# MambaVesselNet: A Hybrid CNN-Mamba Architecture for 3D Cerebrovascular Segmentation

Yanming Chen
University of Nottingham Ningbo China
Ningbo, China
Yanming0117@outlook.com

Ziyu Liu
University of Nottingh am Ningbo China
Ningbo, China
Ziyu.Liu@nottingham.edu.cn

Xiangjian He*
University of Nottingham Ningbo China
Ningbo, China
Sean.He@nottingham.edu.cn

## Abstract

Segmenting vessels in magnetic resonance imaging (MRI) stands as a mainstream approach for evaluating cerebrovascular conditions. Due to the complex semantics and topology of cerebrovascular structures, existing CNN-based segmentation methods often fail to correlate the topological structure and branch vessels, resulting in incomplete segmentation. To address the challenge of global dependencies modelling, transformer architectures have been employed due to their capability of capturing long-range dependencies, and they have shown promise in 3D medical image segmentation. However, the transformer architecture greatly increases the computational burden when processing high-dimensional 3D MRI images. In light of this, a selective state space model (SSM) Mamba has gained recognition for its adeptness in handling long-range dependencies in sequential data, particularly noted for its efficiency and speed in natural language processing applications. Mamba is now widely applied in various computer vision tasks. Based on these findings, in this study, we propose MambaVesselNet, a Hybrid CNN-Mamba network for 3D cerebrovascular segmentation. MambaVesselNet leverages CNNs to capture local features and incorporates the Mamba block at the bottleneck to model long-range dependencies within the whole-volume features. The effectiveness of MambaVesselNet is validated on a public cerebrovascular dataset, and our benchmark demonstrates new state-of-the-art performance.

## Keywords

State space models, Mamba, 3D medical imaging, Cerebrovascular segmentation

**ACM Reference Format:**
Yanming Chen, Ziyu Liu, and Xiangjian He. 2024. MambaVesselNet: A Hybrid CNN-Mamba Architecture for 3D Cerebrovascular Segmentation. In *ACM Multimedia Asia (MMASIA '24), December 3–6, 2024, Auckland, New Zealand*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696409.3700231

*The corresponding author


*MMASIA '24, December 3–6, 2024, Auckland, New Zealand*

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700231

## 1 Introduction

Cerebrovascular angiographies, particularly through magnetic resonance imaging (MRI), are fundamental approaches in diagnosing and planning treatments for cerebrovascular diseases. The segmentation of cerebral vessels in Magnetic Resonance Angiogram (MRA) volumes presents significant challenges due to the complex 3D nature of the vascular structures and complexities in processing high-dimensional and high-resolution images. Traditional segmentation methods, relying on intensity-based features [27, 28], often fell short in accurately delineating cerebral vessels. In the field of deep learning, Fully Convolutional Neural Networks (CNNs), especially those with U-shaped encoder-decoder structures [14, 15, 29], have set new benchmarks in various segmentation tasks. In the classic U-Net architecture, the encoder is used to capture global context by progressively reducing the feature dimensions, while the decoder then enlarges these features back to the original input size for precise pixel or voxel-level segmentation. While CNN-based models excelled in representation learning, their ability to capture whole-volume context was constrained by their localized receptive fields due to the limited kernel size [13]. As shown in Fig. 1, CNN-based methods could only extract the feature maps from limited receptive fields, but due to the special topology structure of a human cerebrovascular structure, some global dependencies between vessels might be ignored.

To address this, several efforts were made using a large kernel convolution [17, 20] to facilitate larger receptive fields. However, the inherent locality of these fields in convolutional architecture continued to restrict the network capacity for learning from broader regions.

Recently, the transformer architecture [26], with its self-attention mechanism for global information extraction, gained attention in the realm of 3D medical imaging because of its ability to model long-range dependencies and grasp the global context [7]. Specifically, transformers represented images through a sequence of 1D patch embeddings and mechanisms to learn the relationships among tokens. This enabled effective long-range information modelling. However, the typically high-resolution 3D volumes produced by the MRA scans resulted in lengthy 1D sequences, imposing significant computational demands on transformer-based approaches.

