Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 3, 2025

MambaFormer3D: A Hybrid Transformer-Mamba Network for 3D Brain Tumor Segmentation

Beibei Hou, Saizong Guan

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China

Abstract: In recent years, Vision Transformers (ViTs) have demonstrated strong performance in 3D medical image segmentation, particularly in capturing global features and modeling long-range dependencies. However, the self-attention mechanism in ViTs faces significant computational challenges due to its quadratic complexity, especially when applied to three-dimensional medical images, resulting in a high number of parameters and increased computational overhead. This limitation hampers their deployment on resource-constrained devices. To address this challenge, this paper proposes a lightweight MambaFormer architecture, which leverages the efficient representation learning capability of the Mamba SSM and the global feature extraction strength of Transformers, achieving linear computational complexity. The synergy between the MALR and LMA modules ensures computational efficiency while significantly enhancing segmentation performance, particularly in brain tumor segmentation tasks. Experimental results demonstrate that MambaFormer3D exhibits superior performance in processing multi-modal BraTS data, showing great potential in computationally constrained environments.

Keywords: Image Segmentation; Transformer; Mamba; Lightweight.

1. Introduction

Medical image segmentation is a crucial technique for accurately delineating anatomical structures or pathological regions, with significant applications in clinical diagnosis, treatment planning, and disease monitoring [1]. Among its various applications, brain tumor segmentation is a vital task that aims to identify and segment different tumor subregions from multi-modal MRI scans, such as T1, T1ce, T2, and FLAIR [2]. These subregions include enhancing tumor, edema, and necrosis [3]. Due to the complex 3D volume of brain tumors, class imbalance, and indistinct boundaries, achieving accurate segmentation poses substantial challenges [4].

In brain tumor segmentation research, gliomas are the primary target, which are categorized into high-grade gliomas (HGG) and low-grade gliomas (LGG) based on their malignancy [5]. HGGs grow rapidly, are highly invasive, and have a poor prognosis, while LGGs grow slowly and have relatively better outcomes. Multi-modal MRI is the mainstream approach for analyzing brain tumors [6], as different modalities provide diverse characteristics of tumor tissues, facilitating a comprehensive understanding of various tumor regions and aiding in accurate diagnosis and treatment planning.

Convolutional Neural Networks (CNNs) [7] are classic methods widely used for medical image segmentation, with popular models like U-Net [8], V-Net [9], and Attention U-Net [10] achieving remarkable results in various tasks. These models utilize convolutional layers to extract spatial features and employ encoder-decoder architectures to capture hierarchical information, making them effective for segmenting anatomical structures or pathological regions.

Despite their success in medical image segmentation, CNNs have notable limitations. CNNs rely on local receptive fields, making it challenging to capture long-range dependencies and global contextual information, which are crucial for accurately segmenting complex and heterogeneous tumor regions.

To address these challenges, Transformer-based models have emerged as a powerful alternative. Transformers [11] utilize a self-attention mechanism to capture global dependencies by considering relationships between all pixels or voxels simultaneously. However, the standard self-attention mechanism has a quadratic computational complexity concerning the input size, leading to a large number of parameters and high computational cost. This complexity makes it challenging to deploy Transformer models on mobile or resource-constrained medical devices.

To address these limitations, researchers have explored alternative approaches like State Space Models (SSMs). SSMs have recently garnered significant attention for their ability to model long-range dependencies while maintaining linear computational complexity. Unlike Transformers, which face challenges due to their quadratic complexity, SSMs offer a more efficient approach to processing sequential data, making them suitable for large-scale medical image segmentation.

Mamba [12], as a representative SSM, has been applied successfully across various domains like natural language processing and computer vision. Leveraging the strengths of Mamba, several medical image segmentation models have emerged. For example, U-Mamba [13] integrates a hybrid CNN-SSM block, effectively combining local feature extraction from CNNs and long-range contextual modeling from SSMs. This model has demonstrated effectiveness in multiple medical segmentation tasks. VM-UNet [14] is a medical image segmentation model based on State Space Models (SSMs). Unlike CNNs, which struggle with long-range dependencies, and Transformers with high computational complexity, VM-UNet uses the Visual State Space (VSS) block to capture extensive contextual information efficiently. Its asymmetrical encoder-decoder structure reduces convolutional layers, lowering computational costs while ensuring accurate segmentation.

