LKM-UNet: Large Kernel Vision Mamba UNet for Medical Image Segmentation

Jinhong Wang[1,2,5], Jintai Chen[3(✉)], Danny Chen[4], and Jian Wu[5,6]

[1] College of Computer Science and Technology, Zhejiang University, China
[2] Institute of Wenzhou, Zhejiang University, China
[3] Computer Science Department, University of Illinois Urbana-Champaign, USA
[4] Department of Computer Science and Engineering, University of Notre Dame, USA
[5] State Key Laboratory of Transvascular Implantation Devices of The Second Affiliated Hospital, Zhejiang University School of Medicine, China
[6] School of Public Health Zhejiang University, China
wangjinhong@zju.edu.cn, jtchen721@gmail.com, dchen@nd.edu, wujian2000@zju.edu.cn

Abstract. In clinical practice, medical image segmentation provides useful information on the contours and dimensions of target organs or tissues, facilitating improved diagnosis, analysis, and treatment. In the past few years, convolutional neural networks (CNNs) and Transformers have dominated this area, but they still suffer from either limited receptive fields or costly long-range modeling. Mamba, a State Space Sequence Model (SSM), recently emerged as a promising paradigm for long-range dependency modeling with linear complexity. In this paper, we introduce a **L**arge **K**ernel vision **M**amba **U**-shape **Net**work, or `LKM-UNet`, for medical image segmentation. A distinguishing feature of our `LKM-UNet` is its utilization of large Mamba kernels, excelling in locally spatial modeling compared to small kernel-based CNNs and Transformers, while maintaining superior efficiency in global modeling compared to self-attention with quadratic complexity. Additionally, we design a novel hierarchical and bidirectional Mamba block to further enhance Mamba's global and neighborhood spatial modeling capability for vision inputs. Comprehensive experiments demonstrate the feasibility and the effectiveness of using large-size Mamba kernels to achieve large receptive fields. Codes are available at https://github.com/wjh892521292/LKM-UNet.

Keywords: Medical Image Segmentation · UNet · Mamba.

1 Introduction

Efficiently segmenting biomedical objects of interest (*e.g.*, lesions) in large-size 2D/3D images significantly enhances downstream clinical practice and biomedical research. Currently, automatic segmentation models, leveraging popular deep

✉: Corresponding Author.

learning backbones such as convolutional neural networks (CNNs) and Transformers [19], have reduced manual annotation requirements but may incur considerable computational costs or overlook specific details. CNN-based models (e.g., UNet [18]) typically seek to extract global patterns by hierarchically stacking small kernels, excelling in pixel-level feature extraction but being ineffective in learning long-range dependencies due to their limited receptive fields [14]. Though recent investigations [3] have shown the effectiveness of large convolution kernels, it often requires specific optimization strategies and complicated inference-time model reformulation. In contrast, Transformer-based algorithms offer powerful long-range modeling but sacrifice pixel-level spatial modeling [21, 20]. Further, a key component, the self-attention module, incurs quadratic complexity and cannot handle excessive tokens [9], resulting in the need to pack pixels into windows and hence sacrificing resolution information. Especially, many studies have shown that Transformers achieve the best trade-off at a 7×7 window size as a smaller window causes more computation and a larger window causes a significant drop in performance [13, 1]. Later studies have investigated CNN-Transformer hybrid models or approaches for intra-patch locality modeling [20, 23]. However, due to the typically larger sizes of medical images compared to natural images, how to reduce the interaction complexity between global patches and how to enlarge the receptive field of local spatial modeling are still open.

