{5}$$
$$y_t = Ch_t \tag{6}$$

This basic SSM performs very poorly in practice due to gradients scaling exponentially in the sequence length. To address this issue, Mamba proposes two key elements: imposing a structure to the matrix A, using the HiPPO theory [24], and including a selection mechanism, i.e., making the parameters B, C, and Δ input-dependent through a linear projection:

$$B = \text{Linear}_N(x)$$
$$C = \text{Linear}_N(x) \tag{7}$$
$$\Delta = \text{SoftPlus}(\text{Parameter} + \text{Broadcast}_D(\text{Linear}_1(x))$$

Such a formulation, together with an efficient implementation of the process by means of a selective scan algorithm that allows the model to filter out irrelevant information, constitutes the so-called S6 model [23].

The original Mamba publication [23] introduces a Mamba block, which is depicted in gray in Fig. 1a. This block comprises an initial residual connection, followed by a layer normalization step in the main flow. Subsequently, the information is split into two branches, each incorporating a linear layer. The left branch includes a one-dimensional convolutional layer followed by a SiLU activation and an S6 model, whereas the tensors flowing on the right branch are processed solely with a SiLU activation. The outputs from these branches are then multiplied element-wise, after which a final linear layer is applied. The output of this linear layer is subsequently summed with the initial residual connection to yield the final block output.

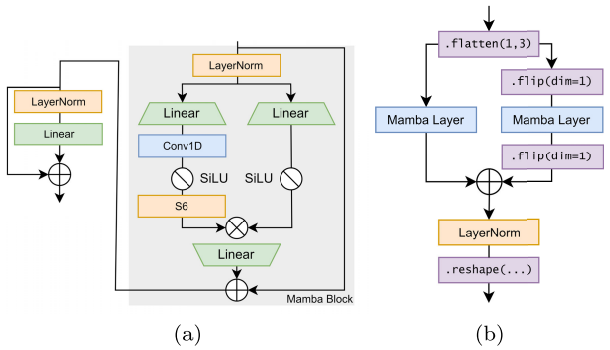


FIGURE 1. From left to right: (a) The unidirectional Mamba layer, which processes input sequences only in the forward direction. The layers within the gray square collectively form the Mamba Block. (b) The bidirectional Mamba layer, consisting of two unidirectional Mamba layers: the left branch processes the forward sequences, while the right branch processes the reversed sequences.

B. VISION MAMBA

Mamba is a sequence-to-sequence model; thus, it is only able to handle 1D sequences. In order to apply it to 2D images and 3D volumes, a 1D sequence flattening of pixels (or voxels) is required. Different from the approach adopted in Vision Transformer, where the quadratic cost of self-attention with respect to the number of pixels prevents their scaling to "realistic" input size and requires extracting patches to reduce the input spatial dimension, Mamba allows us a linear-time sequence modeling of the input, preventing any sampling. Patch down-sampling is a major issue in medical image segmentation, due to the need for voxel-wise details, which is usually enforced by the large medical input data.

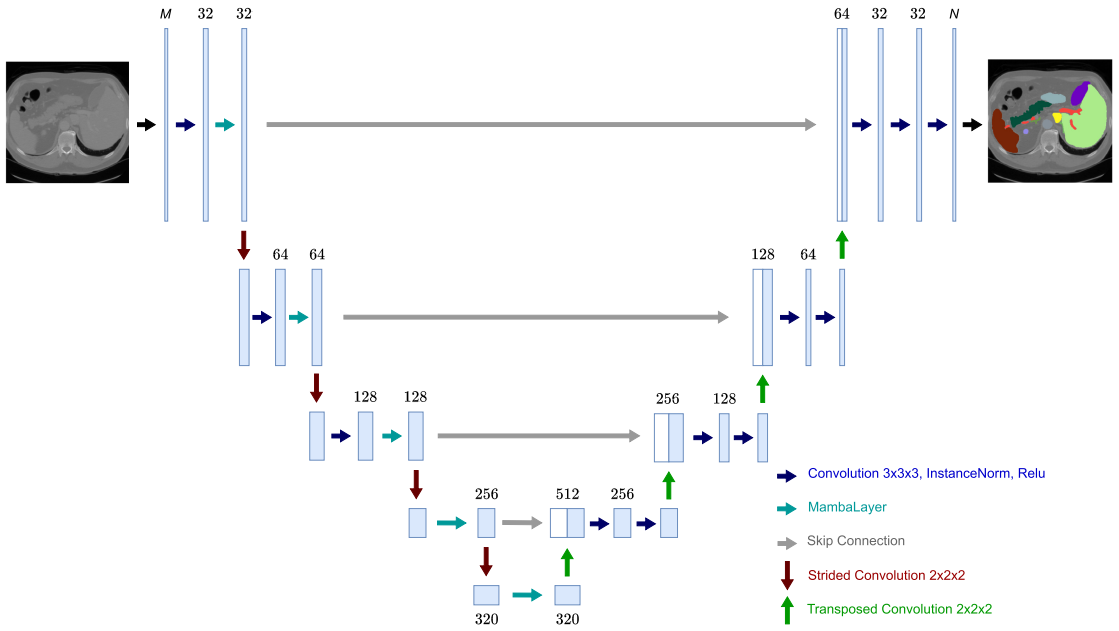


FIGURE 2. U-Net Architecture integrating our proposed Mamba Layers. By properly selecting the Mamba Layers (turquoise arrows), SegMamba, BiSegMamba, and MultiSegMamba are obtained. To obtain SegMambaSkip, the currently displayed Mamba Layers (turquoise arrows) must be replaced by the standard U-Net convolution (blue arrow), and corresponding Mamba Layers must be placed within the skip connections (gray arrows).

One downside of Mamba is that it is not permutation-invariant. In contrast to the transformer self-attention mechanism, 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, resulting in an approximation of the past tokens only. 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. To mitigate the issue, we devised Mamba layers capable of processing tensors along different spatial directions.