While the aforementioned methods effectively leverage the strengths of State Space Models (SSMs) to capture long-range dependencies, there is still room for optimization in terms of lightweight design. To address this, we introduce MambaFormer3D, a lightweight hybrid architecture that fully integrates the efficient sequential processing capabilities of Mamba, the local receptive field advantages of CNNs, and the global contextual modeling power of Transformers. By combining these components, MambaFormer3D not only enhances segmentation performance but also reduces computational complexity, making it a more efficient solution for high-resolution 3D medical image segmentation tasks.

2. Method

2.1. Architecture Overview

This paper proposes an efficient 3D medical image segmentation network, MambaFormer3D, whose overall architecture is illustrated in Figure 1. The network is designed based on the Multi-scale Adaptive Lightweight Representation (MALR) unit and Linear Mamba Attention (LMA), aiming to reduce computational complexity while improving the segmentation accuracy of 3D medical images.

In the encoder path, multi-modal MRI inputs are first converted into 3D patches via Patch Embedding. Subsequently, the MALR unit integrates the strengths of Mamba and Transformer architectures. Within MALR, the traditional self-attention mechanism is replaced by the LMA module to reduce computational complexity while maintaining efficient global information capture capabilities. After multiple downsampling operations, the feature map size gradually decreases, and the channel dimension increases to extract richer semantic information.

Skip connections employ a layer-wise feature fusion strategy to preserve low-level spatial details and combine them with high-level semantic information, thereby enhancing feature representation. In the decoder path, feature recovery is achieved through a combination of transposed convolution and MALR units. The upsampled features are concatenated with skip connection features to improve segmentation precision. Additionally, residual blocks in the skip connections further optimize gradient flow, enhancing network stability and convergence speed.

