Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining

Jiarun Liu[1,2,3*], Hao Yang[1,2,3*], Hong-Yu Zhou[4✉], Yan Xi[1],
Lequan Yu[5], Cheng Li[1], Yong Liang[2], Guangming Shi[2], Yizhou Yu[4],
Shaoting Zhang[6], Hairong Zheng[1], and Shanshan Wang[1✉]

[1] Paul C. Lauterbur Research Center for Biomedical Imaging,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2] Peng Cheng Laboratory
[3] University of Chinese Academy of Sciences
[4] Department of Computer Science, The University of Hong Kong
[5] Department of Statistics and Actuarial Science, The University of Hong Kong
[6] Shanghai Artificial Intelligence Laboratory

Abstract. Accurate medical image segmentation demands the integration of multi-scale information, spanning from local features to global dependencies. However, it is challenging for existing methods to model long-range global information, where convolutional neural networks are constrained by their local receptive fields, and vision transformers suffer from high quadratic complexity of their attention mechanism. Recently, Mamba-based models have gained great attention for their impressive ability in long sequence modeling. Several studies have demonstrated that these models can outperform popular vision models in various tasks, offering higher accuracy, lower memory consumption, and less computational burden. However, existing Mamba-based models are mostly trained from scratch and do not explore the power of pretraining, which has been proven to be quite effective for data-efficient medical image analysis. This paper introduces a novel Mamba-based model, Swin-UMamba, designed specifically for medical image segmentation tasks, leveraging the advantages of ImageNet-based pretraining. Our experimental results reveal the vital role of ImageNet-based training in enhancing the performance of Mamba-based models. Swin-UMamba demonstrates superior performance with a large margin compared to CNNs, ViTs, and latest Mamba-based models. Notably, on AbdomenMRI, Encoscopy, and Microscopy datasets, Swin-UMamba outperforms its closest counterpart U-Mamba by an average score of 2.72%. The code and models of Swin-UMamba are publicly available at: https://github.com/Jiarun-Liu/Swin-UMamba.

Keywords: Medical image segmentation · ImageNet-based pretraining · Long-range dependency modeling.

[*] The first two authors contributed equally. Corresponding authors: ss.wang@siat.ac.cn and whuzhouhongyu@gmail.com.

1 Introduction

Medical image segmentation plays an important role in modern clinical practice such as assisting in diagnoses, formulating treatment plans, and implementing therapies [2,20,27]. In recent years, deep learning has made significant advancements in this field [23,31,6,11] to enhance efficiency, accuracy, and consistency in medical image analysis to make accurate and rapid diagnoses [26,14]. However, accurate medical image segmentation requires integrating local features with their corresponding global dependencies [24]. It is still challenging to efficiently capture complex and long-range global dependencies from image data [22,30]. Convolutional neural networks (CNNs) such as U-Net [23], nnU-Net [11], and SegResNet [21] are commonly employed in medical image segmentation. They are effective at extracting local features but may struggle with capturing global context and long-range dependencies. This is because CNNs are inherently limited by their local receptive fields [17], which restrict their ability to capture information from distant regions in the image. On the other hand, vision transformers (ViTs) have shown the capability in handling global context and long-range dependencies [10]. However, ViTs are constrained by their attention mechanism, suffering from high quadratic complexity for long sequences modeling [4], where high-resolution images are not rare in the medical domain (e.g. whole-slide pathology images, high-resolution MRI/CT scans). Despite the complexity, transformers are prone to overfitting when dealing with limited datasets [15], indicating their data-hungry nature.

Recently, Mamba [4] has demonstrated its efficiency and effectiveness in long-range dependency modeling. Compared with transformers, Mamba scales linearly or near-linearly with sequence length while maintaining the capability of modeling long-range dependencies, offering higher accuracy, lower memory consumption, and less computational burden [33]. Several latest studies [18,33,16,29] have preliminarily explored the effectiveness of Mamba in vision tasks. For instance, Vim [33] proposed a generic vision backbone with bidirectional Mamba blocks, while VMamba [16] introduced a cross-scan module to solve the direction-sensitive problem due to the difference between 1D sequences and 2D images. For medical image segmentation, U-Mamba [18] and SegMamba [29] proposed a task-specific architecture with the Mamba block based on nnU-Net [11] and Swin-UNETR [8], respectively.