Recently, structured state space sequence models (SSMs) [6] (e.g., Mamba [5]) have emerged as a powerful long-sequence modeling approach with linear complexity in terms of input size, shedding light on efficient modeling of both local and global dependencies. In contrast to the conventional self-attention approach, SSMs enable each element in a 1D array (e.g., a text sequence) to interact with any of the previously scanned samples through a compressed hidden state, effectively reducing the quadratic complexity to linear. SSMs were devised to address natural language processing (NLP) tasks [4], but also show effectiveness in computer vision [25]. For medical image segmentation, for example, U-Mamba [5] and SegMamba [22] introduced SSM-CNN hybrid models that directly apply Mamba to efficiently model long-range dependencies at the pixel-level. Although effective, the potential and deficiency of Mamba are still not yet fully explored and resolved, in three aspects. **First**, benefiting from its linear complexity, Mamba possesses more flexibility in space allocation. That is, unlike the stereotypes of previous methods based on small convolution kernels or size-constrained self-attention windows, Mamba is promising in endowing the model with an ability in large receptive field spatial modeling, which has been neglected in the known Mamba-based approaches. **Second**, Mamba is a unidirectional sequence modeling method that lacks positional awareness and focuses more on posterior tokens. **Third**, the original Mamba was proposed for 1D sequence modeling, which is not suitable for computer vision tasks that require spatial-aware understanding. Mainly due to local adjacent spatial pixels becoming discontinuous, Mamba can suffer forgetting problems and inefficient local modeling capabilities.

In this paper, we propose a **L**arge **K**ernel **M**amba **U**-shape **Net**work (`LKM-UNet`) for 2D and 3D medical image segmentation. `LKM-UNet` utilizes the power-

ful sequence modeling capabilities and linear complexity of Mamba to implement large receptive fields by assigning large kernels (windows) to SSM modules. Further, we design a novel hierarchical and bidirectional large kernel Mamba block (LM block) to enhance the representation modeling capability of SSMs. The bidirectional Mamba design is used for location-aware sequence modeling, reducing the weight impact of the input order. The hierarchical Mamba design is composed of two types of SSM operations: (i) pixel-level SSM (PiM) and (ii) patch-level SSM (PaM). PiM captures the neighborhood and local-scope pixel information by large kernel SSM and avoids the forgetting problem that occurs in SSMs due to discontinuity of adjacent patches after tokenization. PaM deals with long-range dependency modeling and global patch interaction.

Our main contributions are as follows. (1) We propose a Large Kernel Mamba UNet (`LKM-UNet`) for 2D/3D medical image segmentation. (2) We assign kernels of large receptive fields to SSM layers to enable the model to possess the capability of large spatial modeling. (3) We design a bidirectional Mamba for location-aware sequence modeling. (4) We propose a novel hierarchical Mamba module composed of pixel-level SSM (PiM) and patch-level SSM (PaM), enhancing local-neighborhood pixel-level and long-range global patch-level modeling.

2 Preliminaries: SSM Models

SSM-based models, namely the structured state space sequence models (S4) and Mamba [5], all evolved from the continuous system that maps a 1-dimensional function or sequence $x(t) \rightarrow y(t) \in \mathcal{R}$ through a hidden state $h(t) \in \mathcal{R}^N$. This process can be represented as the following linear ordinary differential equation:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathcal{R}^{N\times N}$ is a state matrix and $\mathbf{B}, \mathbf{C} \in \mathcal{R}^N$ are projection parameters.

S4 and Mamba are discrete versions of the aforementioned continuous system, which include a timescale parameter $\mathbf{\Delta}$ to transform the continuous parameters $\mathbf{A}, \mathbf{B}$ to discrete parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. Typically, the zero-order hold (ZOH) is employed as the discretization rule and can be defined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \end{aligned} \tag{2}$$

After the discretization, the discretized version of Eq. (1) can be defined as :

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{3}$$

Then the output is computed through a global convolution, defined as:

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\ y &= x * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

where L is the length of the input sequence x and $\overline{\mathbf{K}} \in \mathcal{R}^L$ is a structured convolutional kernel.