To address the limitations of long-sequence modelling, the selective state space model Mamba [9] was proposed to effectively model whole-volume dependencies through a novel selection mechanism that enhances the training and inference efficiency. In the field of computer vision, VM-UNet [24] employed a pure SSM-based model for 2D medical image segmentation and demonstrated competitive performance on skin lesion and organ segmentation tasks. VMamba

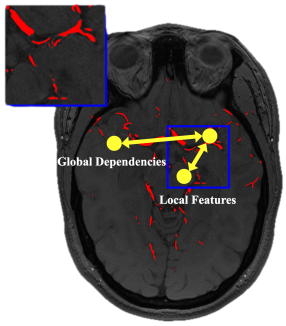


**Figure 1: Local Features and Global Dependencies. This figure shows an MRI scan with highlighted regions indicating local features and global dependencies. The local features are marked within a blue box, and the global dependencies are indicated with yellow arrows connecting various points. This visual representation illustrates the limited scope of local feature extraction and the importance of capturing global dependencies in cerebrovascular segmentation.**

[18] incorporated a Cross-Scan Module (CSM) for better spatial domain traversal and achieved linear complexity without sacrificing global receptive fields.

In this study, we propose a Hybrid CNN-Mamba (MambaVesselNet) architecture that leverages the advantages of both CNN and Mamba block. Specifically, the CNNs-encoder is used to extract local features from 3D images, generating a series of feature maps. At the bottleneck stage, the output from the CNN is transformed into a 1D sequence and then processed through the Mamba block to learn contextual information from the encoded patches. Feature representations derived from this stage are subsequently fed back into a CNN-based decoder to generate the segmentation outputs. This entire network structure adheres to a U-shaped configuration [23], with skip connections implemented between the encoder and decoder. To our knowledge, MambaVesselNet is the first network that utilizes Mamba for segmenting 3D cerebrovascular structures. The main contributions of this work are summarized as follows.

- We propose a novel Hybrid CNN-Mamba model (MambaVesselNet) for 3D medical image Segmentation. To our knowledge, this is the first study that explored the integration of CNN and Mamba for segmenting cerebrovascular structures in MRA scans.
- Different from other Mamba-based vision models using Mamba as a backbone across different scales, we propose a novel architecture that places the Mamba module at the bottleneck to extract information at the same dimension, reducing computational load while achieving better performance.
- We validate the effectiveness of our proposed model on the public cerebrovascular dataset [6]. MambaVesselNet achieves new state-of-the-art performance.

## 2 Related Work

**CNN-based Segmentation Networks.** Fully Convolutional Neural Networks (CNNs) were increasingly applied in 3D vascular segmentation. For instance, Chen et al. [5] used innovative CNN architectures for tasks like intracranial artery and vascular boundary segmentation. Tetteh et al. [25] introduced DeepVesselNet with efficient convolutional filters and a class-balancing loss function. Despite these advancements, challenges like morphological variation and class imbalance in cerebrovascular segmentation in TOF-MRA volumes remained. Yuan and Yang [32] developed a two-stage FCN approach for the segmentation of the aortic vessel tree in CT scans, achieving high accuracy in the MICCAI segmentation challenge. Banerjee et al. [1] proposed a domain-general AI method for the volumetric analysis of cerebrovascular structures across multiple MRA centres using TOF-MRA, employing a multi-task deep CNN with a topology-sensitive loss function to enhance segmentation accuracy. Despite their success in various segmentation tasks, convolutional networks fell short in capturing global context and global spatial dependencies.

**Transformer-based Models.** Transformer-based models made significant strides in the field of medical imaging and computer vision [7, 33] and medical image analysis [30]. To enhance the model's ability to model long-range dependencies, Chen et al. [4] firstly proposed TransUnet, a medical image segmentation framework that combined the global contextual strengths of transformers with the detailed spatial resolution capabilities of U-Net. This concept was further expanded by the introduction of UNETR [12]. UNETR utilized the Vision Transformer (ViT) as its encoder to learn the global context and then merged it with a CNN-based decoder through skip connections at various resolutions. Furthermore, SwinUNETR [11] used the Swin Transformer [19] as an encoder. Different from standard transformers, Swin Transformers introduced an innovative hierarchical design that computed self-attention within non-overlapping local windows, and these windows were shifted across layers, allowing the model to capture multi-scale features while maintaining efficiency. However, contemporary 3D medical imaging modalities, such as magnetic resonance imaging (MRI), often generated highly detailed, multi-volumetric data sets. These high-resolution volumetric scans translated into lengthy 1D sequences for analysis, thus imposing substantial computational burdens on transformer-based frameworks.