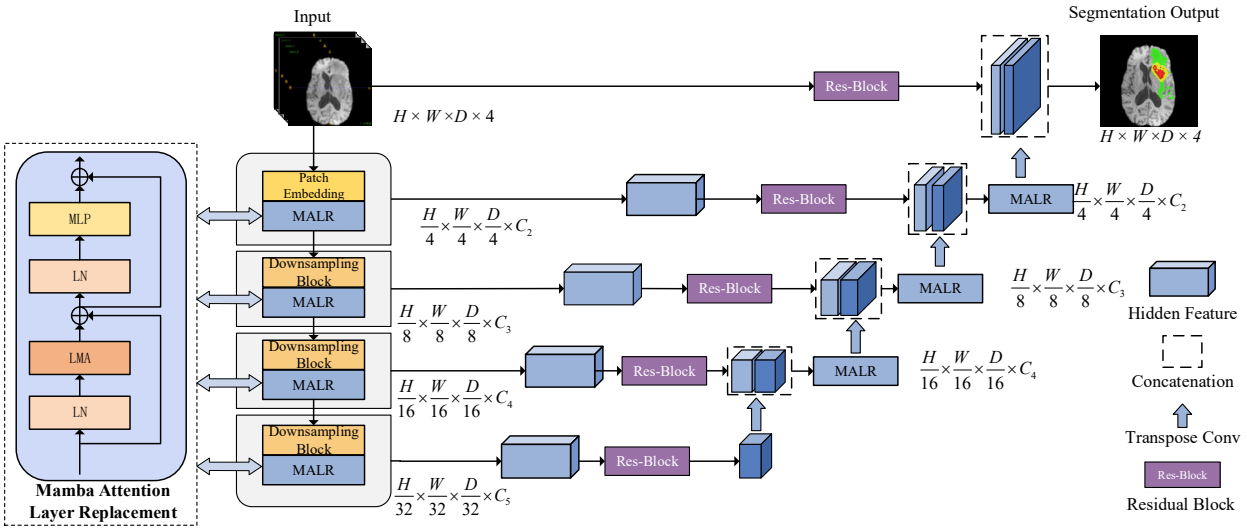

Figure 1. Overall framework of MambaFormer3D

2.2. MALR

First, for the preprocessed multi-modal MRI data of brain tumors (e.g., BraTS dataset), the four modalities (T1, T1ce, T2, FLAIR) are concatenated into multi-channel images and fed into the Patch Embedding layer. The input image is represented as $X \in \mathbb{R}^{H \times W \times D}$, where H, W, and D denote the height, width, and depth of the input data, respectively. Following Swin Transformer, the input is divided into non-overlapping 3D patches, each with a resolution of 4×4×4. This division yields a sequence length of $N = H/4 \times W/4 \times D/4$. After projection to the C-dimensional channel space, the resulting feature map has a size of $H/4 \times W/4 \times D/4 \times C$.

As illustrated in Figure 2, the MALR module takes the feature map matrix X as input. First, spatial feature extraction is performed using a 3×3×3 depthwise separable convolution, which decomposes standard convolution into depthwise and pointwise convolutions. This maintains the 3×3×3 local receptive field while significantly reducing parameter count and computational costs. Subsequently, Layer Normalization (LN) is applied to standardize feature distributions by normalizing feature statistics along the channel dimension, effectively mitigating internal covariate shift and improving training stability.

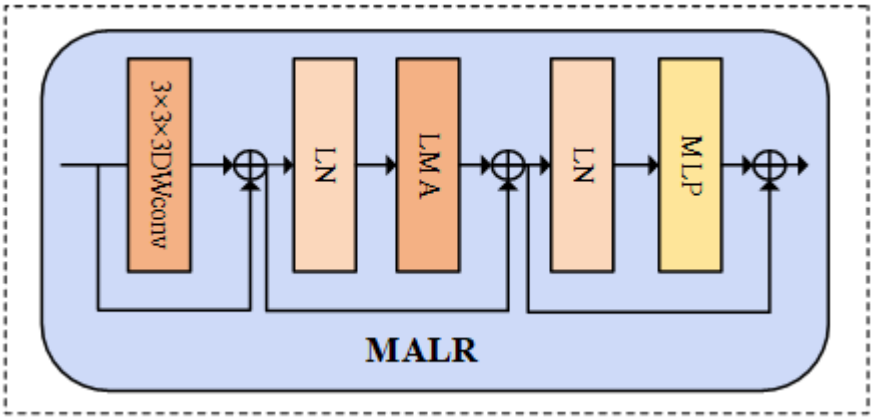

Figure 2. MALR structure

On the normalized features, a two-stage multilayer perceptron (MLP) is employed for cross-channel information fusion. The first fully connected (FC) layer expands the channel dimension by a factor of 4, followed by a GELU activation function. The second FC layer compresses the channel dimension back to its original size. This "expansion-compression" structure enhances the model's capability to represent complex patterns. Notably, the module innovatively introduces Linear Mamba Attention (LMA) as a global context modeling unit. Rooted in state space model theory, LMA achieves long-range dependency modeling with linear time complexity O(N), drastically reducing computational overhead compared to traditional self-attention mechanisms. The computational formulation of MALR is as follows:

$$Y = \mathrm{D}WConv(X) + X \tag{1}$$

$$Z = LMA(LN(Y)) + Y \tag{2}$$

$$W = MLP(LN(Z)) + \mathrm{Z} \tag{3}$$

2.3. LMA

As shown in Figure 3, LMA is a module based on the linear Mamba attention mechanism. The module makes use of Mamba's Selective State-Space Model (SSM) to effectively overcome the squared complexity problem ($O(N^2)$) of Transformer's traditional self-attention computation, while maintaining the capability of long-range dependency modelling. This makes LMA more efficient in processing large-scale 3D medical image data, ensuring accurate feature extraction while reducing computational cost. In the LMA structure, the input feature $X \in \mathbb{R}^{H\times W\times D\times \mathrm{C}}$ is first mapped into a query Q, a key K, and a value V ($Q, K, V \in \mathbb{R}^{HWD\times \mathrm{C}}$) through a linear layer. As shown in Eq. (4), the key K is compressed to a low-dimensional representation by Global Pooling operation to reduce the computational complexity:

