SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation

Zhaohu Xing[1], Tian Ye[1], Yijun Yang[1], Guang Liu[2], and Lei Zhu[1,3] (✉)

[1] The Hong Kong University of Science and Technology (Guangzhou)
zxing565@connect.hkust-gz.edu.cn
[2] Beijing Academy of Artificial Intelligence
[3] The Hong Kong University of Science and Technology

Abstract. The Transformer architecture has demonstrated remarkable ability in modeling global relationships. However, it presents a significant computational challenge when processing high-dimensional medical images. Mamba, as a State Space Model (SSM), has recently emerged as a notable approach for modeling long-range dependencies in sequential data, excelling in the field of natural language processing with its remarkable memory efficiency and computational speed. Inspired by its success, we introduce **SegMamba**, a novel 3D medical image **Seg**mentation **Mamba** model, designed to effectively capture long-range dependencies within whole-volume features at every scale. Our SegMamba, in contrast to Transformer-based methods, excels in whole-volume feature modeling, maintaining superior processing speed, even with volume features at a resolution of $64 \times 64 \times 64$ (The sequential length is about 260k). Comprehensive experiments on three datasets demonstrate the effectiveness and efficiency of our SegMamba. Additionally, to facilitate research in 3D colorectal cancer (CRC) segmentation, we contribute a new large-scale dataset (named CRC-500). The code for SegMamba and information about CRC-500 dataset are available at: https://github.com/gexing/SegMamba.

Keywords: State space model · Mamba · Long-range sequential modeling · 3D medical image segmentation.

1 Introduction

3D medical image segmentation is is an essential task. Accurate segmentation results can reduce the diagnostic burden of diseases for doctors. To improve the segmentation performance, extending model's receptive field is a critical aspect in this task. Conventional convolutional neural networks (CNNs) are not very effective at extracting large range information from high-resolution 3D medical images. Hence, the large-kernel convolution [16] is proposed to model a broader range of features. 3D UX-Net [12] introduces a new architecture, utilizing an convolution block with a large kernel size ($7 \times 7 \times 7$) to facilitate larger receptive

fields. However, CNN-based methods struggle to model global relationships due to the locality of the convolution layer.

Recently, the transformer architecture [22,2,23], utilizing a self-attention module to extract global information, has been extensively explored for 3D medical image segmentation. UNETR [7] employs the Vision Transformer (ViT) [3] as its encoder to learn contextual information, which is then merged with a CNN-based decoder via skip connections at multiple resolutions. SwinUNETR [6] leverages the SwinTransformer [15] as the encoder to extract multi-scale features. It also designs a multi-scale decoder to fuse features from each encoder stage, achieving promising results in 3D medical image segmentation. However, the typically high resolution of 3D medical images can result in significant computational burdens and reduced speed performance for transformer-based methods.

To overcome the challenges of long sequence modeling, Mamba [5], which originates from state space models (SSMs) [10], is designed to model long-range dependencies and enhance the efficiency of training and inference through a selection mechanism and a hardware-aware algorithm. Numerous studies have explored the applications of Mamba in computer vision (CV). U-Mamba [17] integrates the Mamba layer into the encoder of nnUNet [9] to enhance general medical image segmentation. Meanwhile, Vision Mamba [24] introduces the Vim block, which incorporates bidirectional SSM for data-dependent global visual context modeling and position embeddings for location-aware visual understanding. Additionally, VMamba [14] designs a CSM module to bridge the gap between 1-D array scanning and 2-D plain traversing. However, these methods are not specifically designed for 3D medical image segmentation.