1) OUR MAMBA LAYER

Instead of directly including the Mamba Block into our U-Net architecture, by taking inspiration from the ViT architecture [7], we developed a wrapper. The wrapping, consisting of an additional LayerNorm and an MLP head followed by a skip connection, allows us to improve Mamba stability. We denote this module as the Mamba Layer, illustrated in Fig. 1a. Subsequently, we integrated two instances of the Mamba Layer into a unified module. This module, named Bidirectional 3D Mamba Layer, takes as input a 3D volume with dimensions `(B, H, W, D, C)`. It 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 from this layer is reversed to its original order and then summed token by token with the output of the "straight" layer. Finally, the sum is normalized and reshaped back into a 3D volume (Fig. 1b).

The strategies we introduce to integrate the Mamba Layer into nnU-Net are detailed in the following and depicted in Fig. 2.

2) SEGMAMBA

Our initially proposed integration involves the inclusion of a singular (unidirectional) Mamba Layer preceding each pooling convolution and the bottleneck of U-Net. This strategic placement is designed to enhance the overall contextual understanding, addressing the inherent limitations in the global context that convolutions often encounter while limiting the number of additional parameters.

3) SEGMAMBASKIP

One of the universally recognized strength points of the U-Net architecture lies in its skip connections [6], which allow the decoding part of the network to access fine-grained details coming from the encoder. Indeed, as the network compresses the image to extract high-level features, it loses some fine-grained details. Skip connections help by copying the detailed feature maps from earlier layers (before compression) and combining them with the layers that are reconstructing the image. This way, the network gets both the detailed, low-level information and the high-level understanding, which helps produce more accurate and sharper segmentation results. Meanwhile, Mamba has been devised to efficiently select data in an input-dependent manner, thus being capable of filtering out irrelevant information. Hence, we augment the skip connections in the U-Net architecture by inserting an additional Bidirectional 3D Mamba Layer before concatenating the activation map to the corresponding decoder output. The Mamba layer introduced in the skip connection is intended to enrich the information flowing from the encoder to the decoder in the skip branches. What is noteworthy here is that Mamba processing does

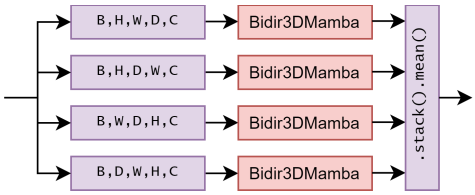


FIGURE 3. To achieve multi-directionality, four Bidirectional 3D Mamba Layers are employed, corresponding to four out of the six possible permutations of the triplet `(H, W, D)`. The outputs of each layer is stacked, and the mean per token is computed.

not degrade spatial information but instead enriches it with contextual awareness. Indeed, the design of the Mamba layer includes a residual connection (right branch of Fig. 1a), meaning that it preserves spatial information while learning to enhance feature quality. The residual connection ensures the original encoder features, i.e., the fine-grained details, remain intact by providing a direct pathway alongside the Mamba layer. As said, rather than replacing features, the Mamba layer placed on the skip connection is meant to enrich the features passed through the skip connections with long-range dependencies or contextual patterns without disrupting spatial fidelity.

4) PANSEGMAMBA

It includes our Bidirectional 3D Mamba Layer before each downsampling step as well as in the U-Net bottleneck. BiSegMamba processes data in both forward and backward directions, ensuring that each voxel receives context from all other voxels in at least one of the two directions. This means that voxels at the start of a sequence in the forward direction (which initially have limited context) are processed with extensive context in the backward direction, and vice versa. By leveraging both directions of a single sequential arrangement, in BiSegMamba we strike a balance between computational efficiency and model effectiveness. The Bidirectional 3D Mamba Layer enables the model to effectively weigh the importance of tokens across various spatial dimensions, without the need to consider all possible permutations. This approach is particularly beneficial when dealing with distant dependencies and selective information processing, enhancing the ability to discern relevant features during downsampling and in bottleneck layers.

5) MULTISEGMAMBA

With this variation, we propose to process all conceivable sequential arrangements of the volume, resulting in a total of six possible permutations for the three spatial dimensions `(H, W, D)` of a 3D volume. This yields a total of 12 distinct sequences, accounting for both the forward and backward directions of the six permutations. The rationale behind seeking multiple directions stems from the necessity for each voxel to exploit spatial information in all conceivable orientations. If we were to consider only a single sequence, such as `(H, W, D).flatten()`, 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. Due to memory constraints, we only encompass 4 out of 6 possible directions,[1] chosen specifically to ensure that each spatial dimension is similarly represented, with each direction appearing at least once at the beginning of the flattened representation and at least once at the end. This design is intended to explore different spatial relationships, mitigating potential biases introduced by fixed directional ordering. While alternative order combinations are available, preliminary experiments showed negligible differences. By incorporating multiple directions, we maintain linear complexity while affording each token superior spatial awareness. This approach ensures that neighboring tokens are indeed proximate in the obtained representation, enhancing the overall spatial awareness of the model. 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 (Fig. 3). This module substitutes the Bidirectional 3D Mamba Layer in BiSegMamba.

TABLE 1. Configuration of the proposed models when trained on the selected datasets.

	BrainTumour	Synapse	ACDC
Spacing	[1, 1, 1]	[3, 0.76, 0.76]	[6.35, 1.52, 1.52]
Median shape	$138 \times 170 \times 138$	$148 \times 512 \times 512$	$13 \times 246 \times 213$
Crop size	$128 \times 128 \times 128$	$48 \times 192 \times 192$	$14 \times 256 \times 224$
Batch size	2	2	4

C. IMPLEMENTATION DETAILS

Details regarding patch shape, batch size, and other pipeline settings are reported in Tab. 1. All of our models have been trained for 300 epochs using RAdam, a learning rate of 0.0003, and a linear learning rate scheduler. For the parameters initialization of the Mamba layers, we scale the weights of residual layers at initialization by a factor of $1/\sqrt{N}$, where N is the number of residual layers. This is the same as in the GPT-2 paper and employed in the Mamba source code. SegMamba, BiSegMamba, and MultiSegMamba variants include 5 Mamba layers, matching the encoder depth, while SegMambaSkip uses 4 layers, aligned with the number of skip connections. The inner dimension of the State-Space Model (i.e., the size of the evolution parameter A, Sec. III-A) is defined as $\min(C, 256)$, where C is the number of tokens' channels in the input sequence, which can be seen in Fig. 2, i.e., 32, 64, 128, 256. Training has been performed on an A100 Nvidia GPU using CUDA 11.8 and PyTorch 2.1.2.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

We conducted experiments using three different well-known datasets: MSD BrainTumor [25], Synapse Multi-organ [4],

[1] In our experiments we employ the following directions: `(H, W, D)`, `(H, D, W)`, `(W, D, H)`, and `(D, W, H)`.