$$K_{pool} = Global\ Pooling(K) \tag{4}$$

This operation effectively reduces computational demands, making attention calculation more efficient. The attention computation is formulated as follows (Equation 5):

$$Attention = Softmax\left(Q \cdot {K_{pool}}^{T}\right) \odot VSS(V) + X \tag{5}$$

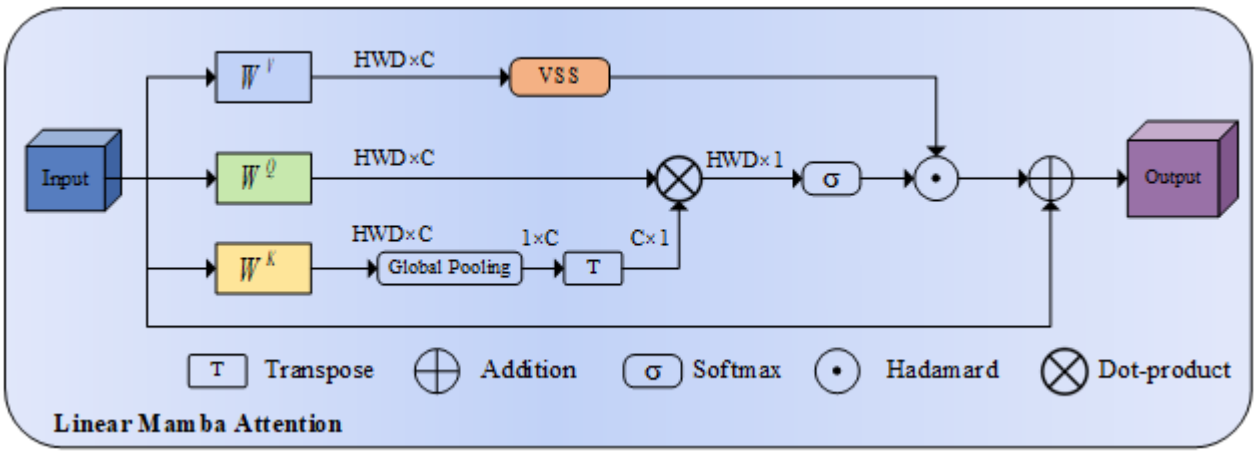

Figure 3. LMA structure

Where ⊙ denotes element-wise multiplication. The Softmax function is applied to the spatial attention scores derived from $Q \cdot K_{pool}$, ensuring normalized positive attention weights that sum to 1 at each spatial position. The VSS (Variable State-Space) operation further enhances global feature modeling capabilities, enabling the model to effectively learn long-range dependencies without increasing computational overhead. Its architecture is illustrated in Figure 4.

For VSS, the input with dimensions (B, N, C) is first mapped to a new feature dimension space through a linear transformation. A 1D convolution layer is then applied to the input sequence data for multiplication operations. After obtaining intermediate features, the SiLU activation function is used for further processing. The SSM part, which is the core of the module, processes the data through a Selective State Space Model (SSM) to improve the model's efficiency. Finally, a linear layer maps the features to the final output dimension C.

By introducing the Selective State Space Model (SSM) and global pooling, LMA successfully reduces the quadratic complexity ($O(N^2)$) of traditional Transformers' attention computation to linear complexity (O(N)). This improvement significantly reduces computational costs and memory usage, enabling LMA to efficiently handle large-scale 3D medical image data while retaining the core advantage of Transformers in modeling long-range dependencies.