Although remarkable performance has been accomplished with these efforts, existing Mamba-based models are mostly trained from scratch. The impact of pretraining for the Mamba-based model in medical image segmentation tasks remains unclear, which has been proven to be quite effective for data-efficient medical image analysis with CNNs [6] and ViTs [7]. This is particularly important in the field of medicine, where medical image datasets are often limited in size and diversity [28]. Understanding the effectiveness of pretraining Mamba-based models in medical image segmentation can offer valuable insights into enhancing the performance of deep learning models in medical imaging applications. However, prior research [18] typically employs a specific architecture with Mamba blocks, which fails to consider the transferability from generic vision models. Consequently, the network structure requires redesigning to integrate the pretrained model. Given the fact that the application of Mamba in vision is relatively new, further experimental evaluation is required. Moreover, there is a need for the scalability and efficiency of Mamba-based models for real-world deployment [32], particularly in resource-constrained environments, which is commonly found in medical practice.

In this paper, we proposed a Mamba-based network Swin-UMamba for 2D medical image segmentation. Swin-UMamba uses a generic encoder to integrate the power of the pretrained vision model with a well-designed decoder for medical image segmentation. In addition, we proposed a variant structure Swin-UMamba† with a Mamba-based decoder, providing fewer parameters and lower FLOPs for efficient applications while maintaining competitive performance. Our contribution can be summarized as follows:

- To the best of our knowledge, we are the first attempt to discover the impact of pretrained Mamba-based networks in medical image segmentation. Our experiment verified that ImageNet-based pretraining plays an important role in medical image segmentation for Mamba-based networks, which sometimes is crucial.
- We propose two Mamba-based networks Swin-UMamba and Swin-UMamba† for medical image segmentation, which are particularly designed to unify the power of pretrained models with different computation requirements towards real-world deployment.
- Our results show that both Swin-UMamba and Swin-UMamba† can outperform previous segmentation models including CNNs, ViTs, and the latest Mamba-based models with notable margin, highlighting the effectiveness of ImageNet-based pretraining and proposed architecture in medical image segmentation tasks.

2 Method

We illustrate the overall architecture of Swin-UMamba in Fig. 1. It is mainly composed of 1) a Mamba-based encoder that was pretrained on the large-scale dataset (i.e. ImageNet) to extract features at different scales, 2) a decoder with several up-sample blocks for predicting segmentation results, and 3) skip connections to bridge the gap between low-level details and high-level semantics. We will introduce the detailed structure of Swin-UMamba in the following sections.

2.1 Mamba-based VSS block

Mamba [4] using space state sequential models (SSMs) [5] to reduce the complexity of attention from quadratic to linear for long-sequence modeling in natural language processing. However, the distinction between 2D visual data and 1D language sequences requires careful consideration when adapting Mamba to vision tasks. For instance, while 2D spatial information is crucial in vision tasks

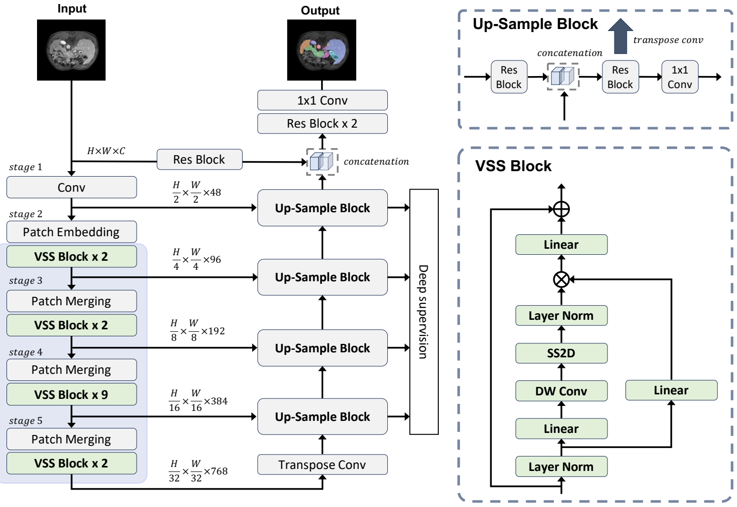

Fig. 1. The overall architecture of Swin-UMamba. Swin-UMamba can leverage the power of vision foundation models by loading the weights of pretrained models. Each block within the blue box was initialized with the ImageNet pretrained weights.