TABLE 2. 5-fold cross-validation results on the BrainTumor dataset. Our proposals are marked with †. Standard deviations for the average scores over the 5 folds are reported. Best results are in **bold** while the second best are underlined. Methods subjected to a one-sided paired samples t-test comparing our best method against the best of the alternatives are highlighted in blue. If the p-value associated with the test is less than 0.05, the result is indicated as statistically significant by ⋆. Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC) scores are reported, alongside the average.

	Model	Average		WT		ET		TC	
		HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑
CNNs	nnU-Net [30]	4.53±0.17	85.74±0.93	4.21	91.15	3.89	80.76	5.47	85.29
	nnU-NetResEnc [30]	4.12±0.16	85.60±0.70	3.71	89.93	3.72	80.86	4.93	86.01
	MedNeXt-M-K3 [49]	6.35±0.21	85.27±0.45	4.59	90.84	6.57	80.88	7.89	84.10
	MedNeXt-M-K5 [49]	6.67±0.30	84.79±0.81	4.93	88.93	6.75	79.97	8.33	85.47
Transformers	TransU-Net [10]	13.18±0.83	64.14±0.84	14.42	70.16	10.80	54.31	14.31	67.94
	TransBTS [50]	9.83±0.22	69.72±0.56	10.32	78.22	10.20	57.26	8.97	73.68
	CoTr [48]	9.96±0.19	68.21±0.42	9.25	74.81	9.58	55.14	11.04	74.67
	UNETR [12]	9.04±0.41	70.92±1.02	8.03	78.85	9.83	58.06	9.25	75.85
	SwinU-Net [11]	9.98±0.33	67.95±0.57	8.85	76.43	10.31	57.21	10.77	70.20
	Swin-UNETR [13]	6.77±0.25	84.07±0.67	7.13	88.92	7.54	79.93	5.63	83.34
	LeViT-UNet-384s [47]	8.56±0.30	70.06±0.87	8.20	77.06	8.60	58.10	8.89	75.03
	MISSFormer [51]	9.21±0.19	83.08±0.39	8.40	88.21	9.57	79.71	9.64	81.33
	nnFormer [45]	4.05±0.25	86.34±0.51	3.47	91.28	4.24	**81.76**	4.46	85.97
Mamba	UMambaBot [44]	**3.80±0.22**	86.35±0.21	3.49	92.10	3.80	80.04	4.10	86.90
	UMambaEnc [44]	4.17±0.18	86.16±0.53	3.63	92.30	4.44	79.72	4.43	86.46
	SegMambaSkip†	4.53±0.20	85.25±0.81	3.61	92.11	5.43	78.85	4.54	84.79
	SegMamba†	3.82±0.11	86.66±0.45	3.66	92.26	3.83	80.77	3.96	86.96
	BiSegMamba†	3.85±0.16	85.75±0.60	**3.38**	**92.43**	**3.46**	79.60	4.70	85.21
	MultiSegMamba†	3.84±0.24	**86.70±0.17**⋆	3.72	92.09	3.88	80.84	**3.93**	**87.18**

TABLE 3. 5-fold cross-validation results on the Synapse Abdomen dataset. Our proposals are marked with †. For space constraints, single class results only report the Dice score. Standard deviations for the average scores over the 5 folds are reported. Best results are in **bold** while the second best are underlined. Methods subjected to a one-sided paired samples t-test comparing our best method against the best of the alternatives are highlighted in blue. If the p-value associated with the test is less than 0.05, the result is indicated as statistically significant by ⋆.

	Model	Average		Aorta	Gallb.	L.Kidn.	R.Kidn.	Liver	Pancr.	Spleen	Stom.
		HD95↓	DSC↑								
CNNs	nnU-Net [30]	10.91±0.69	86.21±1.19	91.65	70.01	86.67	85.75	96.11	83.22	90.69	85.55
	nnU-NetResEnc [30]	7.70±0.42	86.61±1.07	89.94	64.20	90.79	91.18	**97.36**	79.48	92.03	**87.93**
	MedNeXt-M-K3 [49]	18.99±0.53	85.70±0.77	92.44	**72.75**	87.62	86.21	97.15	81.17	90.30	77.93
	MedNeXt-M-K5 [49]	17.30±0.60	86.00±0.86	92.15	71.66	87.89	87.43	96.91	80.26	90.95	80.78
Transformers	TransU-Net [10]	32.27±1.01	77.24±0.91	86.88	62.59	81.35	76.98	94.45	55.57	84.97	75.12
	TransBTS [50]	11.98±0.67	83.27±1.06	91.95	62.24	86.91	87.15	96.67	71.91	91.62	77.70
	CoTr [48]	9.35±0.39	84.67±0.75	**92.77**	63.07	87.98	86.84	92.75	78.63	94.54	80.76
	UNETR [12]	19.15±0.84	78.10±1.12	89.75	55.81	85.71	84.71	94.00	60.23	84.47	70.14
	SwinU-Net [11]	22.02±0.70	79.06±0.73	85.65	66.46	83.03	79.37	94.02	56.57	90.67	76.72
	Swin-UNETR [13]	11.02±0.72	83.64±1.31	91.22	66.48	87.09	86.62	95.99	68.79	95.72	77.19
	LeViT-UNet-384s [47]	16.80±0.81	78.38±0.99	87.52	61.77	84.04	79.87	92.80	59.20	88.84	73.03
	MISSFormer [51]	18.50±0.59	81.87±0.85	86.48	68.92	85.56	81.60	94.24	65.44	91.70	80.99
	nnFormer [45]	11.14±0.48	86.56±0.64	91.63	69.85	86.61	86.55	96.97	**83.68**	90.72	86.44
Mamba	UMambaBot [44]	7.35±0.42	86.88±0.80	89.88	60.14	89.99	94.37	96.81	82.33	95.66	85.88
	UMambaEnc [44]	7.83±0.50	87.82±0.75	89.57	65.20	89.46	94.84	96.97	83.35	**96.80**	86.40
	SegMambaSkip†	6.29±0.47	88.26±0.89	89.64	69.04	93.40	**94.91**	96.80	79.61	96.45	86.19
	SegMamba†	7.91±0.38	87.48±0.77	89.59	62.21	93.65	94.81	96.82	80.72	95.22	86.85
	BiSegMamba†	5.99±0.53	88.29±0.90	91.02	70.12	92.98	94.32	96.94	79.08	96.26	85.58
	MultiSegMamba†	**5.98±0.36**⋆	**88.93±0.21**⋆	91.36	71.78	**94.00**	94.88	95.76	80.65	96.22	86.77

and ACDC [26]. The selected experimental setting aligns with the existing literature on medical image segmentation [11], [12], [45], [46], [47], [48].