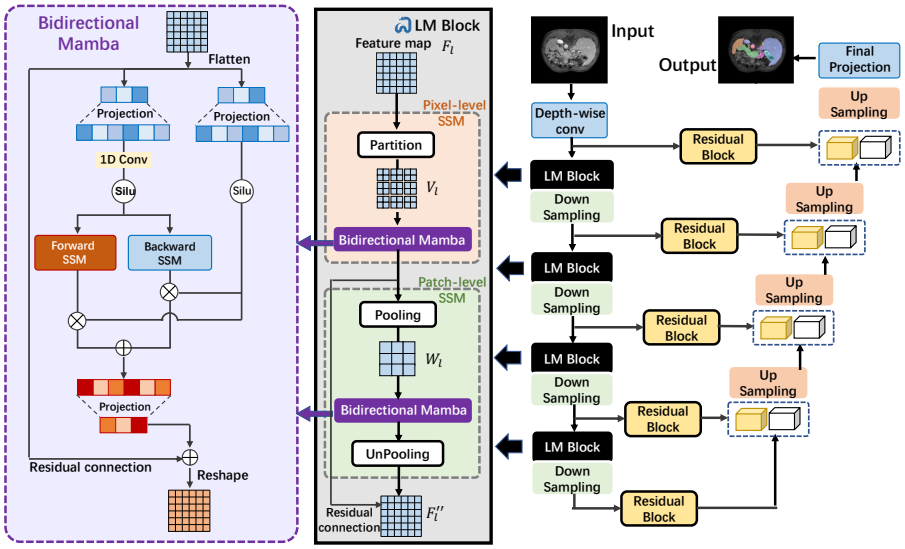

Fig. 1. An overview of our proposed `LKM-UNet`.

3 Method

In this section, we first introduce the overall architecture of `LKM-UNet`. Subsequently, we elaborate the core component, the LM block.

3.1 LKM-UNet

An overview of `LKM-UNet` is given in Fig. 1. Specifically, except for the common UNet composition of a depth-wise convolution, an encoder with downsampling layers, a decoder with upsampling layers, and skip-connection, `LKM-UNet` improves the structure of UNet with proposed large kernel Mamba (LM) blocks inserted into the encoder. Given a 3D input image with a resolution of $C \times D \times H \times W$, the depth-wise convolution first encodes the input into a feature map $\mathbf{F}_0 \in \mathcal{R}^{48 \times \frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}}$. Then the feature map $\mathbf{F}_0$ is fed into each LM block and the corresponding down-sampling layers, and multi-scale feature maps are obtained. An LM block contains two Mamba modules: pixel-level SSM (PiM) and patch-level SSM (PaM). For the l^{th} layer, the process can be formulated as:

$$\mathbf{F'}_l = \text{PiM}(\mathbf{F}_l), \quad \mathbf{F''}_l = \text{PaM}(\mathbf{F'}_l), \quad \mathbf{F}_{l+1} = \text{Down-sampling}(\mathbf{F''}_l), \tag{5}$$

where PiM and PaM denote pixel-level SSM and patch-level SSM, respectively. Down-sampling denotes the down-sampling layers. After each stage, the produced feature map $\mathbf{F}_{l+1}$ is encoded to $(2C_l, \frac{D_l}{2}, \frac{H_l}{2}, \frac{W_l}{2})$, where $C_l, (D_l, H_l, W_l)$ represent the channel and resolution of feature map $\mathbf{F}_l$. As for the decoder part, we adopt a UNet decoder and residual block with skip connections for upsampling and predicting the final segmentation masks.

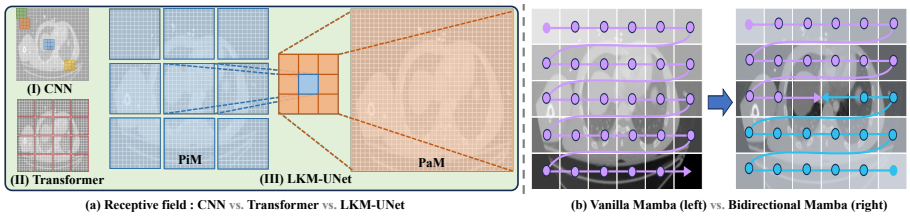

Fig. 2. (a) Respective field comparison among CNN, Transformer, and our proposed `LKM-UNet`. CNNs often use small kernels (like 3×3), and Transformers often use 7×7 sized kernels (windows). Our LKM-UNet can scale up kernel size to 40×40. (b) Scanning order comparison of vanilla Mamba vs. our proposed bidirectional Mamba.