In this paper, we introduce SegMamba, a novel architecture that combines the U-shape structure with Mamba for modeling the whole volume global features at various scales. To our knowledge, this is the first method utilizing Mamba specifically for 3D medical image segmentation. To facilitate the use of Mamba on high-dimensional medical images, we design a tri-orientated Mamba (ToM) module to enhance the sequential modeling of 3D features from three directions. Subsequently, to effectively model the spatial features, we further design a gated spatial convolution (GSC) module to enhance the feature representation in the spatial dimension before each ToM module. Moreover, datasets play an important role in 3D medical imaging. We propose a new large-scale dataset for 3D colorectal cancer segmentation called CRC-500, which consists of 500 3D computed tomography (CT) scans with expert annotations. The dataset will be made available upon request for research purposes. SegMamba exhibits a remarkable capability to model long-range dependencies within volumetric data, while maintaining outstanding inference efficiency, compared to traditional CNN-based and transformer-based methods. Extensive experiments demonstrate the effectiveness of our method.

2 Colorectal Cancer Segmentation Dataset (CRC-500)

Necessity for CRC-500 Dataset Colorectal cancer (CRC) is the third most common cancer worldwide among men and women, the second leading cause

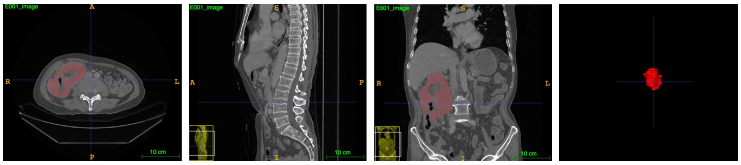

Fig. 1. The data visualization for CRC-500 dataset.

Table 1. Comparison between related datasets and our CRC-500 dataset.

Related Datasets	Rectal Cancer	Colon Cancer	Volume Number	Open-sourced
3D RU-Net [19]	✓	✓	64	✗
MSDenseNet [12]	✓	✓	43	✗
MSD [21]	✗	✓	190	✓
Zhang et al. [7]	✓	✓	388	✗
Our CRC-500	✓	✓	500	✓

of death related to cancer, and the primary cause of death in gastrointestinal cancer [4]. Using deep learning methods to detect the cancer region can assist doctors in making more accurate diagnoses. However, as shown in Table 1, the current 3D colorectal cancer segmentation datasets are small in size. Moreover, only the MSD dataset is publicly available, but it lacks data on rectal cancer. To facilitate research in the 3D colorectal cancer segmentation field, we contribute a new large-scale dataset (named CRC-500). This dataset consists of 500 3D colorectal volumes with corresponding precise annotations from experts. Fig. 1 presents examples in 2D format from our proposed CRC-500 dataset. The details of our CRC-500 will be discussed below.

Dataset Construct The CT scans were acquired from January 2008 to April 2020. All sensitive patient information has been removed. Each volume was annotated by a professional doctor and calibrated by another professional doctor.

Dataset Analysis All the CT scans share the same in-plane dimension of 512×512, and the dimension along the z-axis ranges from 94 to 238, with a median of 166. The in-plane spacing ranges from 0.685×0.685 mm to 0.925×0.925 mm, with a median of 0.826×0.826 mm, and the z-axis spacing is from 3.0 mm to 3.75 mm, with a median of 3.75 mm.

3 Method

SegMamba mainly consists of three components: 1) the 3D feature encoder with multiple tri-orientated spatial Mamba blocks to model the global information at different scales, 2) the 3D decoder based on the convolution layer for predicting segmentation results, and 3) the skip-connections to connect the global multi-scale features to the decoder for feature reuse. Fig. 2 illustrates the overview of the proposed SegMamba. We further describe the details of the encoder and decoder in this section.