1) MSD BRAINTUMOR

The first one is the BrainTumor segmentation dataset from the Medical Segmentation Decathlon (MSD BrainTumor) [25]. It consists of 484 MRI images, each containing four channels: FLAIR, T1w, T1gd, and T2w. The images were annotated with three tumor sub-regions: edema (ED), enhancing tumor (ET), and non-enhancing tumor (NET). To be coherent with the results reported by [12], the segmentation metrics were computed on the classes ET, tumor core (TC, which is the union of ET and NET), and whole tumor (WT, which is the

TABLE 4. 5-fold cross-validation results on the Automatic Cardiac Diagnosis (ACDC) dataset. Our proposals are marked with †. The evaluation metric is the DSC (%). Best results are in bold while the second best are underlined. Methods subjected to a one-sided paired samples t-test comparing our best method against the best of the alternatives are highlighted in blue. If the p-value associated with the test is less than 0.05, the result is indicated as statistically significant by ⋆.

	Model	Average	RV	Myo	LV
CNNs	nnU-Net [30]	91.42	90.10	88.74	95.41
	nnU-NetResEnc [30]	90.84	89.17	88.52	94.84
	MedNeXt-M-K3 [49]	91.64	89.43	89.77	95.72
	MedNeXt-M-K5 [49]	90.70	88.50	88.88	94.73
Transformers	TransUNet [10]	89.75	88.88	84.66	95.70
	TransBTS [50]	91.29	90.42	87.94	95.51
	CoTr [48]	90.90	89.17	88.34	95.18
	UNETR [12]	88.72	85.55	86.48	94.12
	SwinU-Net [11]	89.97	88.29	85.61	**96.01**
	Swin-UNETR [13]	91.36	90.48	87.84	95.75
	LeViT-UNet-384s [47]	90.21	89.78	87.10	93.75
	MISSFormer [51]	87.73	86.55	85.24	91.42
	nnFormer [45]	91.87	**90.78**	89.37	95.46
Mamba	UMambaBot [44]	90.44	87.67	88.76	94.89
	UMambaEnc [44]	90.07	87.34	88.23	94.65
	SegMambaSkip†	91.49	89.58	89.51	95.39
	SegMamba†	91.33	89.37	89.40	95.22
	BiSegMamba†	91.50	89.46	89.66	95.37
	MultiSegMamba†	**92.04**⋆	90.39	**90.29**	95.44

union of ED, ET, and NET). Following the split provided by [12], we employed 95% of the dataset as a training/validation set with 5-fold cross-validation, and 5% for testing purposes. This task is difficult due to the complex, irregular tumor morphology and the need to fuse information from multiple MRI sequences.

2) SYNAPSE MULTI-ORGAN

The second dataset is the Synapse Multi-organ segmentation dataset [4], published within the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. This dataset includes 3 779 axial contrast-enhanced abdominal CT images from 30 abdominal CT scans, with each volume consisting of 85 to 198 slices. We adopted the same split as in [10], with 18 cases for training and 12 cases for testing. In line with our competitors, the evaluation metrics for this dataset were calculated for eight out of thirteen annotated abdominal organs: aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. This segmentation task tests a model's ability to handle widely varying organ sizes and low-contrast boundaries between organs–for example, smaller organs can be hard to discern because of their tiny volume and the ambiguous, noisy CT contrast with surrounding tissues.

3) ACDC

Lastly, the third dataset employed is the ACDC dataset [26]. It comprises 100 cine MRI scans of patients across five pathology groups, each labeled for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). We divided this dataset into 80 samples for training and validation, and 20 test samples, following the split described in [10]. Here, the challenges include poor contrast between blood and myocardium, motion artifacts, and substantial anatomical variability due to different cardiac diseases.

The aforementioned datasets–which spans diverse anatomies (brain, abdomen, heart), imaging modalities (multi-sequence MRI, CT, cine MRI), and class structures (pathological tumor subregions, normal organs, dynamic heart anatomy)–serve as a robust benchmark for evaluating our methods against the state-of-the-art, thereby demonstrating their strong performance and wide generalizability in 3D medical image segmentation.

B. PRE-PROCESSING AND AUGMENTATIONS

Data have been pre-processed leveraging the nnU-Net framework [30]. Such a pre-processing protocol includes Z normalization, foreground cropping, and resampling to a common spacing for all the images. The protocol is tuned based on the dataset fingerprint during the planning phase. Similarly, we employed the data augmentation techniques included in the nnU-Net framework, which include random rotations, scaling, intensity modifications, elastic deformations, and mirroring.

By following these well-established procedures, we align our work with state-of-the-art methodologies and facilitate direct comparisons with other studies in the literature.

C. EVALUATION METRICS

We employ Dice Similarity Coefficient (DSC in %) and the 95th percentile Hausdorff Distance (HD95 in mm), two widely accepted metrics for the segmentation task [52].

The DSC has practically the same meaning as the IoU (Intersection over Union), but the first one is better suited when the region of interest is much smaller than the background. In such a scenario, DSC can be more robust and informative than IoU since more weight is given to the correctly identified region. The DSC metric and its relationship with the IoU are expressed by the following formula:

$$\text{DSC(P,GT)} = \frac{2 \times |P \cap GT|}{|P| + |GT|} = \frac{2 \times IoU}{1 + IoU} \tag{8}$$

where P is the model prediction and GT is the ground truth.