3.2 LM block

LM block is our core component used for further spatial modeling of feature maps of different scales at each stage. Different from the previous methods that use CNN for local pixel-level modeling and Transformer for long-range patch-level dependency modeling, an LM block can accomplish pixel-level and patch-level modeling simultaneously, benefiting from the linear complexity of Mamba. More critically, a lower complexity allows setting larger kernels (windows) to obtain bigger receptive fields, which will improve the efficiency of local modeling, as shown in Fig. 2(a). Specifically, LM block is a hierarchical design that consists of pixel-level SSM (PiM) and patch-level SSM (PaM); the former is used for local-neighborhood pixel modeling and the latter is used for global long-range dependency modeling. Besides, each Mamba layer in an LM block is bidirectional, which is proposed for location-aware sequence modeling.

Pixel-level SSM (PiM). Since Mamba is a continuous model, the discrete nature of input pixels can weaken the correlation modeling of locally adjacent pixels. Hence, we propose a pixel-level SSM to split the feature map into multiple large sub-kernels (sub-windows) and perform SSM operations on the sub-kernels. We first equally divide a whole feature map into non-overlapping sub-kernels for 2D or sub-cubes for 3D. Take 2D for example. Given an input of $H \times W$ resolution, we divide the feature map into sub-kernels of size $m \times n$ each (m and n can be up to 40). Without loss of generality, we assume that H/m and W/n are both integers. Then we have $\frac{HW}{mn}$ sub-kernels, as shown in Pixel-level SSM of Fig. 1. Under this scheme, when these sub-kernels are sent into a Mamba layer, the local adjacent pixels will be input continuously into SSM; thus, the relation between local-neighborhood pixels can be better modeled. Moreover, under the large kernel partition strategy, the receptive fields are enlarged and the model can obtain more details of the local pixels. However, the image is divided into non-overlapping sub-kernels. Hence, we need a mechanism for communication between different sub-kernels, for long-range dependency modeling.

Patch-level SSM (PaM). We introduce a patch-level SSM layer to impart information among different sub-kernels. As shown in Patch-level SSM of Fig. 1, a feature map $\mathbf{F'}_l$ of resolution $H \times W$ first passes through a pooling layer of size $m \times n$ to allow important information for each of the $\frac{HW}{mn}$ sub-kernels to

be summarized into a single representative. Thus, we obtain aggregate maps $\mathbf{Z}_l$ with $\frac{HW}{mn}$ representatives, which are then used to communicate among the sub-kernels through Mamba for global-range dependency modeling. After the interaction in Mamba, we unpooling the aggregate maps back to the same size as the initial feature map $\mathbf{F'}_l$, and apply a residual connection. The process of $\mathbf{F}''_l = \text{PaM}(\mathbf{F}'_l)$ in Eq. (5) can be carried out as:

$$\mathbf{W}_l = \text{Pooling}(F'_l), \quad \mathbf{W}'_l = \text{Bi-Mamba}(\mathbf{W}_l), \quad \mathbf{F''}_l = \text{Unpooling}(\mathbf{W}'_l), \tag{6}$$

where Pooling and Unpooling denote the pooling layer and unpooling layer, respectively. Bi-Mamba denotes the proposed bidirectional Mamba layer.

Bidirectional Mamba (BiM). Different from the vanilla Mamba block which is based on forward-only scan direction SSM layers, each SSM layer (including PiM and PaM) in our LM block is bidirectional. Fig. 2(b) shows the differences. In the original Mamba, as a continuous model, some information forgetting occurs on the elements entered earlier, and the latest elements that enter into Mamba will retain much more information. Thus, the original Mamba with a single scanning direction will focus more on the posterior patches, rather than the center areas of the feature maps often with more organs and lesions. To this end, we propose a bidirectional Mamba structure by performing both forward and backward scanning at the same time and superimposing the output results. The detailed structure is shown in the left part of Fig. 1. BiM has two advantages. First, the model can focus more on the informative patches in the center areas of the image likely with more organs and lesions, rather than the corner areas. Second, for each patch, the absolute position information and relative position information with other patches can be well modeled by the network.

4 Experiments

4.1 Datasets

We conduct experimental comparisons with state-of-the-art methods on two datasets for 2D and 3D segmentation tasks to validate the effectiveness and scalability of `LKM-UNet`.