[16], it is not the primary focus in 1D sequence modeling. Directly adopting Mamba to flattened images would inevitably result in restricted receptive fields, where the relationships against unscanned patches could not be estimated.

Building upon the insights from [16], we incorporate the visual state space (VSS) block as the basic unit in Swin-UMamba. The VSS block addresses the challenges associated with 2D image data by employing 2D-selective-scan (SS2D) based on the selective scan space state sequential model (S6). Given input feature z, the output feature $\bar{z}$ of SS2D can be written as:

$$z_v = expand(z, v) \tag{1}$$

$$\bar{z}_v = S6(z_v) \tag{2}$$

$$\bar{z} = merge(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4) \tag{3}$$

where $v \in V = \{1, 2, 3, 4\}$ is four different scanning directions. $expand(\cdot)$ and $merge(\cdot)$ corresponding to the *scan expand* and *scan merge* operations in [16]. S6 enables each element in a 1D array (e.g., text sequence) to interact with any of the previously scanned samples through a compressed hidden state. We refer to [16] for further details about S6. The overall structure of the VSS block is illustrated in Fig. 1.

2.2 Integrating ImageNet-based pretraining

The primary challenge lies in effectively integrating generic pretrained models into the segmentation task. To this end, we construct an encoder that shares a similar structure with VMamba-Tiny [16], which was pretrained on the extensive ImageNet dataset. It allowed us to integrate the power of the generic vision model to extract information with long-range modeling capability, mimic the risk of overfitting, and establish a robust initialization for Swin-UMamba.

As illustrated in Fig. 1, the encoder of Swin-UMamba can be divided into 5 stages. The first stage is a convolution layer for 2× down-sampling. It differs from VMamba because we prefer a gradual down-sampling process to retain low-level details, which is important for medical image segmentation [23,25]. Subsequent stages follow the design of VMamba-Tiny, where each stage is composed of a patch merging layer for 2× down-sampling and several VSS blocks to process high-level features. Specifically, the patch merging layer in stage 2 was replaced by a 2×2 patch embedding layer. The number of VSS blocks and feature dimensions at each stage are $\{0, 2, 2, 9, 2\}$ and $D = \{48, 96, 192, 384, 768\}$, respectively. We initialize the VSS blocks and patch merging layers with the ImageNet pretrained VMamba-Tiny.

2.3 Swin-UMamba decoder

The decoder of Swin-UMamba follows the commonly used U-shaped architecture. As illustrated in Fig. 1, Swin-UMamba uses an up-sample block with 1) an extra convolution block with a residual connection to process skip connection features, and 2) an additional segmentation head at each scale for deep supervision [13]. Given skip-connected feature z'_l from stage-l and feature z_{l+1} from the last up-sample block, the output feature z_l and the segmentation map y_l at stage-l can be formulated as follows:

$$\hat{z}_l = Res_l^{(2)}(Cat(z_{l+1}, Res_l^{(1)}(z'_l))) \tag{4}$$

$$z_l = DeConv_l(\hat{z}_l), \quad y_l = Conv_l(\hat{z}_l) \tag{5}$$

where $Cat(\cdot)$, $DeConv_l(\cdot)$, $Conv_l(\cdot)$ are the feature concatenation operation, transpose convolution, and 1×1 convolution, respectively. $Res_l^{(1)}(\cdot)$ and $Res_l^{(2)}(\cdot)$ are two convolution blocks with residual connection at stage-l, each $Res(\cdot)$ was composed of two convolution layers with LeakyRELU activation. We use 1×1 convolution to project the feature map dimension d_l into class number K for the final segmentation output.

2.4 Swin-UMamba†: Swin-UMamba with Mamba-based decoder

To further explore the potential of Mamba in medical semantic segmentation, we proposed a variant Swin-UMamba† with a Mamba-based decoder, which can exhibit decent performance with largely reduced complexity.