On the other hand, the HD95 computes the maximum distance between two sets of points, considering the 95th percentile of these distances. In general, the 95th percentile of the distances between boundary points in A and B is defined as follows:

$$\text{d}_{95}(A, B) = \text{x}^{95}_{a \in A} \left\{ \min_{b \in B} d(a, b) \right\} \tag{9}$$

where $x^{95}_{a \in A}\{\}$ denotes the 95th percentile of the elements in the set enclosed within the brackets. Given the set formed by the pixels in the predicted mask (P) and the set of pixels belonging to the ground truth (GT), the Hausdorff distance is determined as the maximum value of the two distances

TABLE 5. Computational comparison on the Synapse dataset. Our proposals are marked with †. The number of parameters is expressed in millions [M] and VRAM in gigabyte [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 is the average across all test volumes.

	Models	Params	GFLOPs	VRAM	Tr.	Inf.
CNNs	nnU-Net [30]	30.64	410.11	7.65	9.20	21.80
	nnU-NetResEnc [30]	57.50	502.49	10.00	10.00	22.20
	MedNeXt-M-K3 [49]	32.65	248.03	15.32	67.60	153.60
	MedNeXt-M-K5 [49]	34.75	308.01	18.85	218.30	416.90
Transf.	TransUNet [10]	96.07	88.91	16.25	26.50	73.90
	CoTr [48]	50.12	369.22	8.10	18.60	41.40
	UNETR [12]	92.49	75.76	15.29	15.40	39.50
	Swin-UNETR [13]	62.83	384.20	13.91	22.00	38.70
	nnFormer [45]	150.50	213.41	9.73	8.20	20.60
Mamba	UMambaBot [44]	41.95	156.32	13.55	22.00	54.20
	UMambaEnc [44]	42.85	231.18	26.42	37.90	89.30
	SegMambaSkip†	62.36	486.92	29.26	12.60	93.50
	SegMamba†	61.49	480.90	25.61	12.70	99.60
	BiSegMamba†	64.75	494.17	27.31	16.50	134.10
	MultiSegMamba†	68.46	527.56	36.92	18.20	149.00

between P and GT and GT and P at the 95th percentile:

$$\text{HD95(P,GT)} = \max\left\{d_{95}(P, GT), d_{95}(GT, P)\right\} \quad (10)$$

By using the 95th percentile, this metric provides a robust evaluation that is less sensitive to outliers or extreme differences between the sets of points.

D. COMPARED METHODS

Performance comparison has been performed on recently proposed methods for medical image segmentation. Specifically, considered competitors can be classified into three main groups: CNN-, Transformer-, and Mamba-based architectures.

In the former group, we include the original nnU-Net [30] configuration making use of the vanilla U-Net architecture (nnU-Net), and its variations based on the U-Net with residual connections in the encoder (nnU-NetResEnc). Furthermore, the transformer-inspired-CNN-modification based on ConvNeXt blocks, MedNeXt [49], has been considered in its two variations K3, and K5. For what concerns Tranformer-based architectures, we compare our proposals with TransU-Net [10], TransBTS [50], CoTr [48], an hybrid architecture combining convolutional and transformer modules, UNETR [12], SwinU-Net [11] and its UNETR-based variation Swin-UNETR [13], LeViT-UNet-384s [47], MISSFormer [51], and the recently published nnFormer [45]. Finally, we include UMamba [44] in its two variations UMamba Bot and UMamba Enc.

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. The nnU-Net five-fold cross-validation schema has always been employed.

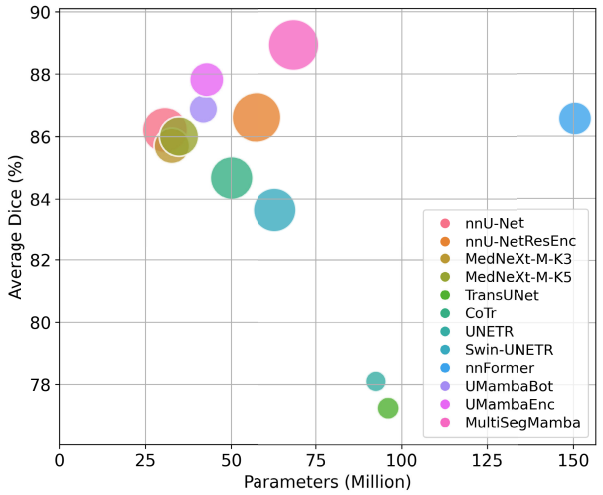


FIGURE 4. Deployment model size and average DSCs across our best model and competitors on Synapse. Circle size indicates GFLOPS.

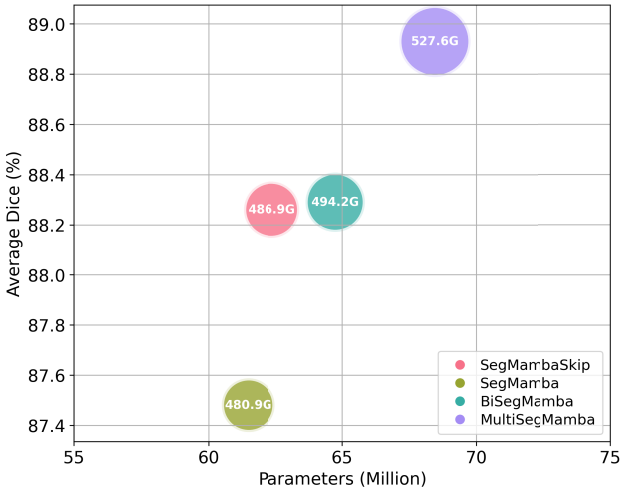


FIGURE 5. Deployment model size and average DSCs across our models on Synapse. Circle size indicates GFLOPS.

E. RESULTS

The experimental results for the BrainTumor, Synapse Abdomen, and Automatic Cardiac Diagnosis (ACDC) datasets are presented in Tab. 2, Tab. 3, and Tab. 4, respectively. Furthermore, to enhance the robustness and reliability of our findings, we conducted a one-sided paired samples t-test comparing our best-performing method (MultiSegMamba in most instances) with the most competitive alternative available in the literature. Experiments subjected to statistical testing in the average DSC and average HD95 columns are highlighted in blue. When the test yields a p-value below 0.05, the method demonstrating statistically superior performance–i.e., higher DSC or lower HD95–is marked with a star symbol (⋆).

