

LightM-UNet: Mamba Assists in Lightweight UNet for Medical Image Segmentation

Weibin Liao^{1,3}, Yinghao Zhu^{2,4}, Xinyuan Wang⁴, Chengwei Pan⁴, Yasha Wang^{1,2} [†], and Liantao Ma^{1,2} [†]

¹Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China

²National Engineering Research Center for Software Engineering, Peking University, Beijing, China

³School of Computer Science, Peking University, Beijing, China

⁴Institute of Artificial Intelligence, Beihang University, Beijing, China

Abstract. UNet and its variants have been widely used in medical image segmentation. However, these models, especially those based on Transformer architectures, pose challenges due to their large number of parameters and computational loads, making them unsuitable for mobile health applications. Recently, State Space Models (SSMs), exemplified by Mamba, have emerged as competitive alternatives to CNN and Transformer architectures. Building upon this, we employ Mamba as a lightweight substitute for CNN and Transformer within UNet, aiming at tackling challenges stemming from computational resource limitations in real medical settings. To this end, we introduce the Lightweight Mamba UNet (**LightM-UNet**) that integrates Mamba and UNet in a lightweight framework. Specifically, **LightM-UNet** leverages the *Residual Vision Mamba Layer* in a pure Mamba fashion to extract deep semantic features and model long-range spatial dependencies, with linear computational complexity. Extensive experiments conducted on two real-world 2D/3D datasets demonstrate that **LightM-UNet** surpasses existing state-of-the-art literature. Notably, when compared to the renowned nnU-Net, **LightM-UNet** achieves superior segmentation performance while drastically reducing parameter and computation costs by 116x and 21x, respectively. This highlights the potential of Mamba in facilitating model lightweighting. Our code implementation is publicly available at <https://github.com/MrBlankness/LightM-UNet>

Keywords: Medical Image Segmentation · Light-weight Model · State Space Models.

1 Introduction

UNet [16], as a well-established algorithm for medical image segmentation, finds extensive application across a spectrum of segmentation tasks pertaining to medical organs and lesions, spanning various modalities of medical images. Its symmetrical U-shaped encoder-decoder architecture, coupled with integral skip connections, has laid the groundwork for segmentation models, spawning a plethora

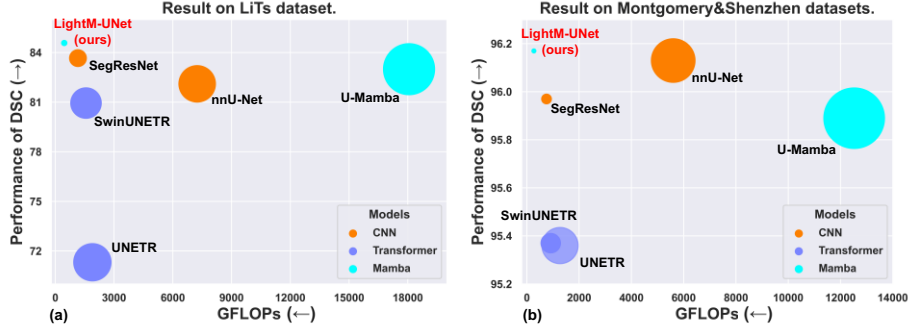


Fig. 1. (a) and (b) respectively show the visualization of comparative experimental results on LiTs [1] and Montgomery&Shenzhen [9] datasets. The central position of the marker indicates the performance of the model, while the size of the marker indicates the number of parameters of the model (larger size indicates a greater number of parameters). Colors in the legend represent the basic architecture these models applied.

of works [8,15,18] predicated on the U-shaped structure. However, being a Convolutional Neural Network-based (CNN-based) model, UNet grapples with the inherent locality of convolution operations, which poses limitations on its capacity to apprehend explicit global and long-range semantic information interactions [2]. Several studies have attempted to mitigate this issue by employing atrous convolutional layers [5], self-attention mechanisms [19], and image pyramids [25]. Nonetheless, these methods still exhibit constraints in modeling long-range dependencies.