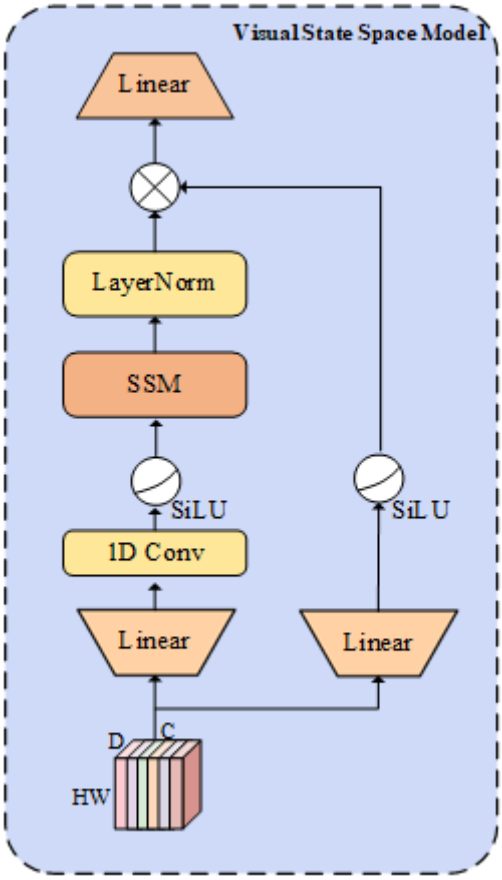

Figure 4. The VSS Modules

3. Experiments

3.1. Datasets

To objectively assess and compare brain tumor segmentation methods, we utilized the official BraTS 2019 datasets, which are widely recognized in brain tumor segmentation research. These datasets provide multi-modal MRI scans with manually annotated ground truth labels, making them well-suited for evaluating segmentation algorithms.

The BraTS 2018 dataset comprises 285 patients for training and 66 unlabeled patients for validation. The BraTS 2019 dataset includes 335 patients in the training set, consisting of 259 high-grade gliomas (HGG) and 76 low-grade gliomas (LGG), along with 125 unlabeled patients in the validation set. Each training sample includes four MRI modalities — T1, T1c, T2, and FLAIR — that capture various aspects of brain tumors. The scans have uniform dimensions of 240 × 240 × 155 mm^3 with a voxel spacing of 1 × 1 × 1 mm^3.

3.2. Evaluation indicators

Brain tumor segmentation is a vital task in medical imaging, requiring accurate and dependable techniques to outline tumor sub-regions. Benchmark datasets like BraTS 2018, BraTS 2019 offer manually labeled ground truth for four categories: Background (label 0), Necrosis and Non-Enhancing Tumor (NCR/NET, label 1), Peritumoral Edema (ED, label 2), and Enhancing Tumor (ET, label 4). For evaluation purposes, these labels are reorganized into three main regions: Whole Tumor (WT, labels 1, 2, and 4), Tumor Core (TC, labels 1 and 4), and Enhancing Tumor (ET, label 1).

To assess segmentation performance, two commonly used metrics are the Dice Similarity Coefficient (Dice score) and the 95% Hausdorff Distance (HD95). The Dice score evaluates the overlap between predicted and true segmentations, while HD95 measures the maximum boundary error, offering complementary perspectives on segmentation accuracy and boundary alignment. The formulas for these metrics are expressed as follows:

$$\text{Dice} = \frac{2|A \cap B|}{|A|+|B|} = \frac{2TP}{2TP+FP+FN} \tag{6}$$

$$\text{HD95} = \max\{\sup_{p \in B} \inf_{q \in A} d(p,q), \sup_{q \in A} \inf_{p \in B} d(p,q)\} \tag{7}$$

Let A and B represent the ground truth and predicted segmentation sets, respectively. The symbols TP, FP, TN, and FN indicate the voxel counts corresponding to true positives, false positives, true negatives, and false negatives.