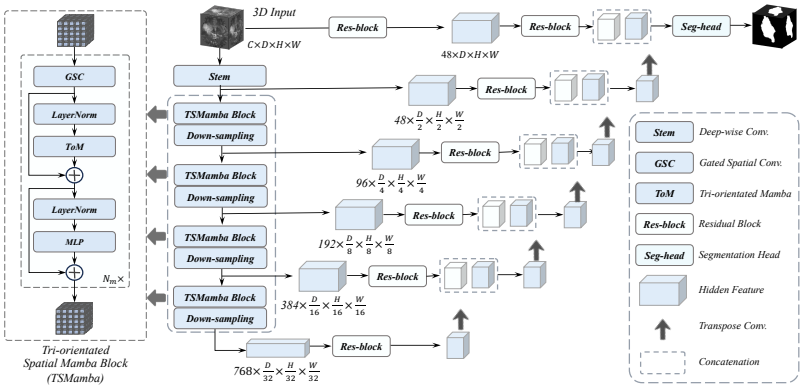

Fig. 2. The overview of the proposed SegMamba.

3.1 Tri-orientated Spatial Mamba (TSMamba) Block

Modeling global features and multi-scale features is critically important for 3D medical image segmentation. Transformer architectures can extract global information, but it incurs a significant computational burden when dealing with overly long feature sequences. To reduce the sequence length, methods based on Transformer architectures, such as UNETR, directly down-sample the 3D input with a resolution of $D \times H \times W$ to $\frac{D}{16} \times \frac{H}{16} \times \frac{W}{16}$. However, this approach limits the ability to encode multi-scale features, which are essential for predicting segmentation results via the decoder. To overcome this limitation, we design a TSMamba block to enable both multi-scale and global feature modeling while maintains a high efficiency during training and inference.

As illustrated in Fig. 2, the encoder consists of a stem layer and multiple TSMamba blocks. For the stem layer, we employ a depth-wise convolution with a large kernel size of $7\times7\times7$, with a padding of $3\times3\times3$, and a stride of $2\times2\times2$. Given a 3D input volume $I \in \mathbb{R}^{C\times D\times H\times W}$, where C denotes the number of input channels, the first scale feature $z_0 \in \mathbb{R}^{48\times\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$ is extracted by the stem layer. Then, z_0 is fed through each TSMamba block and corresponding down-sampling layers. For the m^{th} TSMamba block, the computation process can be defined as:

$$\hat{z}_m^l = GSC(z_m^l), \quad \tilde{z}_m^l = ToM\left(\mathrm{LN}\left(\hat{z}_m^l\right)\right) + \hat{z}_m^l, \quad z_m^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\tilde{z}_m^l\right)\right) + \tilde{z}_m^l, \tag{1}$$

where the GSC and ToM denote the proposed gated spatial convolution module and tri-orientated Mamba module, respectively, which will be discussed next. $l \in \{0, 1, ..., N_m - 1\}$, LN denotes the layer normalization, and MLP represents the multiple layers perception layer to enrich the feature representation.

Gated Spatial Convolution (GSC) The Mamba layer models the feature dependencies by flattening the 3D features into a 1D sequence. Hence, to extract the spatial relationship before the Mamba layer, we design a gated spatial convolution (GSC) module. As shown in Fig. 3 (a), the input 3D features are fed into two convolution blocks (a convolution block contains a norm, a convolution, and

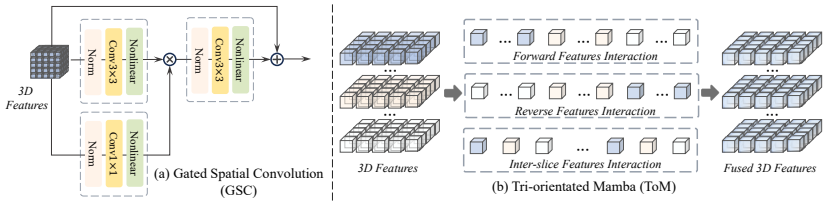

Fig. 3. (a) The gated spatial convolution. (b) The tri-orientated Mamba.

a nonlinear layer), with the convolution kernel sizes being $3 \times 3 \times 3$ and $1 \times 1 \times 1$. Then these two features are multiplied pixel-by-pixel to control the information transmission similar to the gate mechanism [13]. Finally, a convolution block is used to further fuse the features, while a residual connection is utilized to reuse the input features.

$$GSC(z) = z + C^{3\times3\times3}(C^{3\times3\times3}(z) \cdot C^{1\times1\times1}(z)), \tag{2}$$

where z denotes the input 3D features and C denotes the convolution block.