**Selective State Space Models.** To address the computational challenges brought by the high-resolution 3D volumes, the selective state space model (SSM) [9], Mamba was proposed to effectively handle long-range dependencies. Mamba enhanced the efficiency of both training and inference by implementing a selection mechanism. In the context of computer vision, Mamba's application was explored in various contexts. U-Mamba [21] designed a hybrid CNN-SSM block that integrated Mamba into the decoder stage of nnUNet [15]. Thus, U-Mamba also featured a self-configuring capability that facilitated its autonomous adaptation to varying datasets. Vision Mamba (VMamba) [18] combined Vision Transformers' global receptive fields with selective state space models' efficiency, introducing the Cross-Scan Module (CSM) for handling non-causal visual data with linear complexity. In terms of 3D medical image segmentation, SegMamba [31] integrated the Mamba model into Unetr architecture. This novel approach, named Tri-orientated Mamba (ToM), modelled 3D features from multiple directions, enhanced by a Gated Spatial Convolution (GSC) module for spatial feature refinement.

Unlike the above works utilizing Mamba blocks as the backbone or encoder, we observe that this approach can lead to an overemphasis on global contextual information. This overemphasis may result in over-segmentation, where non-vessel areas are incorrectly identified as vessels due to the overshadowing of important local features. Conversely, traditional CNN-based models, while effective at capturing local features and spatial details, struggle to model long-range dependencies and global context, leading to under-segmentation where some vessel structures are missed due to their limited receptive fields (as discussed in Section 6). To address these issues, our model retains the traditional CNN-based U-Net architecture for both the encoder and decoder to leverage its strength in local feature extraction. We incorporate the Mamba block specifically at the bottleneck to effectively model long-range dependencies without compromising the network's ability to accurately localize vessel structures. This strategic placement bridges the gap between high-level feature abstraction and detailed spatial reconstruction, achieving a balance between over-segmentation and under-segmentation.

## 3 Preliminaries

Structured State Space Sequence Models (SSMs) like S4 [10] and Mamba modernize the classical system theory to handle sequential data by mapping a one-dimensional input $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ via intermediate states $h(t) \in \mathbb{R}^N$. According to [9], this process is framed as:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t),$$

where $A$, $B$, and $C$ are matrices that define the dynamics of the system, with $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{N\times 1}$. To facilitate the use of these models in computational settings, particularly in deep learning, the continuous-time models are discretized. The state-space representation is converted into a discrete-time version while allowing the model to be implemented in a computational framework that handles discrete data. The discretization involves the zero-order hold (ZOH) method, transforming continuous-time matrices into discrete-time equivalents as:

$$\tilde{A} = \exp(A\Delta), \quad \tilde{B} = (A\Delta)^{-1}(\exp(A\Delta) - I) \cdot \Delta B,$$

where $\tilde{A}$ and $\tilde{B}$ are the discrete-time equivalents of the continuous-time matrices $A$ and $B$. The parameter $\Delta$ represents the sampling period, which is the interval between consecutive samples in the discrete-time model. The matrix $\exp(A\Delta)$ is the matrix exponential of $A\Delta$, and defines how the state evolves over one sampling period. Additionally, Mamba innovates by integrating a selective scan mechanism, which dynamically adjusts the model's parameters based on the input data $x(t)$, enhancing adaptability and efficiency. This mechanism allows Mamba to selectively propagate or filter information along the sequence, based on the relevance and context of the current input, thus providing a more tailored and efficient computational model.

## 4 Methodology

An overview of MambaVesselNet architecture is presented in Fig. 2. MambaVesselNet follows a contracting-expanding pattern consisting of a stack of convolutional layers as a decoder and concatenated to the decoder via skip connections. Begins with a 3D input of cerebral vascular, MambaVesselNet is structured into four components: 1) the 3D feature encoder that down-samples the input through successive convolutional blocks, 2) a stack of Mamba block bottleneck to capture whole-volume dependencies and detailed spatial features, 3) the 3D feature decoder based on de-convolutions to perform the up-sample operations and generate the segmentation output, and 4) the skip connection to interlink the encoder's multi-scale features with the decoder.