3.3. Ablation Studies

This section validates the effectiveness of the MALR (Multi-scale Adaptive Local-Context Refinement) module in MambaFormer3D through ablation experiments on encoder-decoder configurations. As shown in Table 1, the complete Encoder + Decoder + MALR model (19.94M parameters) achieves state-of-the-art performance on the BraTS 2019 dataset, with a mean Dice coefficient of 0.915 and HD95 of 3.129 mm, surpassing Swin UNETR (62.83M parameters) by

+1.0% in Dice and −21.9% in HD95. When applying MALR only to the encoder (Encoder + MALR, 13.86M parameters), the Dice score improves by 0.8%, and HD95 decreases by 16.5%, demonstrating enhanced local-context modeling capabilities. Integrating MALR into the decoder (Decoder + MALR, 14.97M parameters) further reduces HD95 to 3.381 mm, though with a slight Dice drop (−0.2%), indicating a trade-off between feature resolution and semantic retention. The full model achieves Dice scores of 0.911 (ET), 0.917 (WT), and 0.916 (TC), surpassing Swin UNETR by 0.6%, 0.9%, and 2.2%, respectively, with a notable 1.281 mm reduction in HD95 for tumor core (TC) boundaries.

Notably, MambaFormer3D achieves superior performance with only 31.7% of the parameters of Swin UNETR, demonstrating the efficiency of its hierarchical lightweight design. These results validate the MALR module's ability to adaptively capture multi-scale local context, effectively enhancing the segmentation accuracy for complex boundaries while maintaining computational efficiency. The combination of a lightweight architecture and advanced local-context modeling positions MambaFormer3D as a promising approach for resource-constrained medical applications.

Table 1. Ablation experiments on the BraTS 2019 dataset

Model	Params(M)	Dice	HD95
Swin UNETR	62.83	0.905	4.008
Encoder+MALR	13.86	0.913	3.345
Dncoder+MALR	5.97	0.911	3.381
Encoder and Dncoder+MALR	14.94	0.915	3.129

3.4. Comparative Experiment

In this section, the MambaFormer3D algorithm is compared with the current mainstream brain tumor image segmentation methods, including VNet, UNETR [15], Swin UNETR [16], ADHDC-Net [17], and TransBTS [18], to comprehensively evaluate the performance of each algorithm on the BraTS 2019 dataset. In order to ensure the fairness and scientificity of the comparison, Dice similarity coefficient (Dice), 95% Hausdorff distance (HD95), number of parameters and computational complexity (GFLOPs) are used as evaluation indexes in this section. The specific experimental results are shown in Table 2 and Table 3:

Table 2. Segmentation results of various models on the BraTS 2019 dataset (Dice).

Method	Dice_WT	Dice_TC	Dice_ET	Dice_ave
VNet	0.874	0.883	0.870	0.875
UNETR	0.915	0.895	0.896	0.902
Swin UNETR	0.919	0.899	0.896	0.905
ADHDC-Net	0.785	0.860	0.834	0.826
TransBTS	0.827	0.841	0.839	0.836
MambaFormer3D	0.917	0.916	0.911	0.915

Table 3. Segmentation results of various models on the BraTS 2019 dataset (HD95).

Method	HD95_WT	HD95_TC	HD95_ET	HD95_ave
VNet	6.453	4.500	2.490	4.481
UNETR	5.322	5.075	2.626	4.341
Swin UNETR	4.465	4.410	3.149	4.008
ADHDC-Net	12.703	10.072	4.798	9.191
TransBTS	12.584	16.852	13.083	14.173
MambaFormer3D	3.509	3.276	2.602	3.129

In Table 2, the Dice coefficients of MambaFormer3D are 0.917, 0.916, and 0.911 in the WT, TC, and ET regions, respectively, with an average Dice coefficient of 0.915, which is significantly higher than the other compared methods. Compared with VNet, MambaFormer3D improves the Dice values in the WT, TC and ET regions by 4.3%, 3.3% and 4.1%, respectively. MambaFormer3D improves the average Dice value by 1.3% compared with UNETR. MambaFormer3D improves the average Dice value by 1.2% compared to Swin UNETR. In Table 3, MambaFormer3D also has a significant advantage in HD95 distance, especially outperforming other methods in the WT, TC and ET regions. Compared with VNet, MambaFormer3D reduces the HD95 distance in the WT, TC and ET regions by 2.944 mm, 1.224 mm and 0.112 mm, respectively. Compared with UNETR, MambaFormer3D reduces the HD95 distance in the WT, TC and ET regions by 1.212 mm, 1.799 mm and 0.024 mm. Compared with Swin UNETR, the HD95 distance of MambaFormer3D is reduced by 0.956 mm, 1.134 mm and 0.547 mm in the WT, TC and ET regions, respectively.