Several modifications were made on Swin-UMamba†. First, the up-sample block was replaced by 2× patch expanding layer [3] and two VSS blocks. We found that many parameters and computation burdens were caused by the heavy CNN-based decoder. Second, we changed the encoder back to the original design of VMamba and then removed corresponding skip connections and redundant

Table 1. Dataset information. We follow [18] to perform data processing. Dim indicates the dimension of the processed data in our experiment.

Dataset	Dim	#Training	#Testing	#Targets	Crop size	Epochs
AbdomenMRI[12]	2D	5615	3357	13	$(320, 320)$	100
Endoscopy[1]	2D	1800	1200	7	$(384, 640)$	350
Microscopy[19]	2D	1000	101	2	$(512, 512)$	450

up-sample blocks. The last patch expanding layer in the decoder is 4× up-sample operation, mirroring the 4× patch embedding layer. Deep supervision was applied at resolutions of $\{1\times, \frac{1}{4}\times, \frac{1}{8}\times, \frac{1}{16}\times\}$ as there is no feature at $\frac{1}{2}\times$ scale. Combining all these modifications, the number of network parameters was reduced from 60M to 27M, and the FLOPs were decreased from 68.0G to 18.9G on the AbdomenMRI dataset. Further details of Swin-UMamba† can be found in the supplementary material.

3 Experiments

3.1 Datasets

We evaluate the performance and scalability of Swin-UMamba across three distinct medical image segmentation datasets, encompassing AbdomenMRI [12] (abdominal organs), Endoscopy [1] (instruments), and Microscopy [19] (cell). These datasets are selected across various resolutions and image modalities. We list the information of these datasets in Table 1. The data processing strategy in our experiment was following [18] during training and testing.

3.2 Implemetation details

We implemented Swin-UMamba on top of the well-established nnU-Net framework [11]. The loss function is the sum of Dice loss and cross-entropy loss and we perform deep supervision [13] at each scale. We use an AdamW optimizer with weight decay = 0.05 following [16]. A cosine learning rate decay was adopted with an initial learning rate = 0.0001. We use the ImageNet pretrained VMamba-Tiny model to initialize Swin-UMamba for all datasets. During training, we froze all pretrained parameters for the first 10 epochs to align other modules. Following [18], we disabled the testing time argumentation for a more streamlined and efficient evaluation. For more details, please refer to our code implementation.[7].

3.3 Baselines and evaluation metrics

We select three types of methods as baseline methods for comprehensive evaluation, including CNN-based (nnU-Net [11], SegResNet [21]), transformer-based

[7] https://github.com/JiarunLiu/Swin-UMamba

Table 2. Segmentation results on AbdomenMRI, Endoscopy, and Microscopy (Micro) dataset. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [18]. The number of parameters (param) and FLOPs were computed on the AbdomenMRI dataset. Further results with standard deviation can be found in the supplementary material. ∗: Deep supervision was disabled and we extend the training epochs to 200.

Dataset	param	FLOPs	AbdomenMRI		Endoscopy		Micro	AVG
Metric			DSC	NSD	DSC	NSD	F1	
CNN-based								
nnU-Net	33M	23.3G	0.7450	0.8153	0.6264	0.6412	0.5383	0.6732
SegResNet	6M	24.5G	0.7317	0.8034	0.5820	0.5968	0.5411	0.6510
Transformer-based								
UNETR	87M	42.1G	0.5747	0.6309	0.5017	0.5168	0.4357	0.5320
SwinUNETR	25M	27.9G	0.7028	0.7669	0.5528	0.5683	0.3967	0.5975
nnFormer	60M	50.2G	0.7297	0.7963	0.6135	0.6228	0.5332	0.6591
Mamba-based								
U-Mamba_Bot	63M	45.7G	0.7588	0.8285	0.6540	0.6692	0.5389	0.6899
U-Mamba_Enc	67M	49.9G	0.7625	0.8327	0.6303	0.6451	0.5607	0.6863
w/o ImageNet-based pretraining								
Swin-UMamba	60M	68.0G	0.7054	0.7647	0.5483	0.5632	0.4561	0.6075
Swin-UMamba†	28M	18.9G	0.6653*	0.7312*	0.6402	0.6547	0.5186	0.6420
w/ ImageNet-based pretraining								
Swin-UMamba	60M	68.0G	**0.7760**	**0.8421**	0.6767	0.6922	0.5806	0.7135
Swin-UMamba†	28M	18.9G	0.7705	0.8376	**0.6783**	**0.6933**	**0.5982**	**0.7156**