### 4.1 Encoder

As depicted in Fig. 3, a CNN-based encoder is used in this study to down-sample the input TOF-MRA volumes through successive convolutional blocks. Each block applies two convolutional operations with $3 \times 3 \times 3$ and $2 \times 2 \times 2$ kernels, respectively, followed by layer normalization and a LeakyReLU activation function. The convolutional blocks are defined by $f_i = \text{LeakyReLU}(\text{Norm}_i(W_i * f_{i-1} + b_i))$, where $f_i$ is the output feature map after the $i$-th convolution, $W_i$ is the $i$-th convolutional weight, and $b_i$ is the $i$-th bias. Each Conv block has a residual connection: $f_{\text{out}} = \text{LeakyReLU}(f_i + f_{\text{input}})$. Following each convolutional block, the spatial dimensions are reduced by half and the number of feature channels is increased twofold through a down-sampling process.

### 4.2 Vision Mamba Block

We illustrate the architecture of the Vision Mamba Block (the most left block in Fig. 2), which is designed to capture long-range dependencies and enhance feature representation as:

$$\hat{\mathbf{z}}^l = \text{Mamba}(\text{LayerNorm}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \tag{1}$$

$$\hat{\mathbf{z}}^l_{\text{out}} = \text{MLP}(\text{LayerNorm}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l. \tag{2}$$

The Mamba layer processes the normalized input feature map $\mathbf{z}^{l-1}$, producing an intermediate representation $\hat{\mathbf{z}}^l$. This intermediate representation is then normalized again and passed through an MLP, resulting in the final output feature map $\hat{\mathbf{z}}^l_{\text{out}}$. Residual connections are used in both steps to ensure that the original input features ($\mathbf{z}^{l-1}$ and $\hat{\mathbf{z}}^l$) are added to the corresponding processed features.

### 4.3 Decoder

The decoder utilizes feature representations obtained from the bottleneck and employs skip connections at each resolution level. During each convolution stage $i \in \{0, 1, 2, 3, 4\}$ in the encoder, the resulting feature maps are reshaped to dimensions $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i} \times C_i$, where $C_i = 48 \times 2^i$. In the decoder, these feature maps are upsampled using a deconvolutional layer with a stride of 2, effectively doubling their resolution. These upsampled outputs are then concatenated with the corresponding outputs from the encoder stage via skip connections. The combined features are subsequently processed through a convolutional layer with a stride of 1 to refine the feature representations and generate outputs for each stage. Finally, the segmentation results are produced through a $1 \times 1 \times 1$ convolutional layer.