3.5. Visualization

As shown in Figure 5, this paper compares the brain tumor segmentation effect of MambaFormer3D with other models (VNet, TransBTS, Swin UNETR, UNETR) on the BraTS 2019 dataset using ITK-SNAP visualization software.

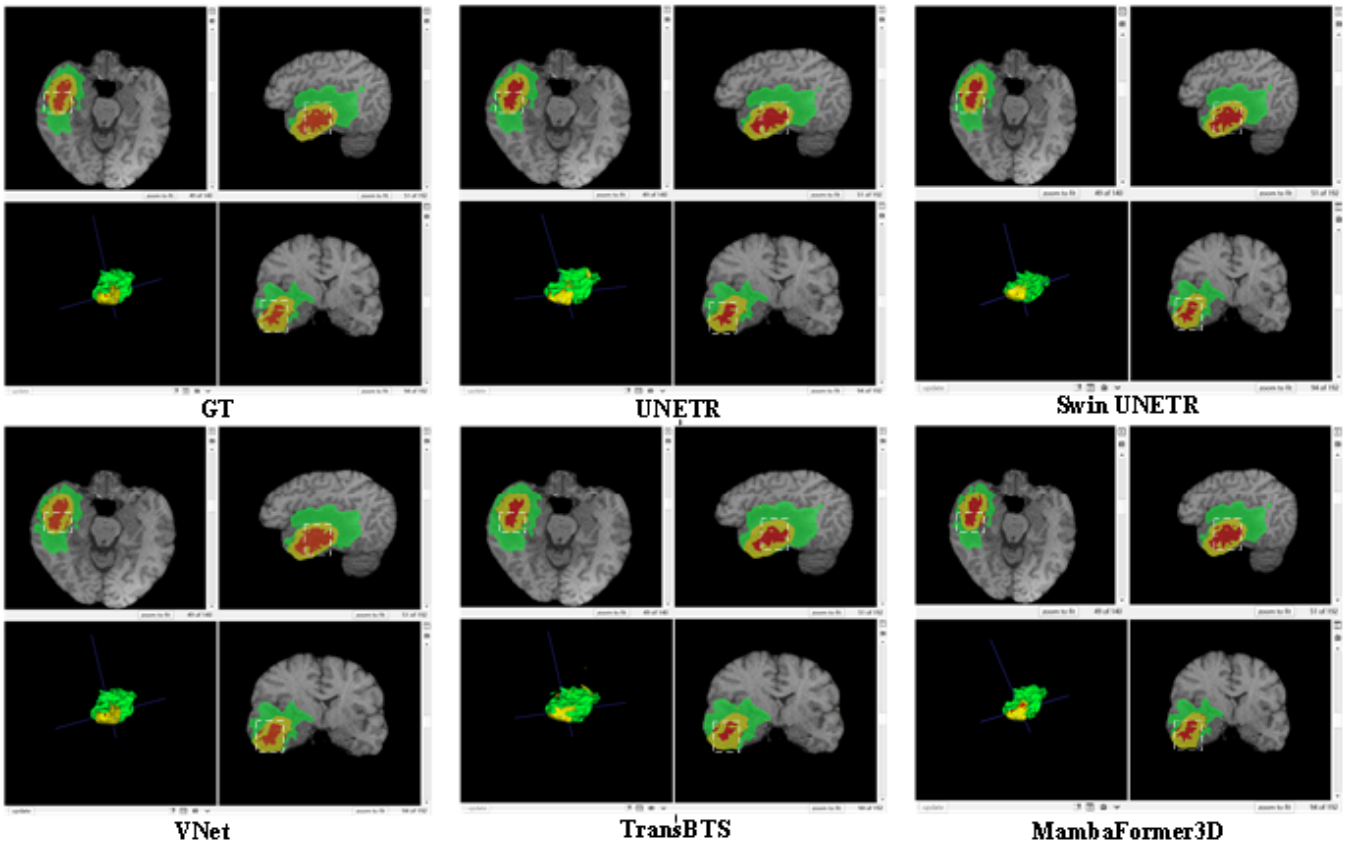

Figure 5. Comparison of segmentation mask and label mask for each algorithm

From the comparison of the generated masks in the dotted box with the ground truth mask (GT), it is obvious that MambaFormer3D outperforms the other models in segmentation accuracy in several brain tumor regions. Especially in the tumor and peri-tumor edema regions, MambaFormer3D is able to accurately capture the subtle structures and boundaries of the tumor, demonstrating its excellent ability in handling complex tumor regions. Compared with other models, MambaFormer3D is able to better maintain the coherence of tumor boundaries, avoiding region breakage or overlap due to mis-segmentation. In addition, MambaFormer3D is also more accurate in segmenting the edema region, especially in the transition area between tumor and edema, and is able to distinguish these regions meticulously, which demonstrates its outstanding advantages in multimodal brain tumor image processing. These results show that MambaFormer3D can better cope with the complex scenarios in brain tumor segmentation tasks, especially in the detail processing and accurate positioning of boundaries, than other existing mainstream methods.

4. Conclusions

The MambaFormer3D model proposed in this paper successfully combines the selective state-space model (SSM) with the linear attention mechanism to achieve both long-range dependency modeling and optimize the computational complexity. Through the synergy of MALR and LMA modules, MambaFormer3D significantly improves the performance of medical image segmentation tasks, especially in brain tumor segmentation, while ensuring high efficiency. The experimental results show that MambaFormer3D has excellent performance in the processing of multimodal MRI data and exhibits greater potential in environments with limited computational resources.

Acknowledgements

Key scientific and technological projects in Henan province (242102211042)