Tri-orientated Mamba (ToM) In TSMamba block, to effectively model the global information of high-dimensional features, we design a tri-orientated Mamba module that computes the feature dependencies from three directions. As shown in Fig. 3 (b), we flatten the 3D input features into three sequences to perform the corresponding feature interaction and obtain the fused 3D features.

$$ToM(z) = Mamba(z_f) + Mamba(z_r) + Mamba(z_s), \tag{3}$$

where $Mamba$ is the Mamba layer to model the global information within a sequence, f denotes forward direction, r denotes reverse direction, and s denotes inter-slice direction.

3.2 Decoder

Our feature encoder, based on the TSMamba block, extracts the multi-scale features. Following many previous studies [12,6,7], we utilize a CNN-based decoder and a skip connections for predicting the segmentation results.

4 Experiments

4.1 Other Dataset

BraTS2023 dataset The BraTS2023 dataset [18,1,11] contains a total of 1,251 3D brain MRI volumes. Each volume includes four modalities (namely T1, T1Gd, T2, T2-FLAIR) and three segmentation targets (WT: Whole Tumor, ET: Enhancing Tumor, TC: Tumor Core).

AIIB2023 dataset The AIIB2023 dataset [20], the first open challenge and publicly available dataset for airway segmentation. The released data include 120 high-resolution computerized tomography (HRCT) scans with precise expert annotations, providing the first airway reference for fibrotic lung disease.

Table 2. Quantitative comparison on BraTS2023 and AIIB2023 datasets. The bold value denotes the best performance.

Methods	BraTS2023								AIIB2023		
	WT		TC		ET		Avg		Airway Tree		
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	IoU ↑	DLR ↑	DBR ↑
SegresNet [19]	92.02	4.07	89.10	4.08	83.66	3.88	88.26	4.01	87.49	65.07	53.91
UX-Net [12]	93.13	4.56	90.03	5.68	85.91	4.19	89.69	4.81	87.55	65.56	54.04
MedNeXt [21]	92.41	4.98	87.75	4.67	83.96	4.51	88.04	4.72	85.81	57.43	47.34
UNETR [7]	92.19	6.17	86.39	5.29	84.48	5.03	87.68	5.49	83.22	48.03	38.73
SwinUNETR [6]	92.71	5.22	87.79	4.42	84.21	4.48	88.23	4.70	87.11	63.31	52.15
SwinUNETR-V2 [8]	93.35	5.01	89.65	4.41	85.17	4.41	89.39	4.51	87.51	64.68	53.19
Ours	**93.61**	**3.37**	**92.65**	**3.85**	**87.71**	**3.48**	**91.32**	**3.56**	**88.59**	**70.21**	**61.33**

Table 3. Quantitative comparison on CRC-500 dataset.

Methods	Dice ↑	HD95 ↓
SegresNet [19]	46.10	34.97
UX-Net [12]	45.73	49.73
MedNeXt [21]	35.93	52.54
UNETR [7]	33.70	61.51
SwinUNETR [6]	38.36	55.05
SwinUNETR-V2 [8]	41.76	58.05
Ours	**48.02**	**30.89**

Table 4. Ablation study for different modules on CRC-500 dataset. LC denotes large-kernel convolution layer.