(UNETR [9], Swin-UNETR [8], nnFormer [31]), and the latest Mamba-based segmentation network U-Mamba [18]. It's worth noting that adopting the pretrained model into U-Mamba is not straightforward due to structural differences from the pretrained model [16]. We report the results of nnFormer [31] based on official implementation and the other baseline results were referenced from [18]. Dice similarity coefficient (DSC) and normalized surface distance (NSD) were used to evaluate segmentation performance on the AbdomenMRI and Endoscopy datasets. For the Microscopy dataset, we use the F1 score following [18].

3.4 Overall performance

Table 2 presents the segmentation performance on three segmentation datasets. Swin-UMamba and Swin-UMamba† outperform all baseline methods, including CNN-based networks, transformer-based networks, and the Mamba-based networks. The superior result demonstrates the great potential of the Mamba-based network in medical image segmentation. Swin-UMmaba and Swin-UMamba† exhibit a remarkable 2.72% and 2.93% improvement over U-Mamba_Enc in average score. Somewhat surprisingly, we observed that Swin-UMamba† outperforms Swin-UMamba on Endoscopy and Microscopy dataset. One possible

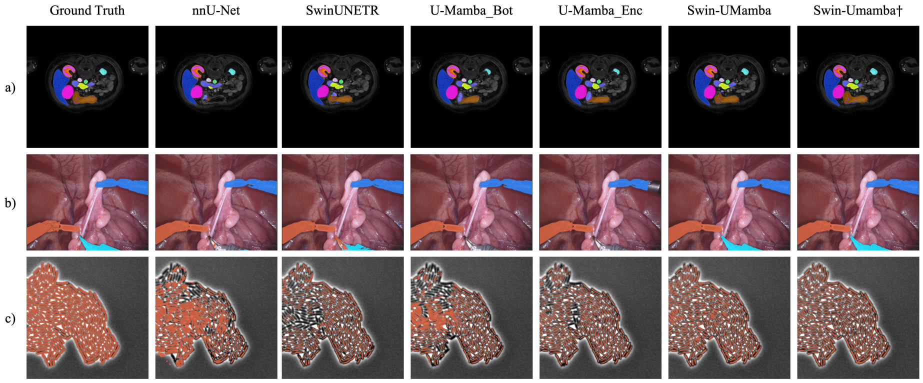

Fig. 2. Result visualization on a) AbdomenMRI, b) Endoscopy, and c) Microscopy.

reason is that Swin-UMamba† has fewer network parameters, making it more robust to small datasets. Besides, Swin-UMamba† exhibits a significantly lower computation burden with the lowest FLOPs among all baseline models. These competitive results demonstrate the potential of the pure Mamba-based network in settings with higher image resolution and limited samples. Fig. 2 shows that Swin-UMamba can accurately recognize target regions.

3.5 The impact of ImageNet-based pretraining

ImageNet-based pretraining shows a crucial role in our experiments, leading to a significant 10.60% average score improvement for Swin-UMamba. This improvement is consistent over different network structures and datasets, since Swin-UMamba† also benefits 7.36% in average score by using ImageNet-based pretraining. It is an effective strategy for mitigating overfitting in small datasets. Swin-UMamba can benefit over 10% on the relatively small Endoscopy and Microscopy datasets. Moreover, ImageNet-based pretraining facilitates faster and more stable training, requiring merely one-tenth of the training iterations compared to baseline methods on the AbdomenMRI dataset. A drastic phenomenon is observed with Swin-UMamba† on the AbdomenMRI dataset. Without ImageNet-based pretraining, Swin-UMamba† fails to converge properly on this dataset with default settings. To address this issue, we disable the deep supervision of Swin-UMamba† in this case. Despite that, Swin-UMamba† outperforms all baseline methods when utilizing the ImageNet pretrained weights. This improvement is particularly noteworthy considering that Swin-UMamba† has less than half of the network parameters and FLOPs compared to U-Mamba.