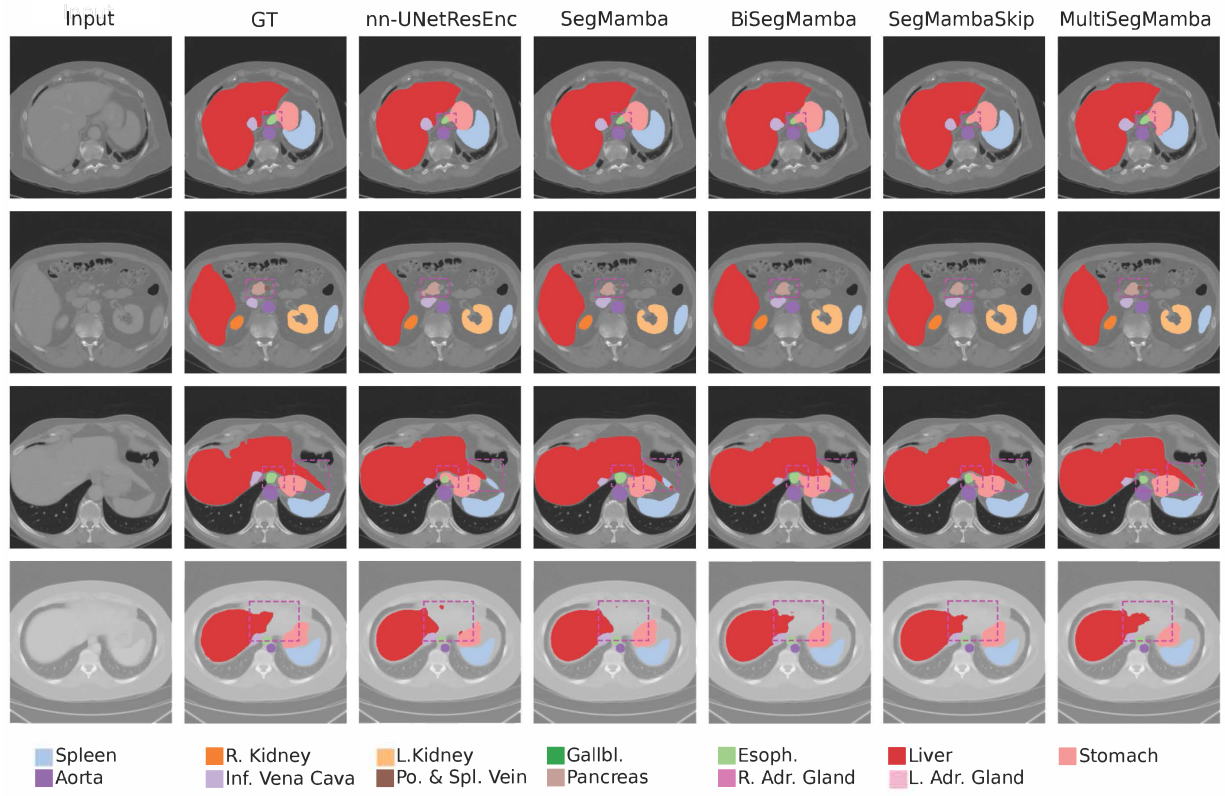


FIGURE 6. Visualization of segmentation results for four sample cases from the Synapse Abdomen evaluation set. Annotation errors are marked with magenta-dashed boxes. The figure is best viewed in color and zoomed in. From left to right: Input, Ground Truth (GT), nnU-NetResEnc, SegMamba, BiSegMamba, SegMambaSkip, and MultiSegMamba.

Overall, our methods are statistically superior to the state-of-the-art, with the exception of one HD95 comparison in the BrainTumor dataset (Tab. 2). In that instance, the statistical analysis does not provide significant evidence to suggest that UMambaBot outperforms SegMamba or vice versa, indicating that their performance is equivalent.

In general, BiSegMamba and MultiSegMamba consistently outperform SegMambaSkip, even if the latter is always competitive with state-of-the-art models and, on some specific classes, proves to be superior to them.

As the results on the BrainTumor dataset show (Tab. 2), SegMamba, BiSegMamba, and MultiSegMamba always outperform SegMambaSkip on average metrics and on most individual classes taken separately. Among the SegMamba models, MultiSegMamba, the one that harnesses more directions, outperforms the other configurations, demonstrating the importance of modeling multiple directions. Excluding nnFormer, our Mamba-based architectures gain more than 3 dice points over best-performing transformer-based architectures and up to 1 dice point over nnU-Net.

For what concerns the Synapse Abdomen dataset (Tab. 3), characterized by a larger number of classes, results show that our model showcases substantial improvements in kidney and spleen segmentation, as well as on average HD95 and DSC, when compared to state-of-the-art architectures. Remarkably, the inclusion of four distinct directions yields a more pronounced improvement in gallbladder segmentation, which is the most difficult to segment. Indeed, the gallbladder is significantly smaller and varies more in shape and position compared to other organs, such as the liver, which is larger and more consistently shaped. Moreover, the close proximity of the gallbladder to other organs and structures in the abdominal cavity increases the complexity of distinguishing it in medical images. Results on gallbladder segmentation show that SegMamba reaches 62.21 Dice points, while its multidirection versions, such as BiSegMamba and MultiSegMamba, improve over it by 8 and 10 points, respectively.

Finally, results on the ACDC dataset are presented in Tab. 4. In such a scenario, MultiSegMamba outperforms all its variants that model fewer directions, and MultiSegMamba and BiSegMamba consistently outperform all the UMamba variations.

F. ABOUT MODEL SIZE

In Tab. 5, a comprehensive computational comparison on the Synapse dataset is reported considering the number of parameters (millions), GFLOPs, and GPU memory. Our proposed models have a higher number of parameters compared to classical 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 (∼60M) are, on average, the double with respect to

TABLE 6. Clinical Metrics of the Left Ventricle for the test set of the ACDC Dataset. Our proposals are marked with †. This table presents a comparative evaluation of the end-diastolic volumes (Dd) and the derived ejection fractions (EF) between the ground truth measurements and the AI predictions. The metrics include the Pearson correlation (ρ), mean absolute error (MAE), and the mean difference with standard deviation (reported as bias $\pm\,\sigma$).