Methods	Modules			Dice ↑	HD95 ↓
	LC	GSC	ToM		
UX-Net [12]	✓			45.73	49.73
M1				45.34	43.01
M2		✓		46.65	37.01
M3			✓	47.22	33.32
Ours	✓		✓	**48.02**	**30.89**

4.2 Implementation Details

Our model is implemented in Pytorch 2.0.1-cuda11.7 and Monai 1.2.0. During training, we use a random crop size of $128 \times 128 \times 128$ and a batch size of 2 per GPU for each dataset. We use cross-entropy loss for all experiments and an SGD optimizer along with a polynomial learning rate scheduler (initial learning rate of 1e-2, a decay of 1e-5). We run 1000 epochs for all datasets and adopt the following data augmentations: additive brightness, gamma, rotation, scaling, mirror, and elastic deformation. All experiments are conducted on a cloud computing platform with four NVIDIA A100 GPUs. For each dataset, we randomly allocate 70% of the 3D volumes for training, 10% for validation, and the remaining 20% for testing.

4.3 Comparison with SOTA Methods

We compare SegMamba against six SOTA segmentation methods, including three CNN-based methods (SegresNet [19], UX-Net [12], MedNeXt [21]), and three transformer-based methods (UNETR [7], SwinUNETR [6], and SwinUNETR-V2 [8]). For a fair comparison, we utilize public implementations of these methods to retrain their networks, thereby generating their best segmentation results. The

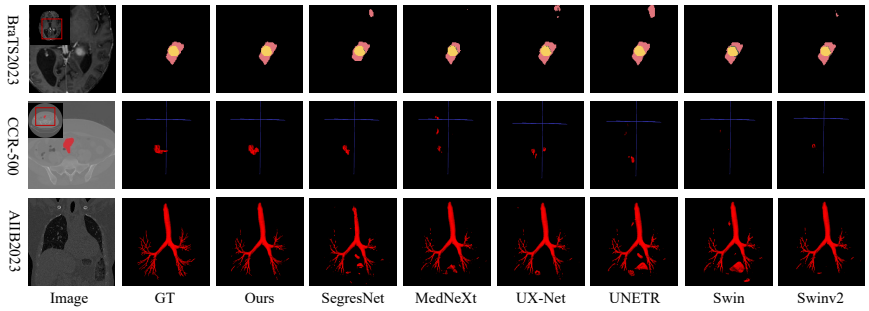

Fig. 4. Visual comparisons of proposed SegMamba and other state-of-the-art methods. Swin denotes SwinUNETR and Swinv2 denotes SwinUNETR-V2.

Table 5. Ablation study for different global modeling modules. TM denotes training memory, IM denotes inference memory, IT denotes inference time, and OOM represents out of memory.

Methods	Core module	Input resolution	Sequence length	TM (M)	IM (M)	IT (case/s)	Is Global
M4	Large-kernel convolution	128^3	262144	18852	5776	1.92	✗
M5	SwinTransformer	128^3	262144	34000	9480	1.68	✗
M6	Self-attention	128^3	262144	OOM	-	-	✓
Ours	TSMamba	128^3	262144	**17976**	6279	**1.51**	✓

Dice score (Dice) and 95% Hausdorff Distance (HD95) are adopted for quantitative comparison on BraTS2023 and CCR-500 datasets. Following [20], the Intersection over union (IoU), Detected length ratio (DLR), and Detected branch ratio (DBR) are adopted on AIIB2023 dataset.

BraTS2023 The segmentation results of gliomas for BraTS2023 dataset are listed in Table 2. UX-Net, a CNN-based method, achieves the best performance among the comparison methods, with an average Dice of 89.69% and an average HD95 of 4.81. In comparison, our SegMamba achieves the highest Dices of 93.61%, 92.65%, and 87.71%, and HD95s of 3.37, 3.85, and 3.48 on WT, TC, and ET, respectively, showing better segmentation robustness.

AIIB2023 For this dataset, the segmentation target is the airway tree, which includes many tiny branches and poses challenges in obtaining robust results. As shown in Table 2, our SegMamba achieves the highest IoU, DLR, and DBR scores of 88.59%, 70.21%, and 61.33%, respectively. This also indicates that our SegMamba exhibits better segmentation continuity compared to other methods.

CRC-500 The results on CRC-500 dataset are listed in Table 3. In this dataset, the cancer region is typically small; however, our SegMamba can accurately detect the cancer region and report the best Dice and HD95 scores of 48.02% and 30.89, respectively.