4 Conclusion

This study aims to reveal the impact of ImageNet-based pretraining for Mamba-based models in 2D medical image segmentation. We proposed a novel Mamba-based model, Swin-UMamba, and its variant, Swin-UMamba†, both capable of leveraging the power of pretrained models for segmentation tasks. Our experiments on various medical image segmentation datasets suggest that ImageNet-based pretraining for Mamba-based models offers several advantages, including superior segmentation accuracy, stable convergence, mitigation of overfitting issues, data efficiency, and lower computational resource consumption. We believe that our findings highlight the importance of pretraining in enhancing the performance and efficiency of Mamba-based models in vision tasks.

Acknowledgments. This research was partly supported by the National Key R&D Program of China (2023YFA1011400), National Natural Science Foundation of China (62222118, U22A2040), Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (2022B1212010-011), Shenzhen Science and Technology Program (RCYX20210706092104034, JCYJ20220531100213029), the major key project of Peng Cheng Laboratory under grant PCL2023AS1-2, Key Laboratory for Magnetic Resonance and Multimodality Imaging of Guangdong Province (2023B1212060052), and Youth Innovation Promotion Association CAS.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
2. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S.E., et al.: A population-based phenome-wide association study of cardiac and aortic structure and function. Nature medicine **26**(10), 1654–1662 (2020)
3. Cao, H., Wang, Y., Chen, J., Dongsheng Jiang, Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision – ECCV 2022 Workshops. pp. 205–218 (2023)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
5. Gu, A., Goel, K., Re, C.: Efficiently modeling long sequences with structured state spaces. In: International Conference on Learning Representations (2021)
6. Guo, J., Zhou, H.Y., Wang, L., Yu, Y.: UNet-2022: Exploring dynamics in non-isomorphic architecture. In: Medical Imaging and Computer-Aided Diagnosis. pp. 465–476. Springer Nature (2023)
7. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence **45**(1), 87–110 (2022)
8. 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)
9. 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)

10. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International Conference on Machine Learning. pp. 12633–12646. PMLR (2023)
11. 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)
12. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
13. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. pp. 562–570. PMLR (2015), ISSN: 1938-7228
14. Li, C., Li, W., Liu, C., Zheng, H., Cai, J., Wang, S.: Artificial intelligence in multiparametric magnetic resonance imaging: A review. Medical Physics **49**(10), e1024–e1054 (2022)
15. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open (2022)
16. 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)
17. 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)
18. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
19. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al.: The multi-modality cell segmentation challenge: towards universal solutions. arXiv preprint arXiv:2308.05864 (2023)
20. Mei, X., Lee, H.C., Diao, K.y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al.: Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature medicine **26**(8), 1224–1228 (2020)
21. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 311–320 (2019)
22. Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., Wang, S.: X-Net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 247–255 (2019)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241 (2015)
24. Sinha, A., Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE Journal of Biomedical and Health Informatics **25**(1), 121–130 (2021)
25. Sun, H., Li, C., Liu, B., Liu, Z., Wang, M., Zheng, H., Feng, D.D., Wang, S.: AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. Physics in Medicine & Biology **65**(5), 055005 (feb 2020)
26. Tang, H., Chen, X., Liu, Y., Lu, Z., You, J., Yang, M., Yao, S., Zhao, G., Xu, Y., Chen, T., et al.: Clinically applicable deep learning framework for organs at risk delineation in CT images. Nature Machine Intelligence **1**(10), 480–491 (2019)

27. Tang, H., Zhang, C., Xie, X.: Automatic pulmonary lobe segmentation using deep learning. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 1225–1228. IEEE (2019)
28. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021)
29. 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)
30. Yang, H., Huang, W., Qi, K., Li, C., Liu, X., Wang, M., Zheng, H., Wang, S.: CLCI-Net: Cross-level fusion and context inference networks for lesion segmentation of chronic stroke. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 266–274 (2019)
31. 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 **32**, 4036–4045 (2023)
32. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-UNet: A dimension-fusion u shape network for chronic stroke lesion segmentation. IEEE/ACM Transactions on Computational Biology and Bioinformatics **18**(3), 940–950 (2021)
33. 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)