Model	Left Ventricle Dd			Left Ventricle EF		
	ρ	bias $\pm\,\sigma$ ml	MAE ml	ρ	bias $\pm\,\sigma$ %	MAE %
nnU-NetResEnc [30]	0.996	-3.5±8.6	6.415	0.990	-1.7±2.8	2.392
nnFormer [45]	0.996	-2.6±8.0	5.576	0.990	-1.6±2.7	2.461
SegMambaSkip†	0.996	-1.8±7.6	5.133	0.990	-1.8±2.8	2.602
SegMamba†	0.995	-3.9±8.9	6.771	0.979	-1.0±4.0	3.058
BiSegMamba†	0.995	-4.3±8.9	7.451	0.976	1.5±4.3	3.226
MultiSegMamba†	0.996	-2.9±8.0	5.936	0.991	-1.9±2.7	2.496

nnU-Net (∼31M), comparable to those of nnU-NetResEnc (∼57M), and much lower than those of transformer-based models (from ∼95M of TransU-Net and UNETR, up to 150M of nnFormer). On the ACDC and BrainTumor datasets, the metrics reported in Tab. 5 follow the same trends: parameter counts remain stable across datasets (aside from minor variations due to differing input/output channels), and GFLOPs and VRAM scale with batch and patch sizes.

As an additional visualization, Fig. 4 and Fig. 5 are provided. The former compares the average Dice score, model size, and computational complexity between our largest proposed model, MultiSegMamba, and several state-of-the-art methods on the Synapse dataset. Notably, MultiSegMamba surpasses the performance of all competitors while maintaining a comparable number of parameters and similar computational cost. Fig. 5, instead, provides a similar comparison between the proposed Mamba-based variants, highlighting the proportionality between computational cost, number of targeted directions, and segmentation performance.

G. CLINICAL METRICS

In datasets such as ACDC, the integration of additional clinical metrics can substantially enhance the evaluation of our predictions. In particular, accurate segmentation of the left ventricle is of paramount importance, as it underpins the calculation of the ejection fraction, a parameter that quantifies, as a percentage, the volume of blood pumped out by the left ventricle with each contraction and is a crucial indicator of cardiac function.

For each patient in the ACDC dataset, two MRI scans are provided: one corresponding to the end-diastolic phase and one to the end-systolic phase, with the former preceding the latter. Accordingly, the volume of the left ventricle is measured in these two distinct phases, denoted as Dd (end-diastolic volume) and Ds (end-systolic volume). The ejection fraction (EF) is then computed as follows:

$$EF = \frac{Dd - Ds}{Dd} \times 100\,\% \tag{11}$$

Subsequently, we compared the end-diastolic volumes (Dd) and the derived ejection fractions (EF) from both ground truth and AI predictions using statistical metrics, including Pearson correlation (ρ), mean absolute error (MAE), and the mean difference with standard deviation (reported as $bias \pm \sigma$). The results of these analyses are presented in Tab. 6.

An examination of the aforementioned table reveals that all methods demonstrate exceptional precision in predicting both the left ventricular end-diastolic volume (Dd) and the corresponding ejection fraction (EF). For the Dd measurements, the Pearson correlation coefficient exceeds 0.995 for all models, and the mean absolute error (MAE) is minimal, with a worst-case value of 7.5 ml. Given that the average ground truth Dd is 196 ml, this corresponds to a relative error of 3.8%. Similarly, for the ejection fraction, the Pearson correlation coefficient is above 0.976 for all models, and the MAE is comparably low, with a worst-case error of 3.2%.

H. QUALITATIVE EVALUATION

Fig. 6 depicts a qualitative comparison of the four variations of the proposed architecture. The comparison is performed on samples taken from the Synapse Abdomen evaluation set. As can be seen, all of our Mamba-based variations perform qualitatively similarly, but the ones leveraging multiple directions are less prone to errors when dealing with fine-grained details. This confirms the quantitative results previously discussed.

V. 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. Additionally, 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, either in skip connections or prior to pooling operations, utilizing both single-directional, bi-directional, and multi-directional implementations. The overall framework blends Convolutions and State-Space Models, leveraging the former for encoding precise spatial information, while addressing the latter to model long-range voxel-level interactions. Mamba provides a global context alongside voxel-wise precision, the former missing in traditional convolutional layers due to limited receptive fields and the latter absent in transformers due to 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 in image encoding and segmentation.

A. LIMITATIONS AND FUTURE WORKS

Despite the advancements made with the Mamba model, two key limitations can be identified and should be addressed in future research.

First, as Mamba is inherently a causal model, its application to non-causal visual data requires modification. Specifically, we tried to solve this problem by processing each sequence both forward and backward. However, this introduces redundancy, increasing the risk of overfitting. We believe that more efficient approaches could be developed to address this issue.

Second, to capture spatial relationships, we unfold image patches from multiple directions, but more effective methods, such as identifying optimal scanning paths or partitioning larger volumes into smaller neighborhoods, may exist. Furthermore, employing too many directions can significantly increase computational demands and redundancy, as mentioned before.

REFERENCES

[1] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, Jan. 2021.

[2] F. Bolelli et al., "Segmenting the inferior alveolar canal in CBCTs volumes: The ToothFairy challenge," *IEEE Trans. Med. Imag.*, vol. 44, no. 4, pp. 1890–1906, Apr. 2025.

[3] F. Bolelli, K. Marchesini, N. van Nistelrooij, L. Lumetti, V. Pipoli, E. Ficarra, S. Vinayahalingam, and C. Grana, "Segmenting maxillofacial structures in CBCT volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 1–10.

[4] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge*, vol. 5, 2015, p. 12.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[9] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas, "A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark," 2022, *arXiv:2203.00131*.

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[11] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2021, pp. 205–218.

[12] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

[13] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, Jan. 2022, pp. 272–284.

[14] L. Lumetti, V. Pipoli, F. Bolelli, E. Ficarra, and C. Grana, "Enhancing patch-based learning for the segmentation of the mandibular canal," *IEEE Access*, vol. 12, pp. 79014–79024, 2024.

[15] L. Lumetti, V. Pipoli, F. Bolelli, E. Ficarra, and C. Grana, "Location matters: Harnessing spatial information to enhance the segmentation of the inferior alveolar canal in CBCTs," in *Proc. 27th Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2024, pp. 108–123.