Visual Comparisons To compare the segmentation results of different methods more intuitively, we choose six comparative methods for visual comparison on three datasets. As depicted in Fig. 4, our SegMamba can accurately detect the boundary of each tumor region on BraTS2023 dataset. Similar to BraTS2023 dataset, our method accurately detects the cancer region on CRC-500 dataset. The segmentation results show better consistency compared to other state-of-the-art methods. Finally, on AIIB2023 dataset, our SegMamba can detect a greater number of branches in the airway and achieve better continuity.

4.4 Ablation Study

The Effectiveness of GSC and ToM modules As shown in Table 4, M1 is our basic method, which only contains the original Mamba layer. In M2, we introduce our GSC module. Compared to M1, M2 achieves the Dice of 46.65% and HD95 of 37.01, with an improvement of 2.88% and 13.95%. This demonstrates that the GSC module can improve the segmentation performance by modeling the spatial features before ToM module. Then, in M3, we introduce the ToM module, which model the global information from three directions. M3 reports the Dice and HD95 of 47.22% and 33.32, with an improvement of 1.22% and 9.97% compared to M2. Finally, our SegMamba introduce both GSC and ToM modules, achieving the state-of-the-art performance, with the Dice and HD95 of 48.02% and 30.89.

The high efficiency of TSMamba We verify the high efficiency of the TSMamba block through an ablation study presented in Table 5. M4 is UX-Net [12], which utilizes large-kernel convolution as its core module. M5 is SwinUNETR [6], which uses the SwinTransformer as its core module. Both improve receptive field by computing long range pixels, but they cannot compute the relationship within a global range. In M6, we use self-attention, a global modeling layer, as the core module, but it is infeasible due to the computational burden. In comparison, our method uses a Mamba-based global modeling module (TSMamba), and achieves a better training memory (TM) and inference time (IT), even though the maximum flattened sequence length reaches 260k.

5 Conclusion

In this paper, we propose the first general 3D medical image segmentation method, based on the Mamba, called SegMamba. We design a tri-orientated Mamba (ToM) module to enhance the sequential modeling of 3D features. Then, to effectively model the spatial features, we further design a gated spatial convolution (GSC) module to enhance the feature representation in the spatial dimension before each ToM module. Moreover, we propose a new large-scale dataset for 3D colorectal cancer segmentation called CRC-500, to facilitate the related research. SegMamba exhibits a remarkable capability to model long-range dependencies within volumetric data, while maintaining outstanding inference efficiency, compared to traditional CNN-based and transformer-based methods. Extensive experiments demonstrate the effectiveness of our method.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Granados-Romero, J.J., Valderrama-Treviño, A.I., Contreras-Flores, E.H., Barrera-Mera, B., Herrera Enríquez, M., Uriarte-Ruíz, K., Ceballos-Villalba, J.C., Estrada-Mata, A.G., Alvarado Rodríguez, C., Arauz-Peña, G.: Colorectal cancer: a review. Int J Res Med Sci **5**(11), 4667 (2017)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
8. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
11. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). ArXiv (2023)
12. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
13. Liu, H., Dai, Z., So, D., Le, Q.V.: Pay attention to mlps. Advances in Neural Information Processing Systems **34**, 9204–9215 (2021)
14. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)

16. Luo, P., Xiao, G., Gao, X., Wu, S.: Lkd-net: Large kernel convolution network for single image dehazing. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1601–1606. IEEE (2023)
17. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
19. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
20. Nan, Y., Xing, X., Wang, S., Tang, Z., Felder, F.N., Zhang, S., Ledda, R.E., Ding, X., Yu, R., Liu, W., et al.: Hunting imaging biomarkers in pulmonary fibrosis: Benchmarks of the aiib23 challenge. arXiv preprint arXiv:2312.13752 (2023)
21. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 140–150. Springer (2022)
24. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)