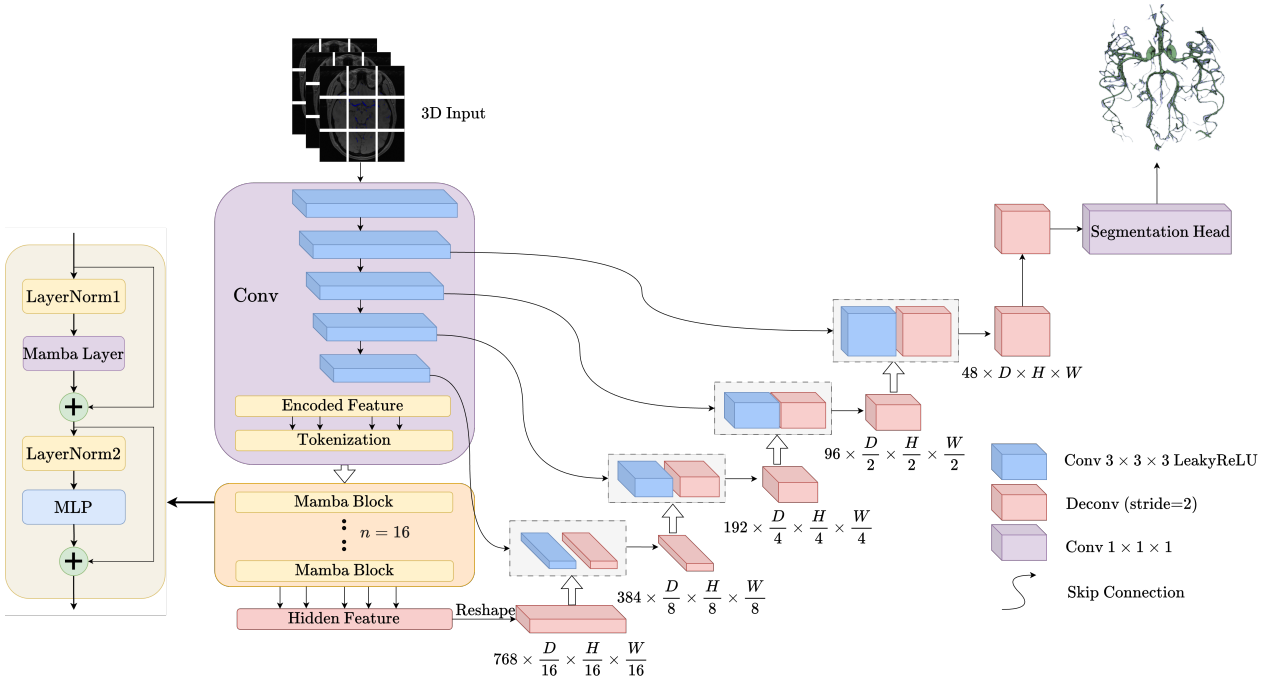


**Figure 2: Overview of MambaVesselNet architecture. The model takes a 3D patch from input images (channel size $C = 4$ for MRI images), and down-sample it through successive convolutional blocks, with each time reducing the spatial resolution by half and doubling the feature channels. After down-sampling and feature extraction operations, the output feature maps are reshaped into a sequence and fed into the Mamba Blocks ($n = 16$). Finally, the feature decoder performs the up-sampling operation and restores the reduced spatial dimensions images to their original size, and the skip connection is used to bridge the multi-scale features from the encoder and the decoder.**

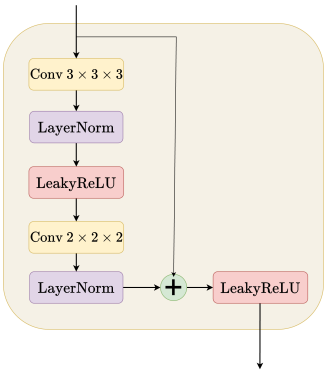


**Figure 3: Encoder architecture. The figure illustrates the encoder architecture. It begins with a convolutional layer (Conv $3 \times 3 \times 3$) followed by layer normalization (LayerNorm) and a LeakyReLU activation function. This sequence is repeated with a Conv $2 \times 2 \times 2$ layer, LayerNorm, and LeakyReLU. A residual connection is shown, linking the input to the output through an addition operation.**

## 4.4 Loss Function

The loss function DiceCELoss [1] [16] combines soft dice loss [22] and cross-entropy loss, computed in a voxel-wise manner based on:

$$\mathcal{L}(G, Y) = \lambda_{\text{dice}} \left( 1 - \frac{2 \sum_{i=1}^{I} G_i Y_i}{\sum_{i=1}^{I} G_i^2 + \sum_{i=1}^{I} Y_i^2} \right) + \lambda_{\text{ce}} \left( -\frac{1}{I} \sum_{i=1}^{I} G_i \log(Y_i) \right), \tag{3}$$

where $G_i$ represents the ground truth in one-hot encoded form, $Y_i$ denotes the probability output by the network for each class, and $I$ is the total number of voxels. The parameters $\lambda_{\text{dice}}$ and $\lambda_{\text{ce}}$ are the weighting factors for the Dice and cross-entropy components. In our implementation, both $\lambda_{\text{dice}}$ and $\lambda_{\text{ce}}$ are set to 1.0 by default, giving equal weight to both loss components.

## 5 Experiments

This section presents the experimental methodology and results for evaluating the efficacy of our benchmark MambaVesselNet in segmenting cerebrovascular structures from MRI images.

[1] https://docs.monai.io/en/stable/losses.html

## 5.1 Dataset

The IXI dataset [8], which is publicly accessible and includes over 600 MR images from healthy experimental subjects, is employed for model accuracy evaluation. For our work, we specifically use 45 Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) images that were annotated by Chen et al. [6]. We use these images captured with a 1.5 T Siemens Magnetom Vision MR scanner, with the primary scanning parameters being a 39 ms repetition time and a 7 ms echo time. The images feature a resolution of $1024\times1024\times92$, with each voxel measuring $0.264 \times 0.264 \times 0.8$ mm$^3$. For the division of our dataset, based on the work of Chen et al. [2, 3], we select 30 cases from the IXI dataset as the training set and the remaining 15 cases as the testing set. In each training step, we randomly sample the input images with volume sizes of $64 \times 64 \times 64$.