Abdomen CT. Abdomen CT is a publicly available 3D multi-organ segmentation dataset comprising 100 CT cases from the MICCAI 2022 FLARE Challenge [16], including 13 types of abdominal organs (liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum). The size of a 3D CT image is $40 \times 224 \times 192$. 50 cases from the MSD Pancreas dataset with annotations from AbdomenCT-1K are used for training, and another 50 cases from different medical centers [2] are used for testing.

Abdomen MR. Abdomen MR is a publicly available 2D segmentation dataset comprising 110 MRI cases from the MICCAI 2022 AMOS Challenge [11], including 13 types of abdominal organs (the same as the Abdomen CT dataset). The size of a 2D MRI image is 320×320. Following the previous work [15, 22], 60 annotated cases are used for training, and another 50 cases are used for testing.

Table 1. Quantitative segmentation results on the 3D Abdomen CT dataset and 2D Abdomem MR dataset.

Method	3D Abdomen CT		2D Abdomen MR	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑
SegResNet [17]	79.27	82.57	73.17	80.34
nnUNet [10]	86.15	89.72	74.50	81.53
UNTER [8]	68.24	70.04	57.47	63.09
SwinUNTER [7]	75.94	76.63	70.28	76.69
nnFormer [24]	78.34	81.45	72.79	79.63
U-Mamba (concurrent) [15]	86.38	89.80	76.25	83.27
`LKM-UNet`**(Ours)**	**86.82**	**90.02**	**77.35**	**83.80**

4.2 Implementation Setup

Our `LKM-UNet` is implemented on PyTorch 1.9.0 based on the nnU-Net framework. All the experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. The batch size in training is 2 for the 3D dataset (Abdomen CT) and 24 for the 2D dataset (Abdomen MR). The Adam [12] optimizer with momentum = 0.99 is used. The initial learning rate is 0.01 with a weight decay of 3e-5. The maximum training epoch number is 1000. For the Abdomen CT dataset, the stage is 6 but the dimensions are not consistent; thus we set the rectangle kernel size with three dimensions to [20, 28, 24], [20, 28, 24], [10, 14, 12], [10, 14, 12], [5, 7, 6], and [5, 7, 6] for each stage. For the Abdomen MR dataset, the stage is 7 and the kernel size is 40, 20, 20, 10, 10, 5, and 5 for each stage.

4.3 Overall Performances

The baseline models include three types of representative networks: CNN-based networks (nnU-Net [10] and SegResNet [17]), Transformer-based networks (UNETR [8], SwinUNETR [7], nnFormer [24]), and the latest Mamba-based network (U-Mamba [15]). For fair comparison, we also implement all the models in the nnU-Net framework, and use the default image pre-processing. Table 1 presents the results. Compared to both CNN-based and Transformer-based segmentation methods, our proposed `LKM-UNet` achieves improved performances in both DSC and NSD, which indicates that the global modeling capabilities of Mamba are critical to medical image segmentation. Note that, compared to U-Mamba which simply applies Mamba as a global modeling adapter, `LKM-UNet` exhibits improvements over U-Mamba, validating the effectiveness of our bidirectional and hierarchical Mamba designs. These results also demonstrate the potential of Mamba in global and local feature modeling with larger receptive fields.

4.4 Is the Kernel Size of `LKM-UNet` Important?

To validate Mamba's large spatial modeling potential, we explore `LKM-UNet`'s performance in different kernel size settings. Table 2 shows the results on the

Table 2. Performances of `LKM-UNet` in three different kernel size settings. The kernel size sequence indicates the kernel size in each stage (the total number of stages is 7).

Kernel size	[10, 5, 5, 5, 5, 5, 5]	[20, 10, 10, 10, 5, 5, 5]	[40, 20, 20, 10, 10, 5, 5]
DSC	75.89	76.45	**77.35**
NSD	82.26	82.78	**83.80**

Table 3. Performances of `LKM-UNet` with different sub-modules. PiM = Pixel-level SSM. PaM = Patch-level SSM. BiM = Bidirectional Mamba.