Doctoral Fund Project of Henan Polytechnic University (B2022-14)

References

[1] Ramesh K K D, Kumar G K, Swapna K, et al. A review of medical image segmentation algorithms[J]. EAI Endorsed Transactions on Pervasive Health & Technology, 2021, 7(27).

[2] Abidin Z U, Naqvi R A, Haider A, et al. Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey[J]. Frontiers in Bioengineering and Biotechnology, 2024, 12: 1392807.

[3] Zhang S, Edwards A, Wang S, et al. A prior knowledge based tumor and tumoral subregion segmentation tool for pediatric brain tumors[J]. arXiv preprint arXiv:2109.14775, 2021.

[4] Kaifi R. A review of recent advances in brain tumor diagnosis based on AI-based classification[J]. Diagnostics, 2023, 13(18): 3007.

[5] Malmer B, Henriksson R, Grönberg H. Different aetiology of familial low-grade and high-grade glioma? A nationwide cohort study of familial glioma[J]. Neuroepidemiology, 2002, 21(6): 279-286.

[6] Zhang W, Wu Y, Yang B, et al. Overview of multi-modal brain tumor mr image segmentation[C]//Healthcare. MDPI, 2021, 9(8): 1051.

[7] O'shea K, Nash R. An introduction to convolutional neural networks[J]. arXiv preprint arXiv:1511.08458, 2015.

[8] Ronneberger, Olaf et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” ArXiv abs/1505.04597 (2015): n. pag.

[9] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-571.

[10] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[12] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.

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

[14] Ruan J, Li J, Xiang S. Vm-unet: Vision mamba unet for medical image segmentation [J]. arXiv preprint arXiv: 2402.02491, 2024.

[15] Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 574-584.

[16] Hatamizadeh A, Nath V, Tang Y, et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images[C]//International MICCAI brainlesion workshop. Cham: Springer International Publishing, 2021: 272-284.

[17] Liu H, Huo G, Li Q, et al. Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation [J]. Expert Systems with Applications, 2023, 214: 119166.

[18] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer(2021)