## 5.2 Evaluation Metrics

To assess the performance of our segmentation model, we use Dice score [34]. The Dice score quantifies the degree of overlap between the predicted segmentation and the actual ground truth by:

$$\text{Dice}(G, P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}, \tag{4}$$

where $G_i$ is the ground truth and $P_i$ is prediction for voxel $i$, with $I$ being the total number of voxels. Additionally, Precision is used to measure the proportion of true positive predictions among all predicted positives and Recall measures the proportion of true positive predictions among all actual positives.

## 5.3 Implementation Details

MambaVesselNet implementation is based on MONAI 1.2.3[2], Pytorch 1.13.1[3] and Cuda 11.6. The adaptive moment (Adam) optimizer updates the parameters along with the Cosine Annealing Learning Rate Scheduler. The learning rate is initialized at $1 \times 10^{-4}$ and annealed down to $1 \times 10^{-7}$ over the training. 5000 iterations are used to train the models and the model with the best validation accuracy is saved for testing. We use a patch size of $64\times64\times64$ and the batch size is set to 2 per GPU. All experiments are conducted on two NVIDIA A5000 GPUs.

## 5.4 Quantitative Evaluation

In this evaluation, we benchmark MambaVesselNet against five state-of-the-art segmentation models to assess its performance. These include two CNN-based approaches, UNet3D [14] and nnUNet [15], and two transformer-based methods, UNETR [12] and SwinUNETR [19]. Additionally, we compare it with another Mamba-based model, SegMamba [31] specialized in medical image segmentation. All models are trained and tested under the same data augmentation settings, and the public implementations of these models from MONAI are used to generate the best results.

As shown in Table 1, MambaVesselNet outperforms all other models in terms of Dice score and Precision. Specifically, MambaVesselNet achieves the highest Dice score of 0.870 and the highest Precision of 0.889. Although its Recall of 0.859 is slightly lower than that of SwinUNETR (0.874) and SegMamba (0.870), the higher Precision of MambaVesselNet contributes to a better overall Dice score. This balance between Precision and Recall is crucial, as it reflects the model's ability to accurately identify vessel structures while minimizing false positives. MambaVesselNet increases the Dice score by approximately 3.9% compared to UNet3D, 2.1% compared to nnUNet, and 0.6% compared to the previous Mamba model SegMamba. The superior performance of MambaVesselNet underscores the effectiveness of our proposed Hybrid CNN-Mamba architecture. Additionally, the Swin-Transformer-based model, SwinUNETR, achieves the highest Recall but with lower Precision, resulting in a lower Dice score compared to MambaVesselNet.

| Models | Precision ↑ | Recall ↑ | Dice ↑ |
|---|---|---|---|
| Unet3D[14] | 0.863 | 0.805 | 0.831 |
| nnUNet[15] | 0.856 | 0.845 | 0.849 |
| UNETR[12] | 0.835 | 0.842 | 0.836 |
| SwinUNETR[19] | 0.845 | **0.874** | 0.857 |
| SegMamba[31] | 0.863 | 0.870 | 0.864 |
| MambaVesselNet | **0.889** | 0.859 | **0.870** |

**Table 1: Quantitative comparisons of segmentation performance.**

## 5.5 Qualitative Results

2D cerebrovascular segmentation visualization are presented in Fig. 4. It shows that the Mamba-based methods, MambaVesselNet and SegMamba, identify finer cerebral vessels more accurately compared to traditional CNN-based approaches. In the central region of the magnified images, varying degrees of false positives are visible in all baseline methods but our benchmark MambaVesselNet exhibits fewer false positive markings and produces the segmentation closest to ground truth.

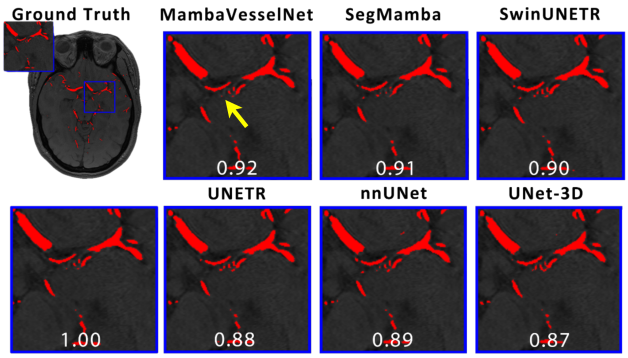


**Figure 4: 2D qualitative comparison of different baselines on the IXI dataset. The first image shows the ground truth, with a blue box zooming into a part of the region. It is followed by MambaVesselNet (Dice score 0.92), SegMamba (Dice score 0.91), SwinUNETR (Dice score 0.90), UNETR (Dice score 0.88), nnUNet (Dice score 0.89), and UNet-3D (Dice score 0.87). Each image highlights the segmentation quality in comparison to the ground truth. The Dice score shown represents the segmentation performance on this specific image.**