Method	Baseline	Only PiM	Only PaM	PiM + BiM	PaM + BiM	PiM + PaM	PiM + PaM + BiM
DSC	74.50	76.82	76.22	76.90	76.73	77.10	**77.35**
NSD	81.53	83.05	82.59	83.31	82.94	83.54	**83.80**

Abdomen MR dataset. Comparing the performances of the three kernel-size settings, one can see that `LKM-UNet` with larger kernel sizes achieves better performances. This indicates that large receptive fields are critical for medical image segmentation which can be achieved with Mamba due to its linear complexity.

4.5 Ablation Study

We conduct ablation experiments on the Abdomen MR dataset to validate the effect of each key component in our `LKM-UNet`, shown in Table 3. Both PiM and PaM provide improvements for `LKM-UNet` over the baseline model, validating the superiority of PiM and PaM in local pixel-level modeling and global modeling, respectively. Notably, the model with PiM gains more improvements, suggesting that enlarging the receptive field of local modeling is a key to improving model performance. After introducing BiM, the performance of `LKM-UNet` further improves, which shows the importance of bidirectional Mamba for location-aware modeling. Finally, `LKM-UNet` with all the components achieves the best performance, further demonstrating our method's effectiveness and its components.

4.6 Effective Receptive Field Visualization

To show more details of receptive field, we exhibit the effective receptive field (ERF) [14] of other methods and LKM-UNet in Fig. 3. CNN-based methods focus more on local feature extraction, while Transformer-based methods have a wider range of ERF. Although U-Mamba utilize Mamba to obtain globally ERF, it weakens some local attentions. By contrast, our proposed LKM-UNet with large kernel Mamba achieves larger ERF both in global and local aspects.

5 Conclusions

In this paper, we introduced a new Mamba-based UNet model for medical image segmentation, achieving large kernel spatial modeling. Further, we designed a bidirectional and hierarchical SSM to enhance the capacities of Mamba in local and global feature modeling. Comprehensive experiments on multi-organ segmentation datasets demonstrated the effectiveness of our proposed method.

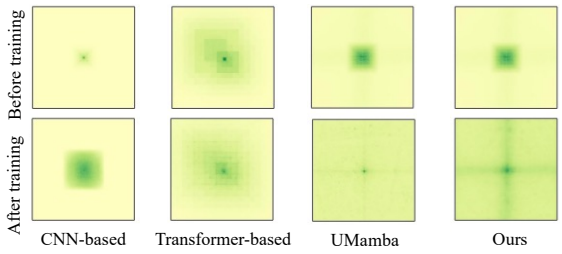

Fig. 3. Effective respective field visualization among CNN, Transformer, U-Mamba and our proposed `LKM-UNet`.

Acknowledgments. This research was partially supported by National Natural Science Foundation of China under grants No. 62176231, No. 62106218, No. 82202984, No. 92259202 and No. 62132017, Zhejiang Key R&D Program of China under grant No. 2023C03053.[7]

References

1. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision Transformers. Advances in Neural Information Processing Systems **34**, 9355–9366 (2021)
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (TCIA): Maintaining and operating a public information repository. Journal of Digital Imaging **26**, 1045–1057 (2013)
3. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVGG: Making VGG-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13733–13742 (2021)
4. Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052 (2022)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
7. 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 (2021)
8. 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)
9. Hua, W., Dai, Z., Liu, H., Le, Q.: Transformer quality in linear time. In: International Conference on Machine Learning. pp. 9099–9117. PMLR (2022)

[7] The authors have no competing interests to declare that are relevant to the content of this article.

10. 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)
11. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. 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)
14. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems **29** (2016)
15. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
16. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., et al.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: The FLARE22 challenge. arXiv preprint arXiv:2308.05862 (2023)
17. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Revised Selected Papers, Part II 4. pp. 311–320. Springer (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
19. 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)
20. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CvT: Introducing convolutions to vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22–31 (2021)
21. Wu, Y., Liao, K., Chen, J., Wang, J., Chen, D.Z., Gao, H., Wu, J.: D-Former: A U-shaped dilated transformer for 3D medical image segmentation. Neural Computing and Applications **35**(2), 1931–1944 (2023)
22. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
23. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 579–588 (2021)
24. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnFormer: Volumetric medical image segmentation via a 3D Transformer. IEEE Transactions on Image Processing (2023)
25. 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)