[16] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2022, pp. 4296–4313.

[17] V. Pipoli, G. Attanasio, M. Lovino, and E. Ficarra, "Squeeze and learn: Compressing long sequences with Fourier transformers for gene expression prediction," in *Proc. 20th Int. Conf. Intell. Comput. (ICIC)*, 2024, pp. 857–867.

[18] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020, *arXiv:2006.04768*.

[19] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, Ł. Kaiser, D. Belanger, L. J. Colwell, and A. Weller, "Rethinking attention with performers," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2020, pp. 1–14.

[20] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[21] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[22] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

[23] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[24] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "HiPPO: Recurrent memory with optimal polynomial projections," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 1474–1487.

[25] M. Antonelli et al., "The medical segmentation decathlon," *Nature Commun.*, vol. 13, no. 1, p. 4128, Jul. 2022.

[26] O. Bernard, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[27] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet–A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[28] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quant. Imag. Med. Surg.*, vol. 10, no. 6, pp. 1275–1285, Jun. 2020.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[30] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[31] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2020, pp. 1–12.

[32] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state-space layers," 2021, *arXiv:2110.13985*.

[33] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. Van Der Wal, "Pythia: A suite for analyzing large language models across training and scaling," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 2397–2430.

[34] B. Wang and A. Komatsuzaki. (2021). *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. [Online]. Available: https://github.com/kingoflolz/mesh-transformer-jax

[35] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. Victoria Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and

L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[36] M. Poli, S. Massaroli, É. Nguyen, D. Y. Fu, T. Dao, S. A. Baccus, Y. Bengio, S. Ermon, and C. Ré, "Hyena hierarchy: Towards larger convolutional language models," in *Proc. 40th Int. Conf. Mach. Learn.*, Jan. 2023, pp. 1–12.

[37] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2024, pp. 103031–103063.

[38] H. Gong, L. Kang, Y. Wang, X. Wan, and H. Li, "NnMamba: 3D biomedical image segmentation, classification and landmark detection with state space model," 2024, *arXiv:2402.03526*.

[39] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: UNet-like pure visual mamba for medical image segmentation," 2024, *arXiv:2402.05079*.

[40] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision mamba UNet for medical image segmentation," 2024, *arXiv:2402.02491*.

[41] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation," 2024, *arXiv:2401.13560*.

[42] H. Zhang, Y. Zhu, D. Wang, L. Zhang, T. Chen, Z. Wang, and Z. Ye, "A survey on visual mamba," *Appl. Sci.*, vol. 14, no. 13, p. 5683, Jun. 2024.

[43] L. Lumetti, V. Pipoli, K. Marchesini, E. Ficarra, C. Grana, and F. Bolelli, "Accurate 3D medical image segmentation with mambas," in *Proc. IEEE 22nd Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2025, pp. 1–5.

[44] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," 2024, *arXiv:2401.04722*.

[45] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "NnFormer: Volumetric medical image segmentation via a 3D transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023.

[46] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. Shahbaz Khan, "UNETR++: Delving into efficient and accurate 3D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 9, pp. 3377–3390, Sep. 2024.

[47] G. Xu, X. Zhang, X. He, and X. Wu, "LeViT-UNet: Make faster encoders with transformer for medical image segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Dec. 2023, pp. 42–53.

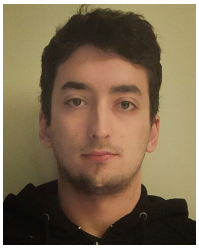
[48] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Intervent*. Cham, Switzerland: Springer, Jan. 2021, pp. 171–180.

[49] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. Maier-Hein, "MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*. Cham, Switzerland: Springer, Jan. 2023, pp. 405–415.

[50] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*. Cham, Switzerland: Springer, Jan. 2021, pp. 109–119.

[51] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[52] L. Maier-Hein et al., "Metrics reloaded: Recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 195–212, Feb. 2024.



LUCA LUMETTI received the B.Sc. and M.Sc. degrees in computer engineering from the Universitá degli Studi di Modena e Reggio Emilia, Italy, where he is currently pursuing the Ph.D. degree with the AImageLab. His research interests include artificial intelligence, computer vision, and medical imaging.

VITTORIO PIPOLI received the B.Sc. degree in computer engineering and the M.Sc. degree in data science and engineering from the Politecnico di Torino, Italy. He is currently pursuing the Ph.D. with the AImageLab, University of Modena and Reggio Emilia. His research interests include artificial intelligence, computer vision, and medical imaging.

KEVIN MARCHESINI received the B.Sc. and M.Sc. degrees in computer engineering from the Universitá degli Studi di Modena and Reggio Emilia. He is currently pursuing the Ph.D. degree with the AImageLab, with research interests in artificial intelligence and computer vision in medical imaging applications. He completed his master's thesis on endoscopic video segmentation during a six-month internship at the Orsi Academy, Belgium.

ELISA FICARRA received the Ph.D. degree in systems and computer engineering from the Politecnico di Torino, Turin, Italy, in 2006. She is currently a Full Professor with the Dipartimento di Ingegneria "Enzo Ferrari," Universitá degli Studi di Modena e Reggio Emilia, Italy. Her research interests include biological image processing and bioinformatics, high-throughput sequencing analysis, and artificial intelligence for biological and smart manufacturing applications.

COSTANTINO GRANA (Member, IEEE) received the degree from the Universitá degli Studi di Modena e Reggio Emilia, Italy, in 2000, and the Ph.D. degree in computer science and engineering, in 2004. He is currently a Full Professor with the Dipartimento di Ingegneria "Enzo Ferrari," Universitá degli Studi di Modena e Reggio Emilia. He has published six book chapters, 47 articles in international peer-reviewed journals, and more than 130 papers at international conferences. His research interests include medical imaging, optimization of image processing algorithms, and computer vision applications.

FEDERICO BOLELLI (Associate Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree from the Universitá degli Studi di Modena e Reggio Emilia, Italy. He is currently a Tenure-Track Assistant Professor with the Dipartimento di Ingegneria "Enzo Ferrari," Universitá degli Studi di Modena e Reggio Emilia. His research interests include image processing, algorithms and optimization, and medical imaging.