[2]https://monai.io/
[3]http://pytorch.org/

## 6 Discussion

Our experiments on IXI cerebrovascular datasets have obtained superior performance over both CNN-based and transformer-based approaches. Specifically, our model has achieved the highest Dice score and Precision across multiple benchmarks. By leveraging the Mamba block at the bottleneck of the network, MambaVesselNet has effectively captured long-range dependencies. This approach has improved segmentation accuracy without incurring additional computational burdens.

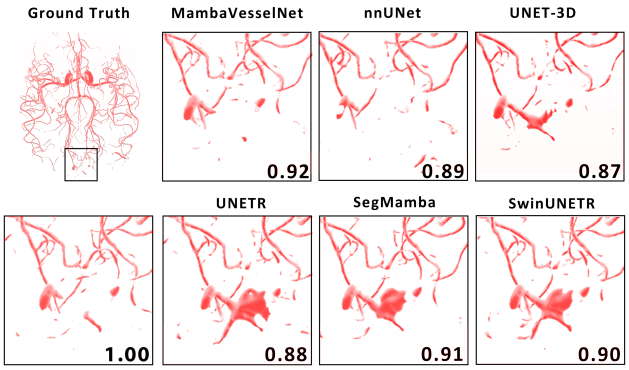


**Figure 5: This figure shows a 3D qualitative comparison of segmentation results on the IXI dataset. The first image displays the ground truth segmentation. The subsequent images show the results from MambaVesselNet (Dice score 0.92), nnUNet (Dice score 0.89), UNET-3D (Dice score 0.87), UNETR (Dice score 0.88), SegMamba (Dice score 0.91), and SwinUNETR (Dice score 0.90). Each model's segmentation quality is compared to the ground truth. The Dice score shown represents the segmentation performance on this specific image.**

As shown in Fig. 5, the qualitative evaluation reveals distinct differences in segmentation performance among various models. CNN-based models, such as nnUNet and UNet3D, tend to exhibit under-segmentation, missing some vessel structures due to their limited receptive fields and inability to capture long-range dependencies. However, they produce fewer false positives because they focus on local features and are less likely to misclassify non-vessel regions. In contrast, transformer-based models (UNETR and SwinUNETR) and the previous Mamba-based model (SegMamba), which utilize transformers or Mamba components as the backbone or encoder, excel at capturing global contextual information. This strength allows them to recognize long-range dependencies between vessel structures, leading to more complete segmentation. However, the overemphasis on global features can overshadow critical local details, resulting in over-segmentation where non-vessel areas are incorrectly segmented as vessels. Our proposed MambaVesselNet addresses these challenges by maintaining a traditional CNN-based design in both the encoder and decoder to ensure precise local feature extraction and spatial localization. By introducing the Mamba block exclusively at the bottleneck, we enable the model to capture essential global dependencies without letting them dominate the entire network. This design choice allows us to harness the benefits of both local and global information, effectively balancing high-level feature abstraction with detailed spatial reconstruction. Consequently, MambaVesselNet achieves superior segmentation performance by mitigating the over-segmentation seen in transformer or Mamba backbone models and the under-segmentation observed in purely CNN-based models, closely matching the ground truth.

## 7 Conclusion

This paper has proposed a hybrid CNN-Mamba model, MambaVesselNet, for the segmentation of volumetric MRI images. The proposed model has incorporated a U-shaped design, featuring a fully CNN-based encoder and decoder interconnected with skip connections. It has employed the selective state space model Mamba as a bottleneck to effectively learn long-range dependencies and capture global contextual representations. The effectiveness of MambaVesselNet has been validated on a public cerebrovascular IXI dataset in MRI modality, and the results have shown that it achieves new state-of-the-art performance, surpassing both CNN-based and transformer-based models as well as the previous Mamba-based models. Our discussion has demonstrated that our hybrid CNN-Mamba architecture effectively combines the advantages of CNNs and Mamba blocks, achieving a balance between over-segmentation and under-segmentation. In conclusion, this study has explored a novel architecture of integrating state space models with convolutional networks, thereby promoting the model's capability to handle complex 3D medical imaging tasks by combining local feature extraction with global contextual understanding.

## Acknowledgements

This work is partially supported by the Yongjiang Technology Innovation Project (2022A-097-G), and the Ningbo 2025 Key R&D Project (2023Z223).

## References

[1] Subhashis Banerjee, Fredrik Nysjö, Dimitrios Toumpanakis, Ashis Kumar Dhara, Johan Wikström, and Robin Strand. 2024. Streamlining neuroradiology workflow with AI for improved cerebrovascular structure monitoring. *Scientific Reports* 14, 1 (2024), 9245.

[2] Cheng Chen, Yunqing Chen, Shuang Song, Jianan Wang, Huansheng Ning, and Ruoxiu Xiao. 2023. Cerebrovascular Segmentation in TOF-MRA with Topology Regularization Adversarial Model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4250–4259.

[3] Cheng Chen, Kangneng Zhou, Tong Lu, Huansheng Ning, and Ruoxiu Xiao. 2023. Integration-and separation-aware adversarial model for cerebrovascular segmentation from TOF-MRA. *Computer Methods and Programs in Biomedicine* 233 (2023), 107475.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

[5] Li Chen, Yanjun Xie, Jie Sun, Niranjan Balu, Mahmud Mossa-Basha, Kristi Pimentel, Thomas S Hatsukami, Jenq-Neng Hwang, and Chun Yuan. 2017. 3D intracranial artery segmentation using a convolutional autoencoder. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 714–717.

[6] Ying Chen, Darui Jin, Bin Guo, and Xiangzhi Bai. 2022. Attention-assisted adversarial model for cerebrovascular segmentation in 3D TOF-MRA volumes. *IEEE Transactions on Medical Imaging* 41, 12 (2022), 3520–3532.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Katherine R Gray, Robin Wolz, Rolf A Heckemann, Paul Aljabar, Alexander Hammers, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. 2012. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage* 60, 1 (2012), 221–229.

[9] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[10] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems* 34 (2021), 572–585.

[11] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*. Springer, 272–284.

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 574–584.

[13] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. 2019. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3464–3473.

[14] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1055–1059.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 2 (2021), 203–211.

[16] Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 1–7. https://doi.org/10.1109/CIBCB48159.2020.9277638

[17] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 2022. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076* (2022).

[18] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. VMamba: Visual State Space Model. arXiv:2401.10166 [cs.CV] https://arxiv.org/abs/2401.10166

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[20] Pinjun Luo, Guoqiang Xiao, Xinbo Gao, and Song Wu. 2023. LKD-Net: large kernel convolution network for single image dehazing. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1601–1606.

[21] Jun Ma, Feifei Li, and Bo Wang. 2024. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv:2401.04722 [eess.IV]

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[24] Jiacheng Ruan and Suncheng Xiang. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491* (2024).

[25] Giles Tetteh, Velizar Efremov, Nils D Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Björn H Menze. 2020. Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *Frontiers in Neuroscience* 14 (2020), 1285.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[27] Dale L Wilson and J Alison Noble. 1999. An adaptive segmentation algorithm for time-of-flight MRA data. *IEEE transactions on medical imaging* 18, 10 (1999), 938–945.

[28] Xunlei Wu, Vincent Luboz, Karl Krissian, Stephane Cotin, and Steve Dawson. 2011. Segmentation and reconstruction of vascular structures for 3D real-time simulation. *Medical image analysis* 15, 1 (2011), 22–34.

[29] Zhitao Xiao, Bowen Liu, Lei Geng, Fang Zhang, and Yanbei Liu. 2020. Segmentation of lung nodules using improved 3D-UNet neural network. *Symmetry* 12, 11 (2020), 1787.

[30] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. 2021. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 171–180.

[31] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. 2024. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation. arXiv:2401.13560 [cs.CV]

[32] Shaofeng Yuan and Feng Yang. 2023. Segmentation of Aortic Vessel Tree in CT Scans with Deep Fully Convolutional Networks. *arXiv preprint arXiv:2305.09833* (2023).

[33] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6881–6890.

[34] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. 2004. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology* 11, 2 (2004), 178–189.