In efforts to endow UNet with the capacity to apprehend global information, recent studies [2,7,6] have delved into integrating Transformer architectures [22], leveraging self-attention mechanisms to capture global information by treating the image as a sequence of contiguous patches. Although effective, Transformer-based solutions introduce quadratic complexity concerning image sizes owing to the self-attention mechanism, leading to a substantial computational overhead, particularly for tasks requiring dense predictions such as medical image segmentation. This overlooks the imperative of computational constraints in real-world medical settings, falling short of fulfilling the necessity for models characterized by low parameters and minimal computational load in mobile healthcare segmentation tasks [18]. In summary, the unresolved inquiry persists: “How can UNet be endowed with the capability to accommodate long-range dependencies without incurring additional parameters and computational burden?”

Recently, State Space Models (SSMs) have garnered considerable attention among researchers. Expanding upon the groundwork laid by classical SSM research [10], modern SSMs (e.g., Mamba [4]) not only establish long-range dependencies but also demonstrate linear complexity concerning input size, making Mamba a strong competitor to CNN and Transformer on the lightweight road of UNet. Some contemporary endeavors, such as U-Mamba [14], have proposed a hybrid CNN-SSM block, amalgamating the local feature extraction capabil-

ity of convolutional layers with SSM’s proficiency in capturing longitudinal dependency relationships. However, U-Mamba [14] introduces a substantial number of parameters and computational load (173.53M parameters and 18,057.20 GFLOPs), rendering it challenging to deploy in mobile healthcare settings for medical segmentation tasks. Therefore, in this study, we introduce **LightM-UNet**, a lightweight U-shaped segmentation model based on Mamba, which achieves state-of-the-art performance while significantly reducing parameter and computation costs (as depicted in Fig. 1). The contributions of this work are threefold.

1. We introduce **LightM-UNet**, a lightweight fusion of UNet and Mamba, boasting a mere parameter count of **1M**. Through validation on both 2D and 3D real-world datasets, **LightM-UNet** surpasses existing state-of-the-art models. In comparison to the renowned nnU-Net [8] and contemporaneous U-Mamba [14], **LightM-UNet** reduces the parameter count by $116\times$ and $224\times$, respectively.
2. Technically, we propose the *Residual Vision Mamba Layer (RVM Layer)* to extract deep features from images in a pure Mamba manner. With minimal introduction of new parameters and computational overhead, we further enhance the capability of SSM to model long-range spatial dependencies in visual images by utilizing *residual connections* and *adjustment factors*.
3. Insightfully, in contrast to contemporaneous endeavors [14,23,17] that integrate UNet with Mamba, we advocate for employing Mamba as a lightweight substitute for CNN and Transformer within UNet, aiming at tackling challenges stemming from computational resource limitations in real medical settings. To our knowledge, this represents the pioneering effort introducing Mamba into UNet as a lightweight optimization strategy.

2 Methodologies

While **LightM-UNet** supports both 2D and 3D versions of medical image segmentation, for convenience, this manuscript describes the methodology using the 3D version of **LightM-UNet**. — **A reading-friendly reminder.**

2.1 Architecture Overview

The overall architecture of the proposed **LightM-UNet** is illustrated in Fig. 2. Given an input image $I \in \mathbb{R}^{C \times H \times W \times D}$, where C , H , W , and D denote the number of channels, height, width, and number of slices of the 3D medical image, respectively. **LightM-UNet** commences by employing a depthwise convolution (DWConv) layer for shallow feature extraction, generating the shallow feature map $F_S \in \mathbb{R}^{32 \times H \times W \times D}$, where 32 denotes a fixed number of filters. Subsequently, **LightM-UNet** incorporates three consecutive Encoder Blocks to extract deep features from the images. Post each Encoder Block, the number of channels in the feature maps doubles, while the resolution halves. Consequently, **LightM-UNet** extracts deep features $F_D^l \in \mathbb{R}^{(32 \times 2^l) \times (H/2^l) \times (W/2^l) \times (D/2^l)}$ at the

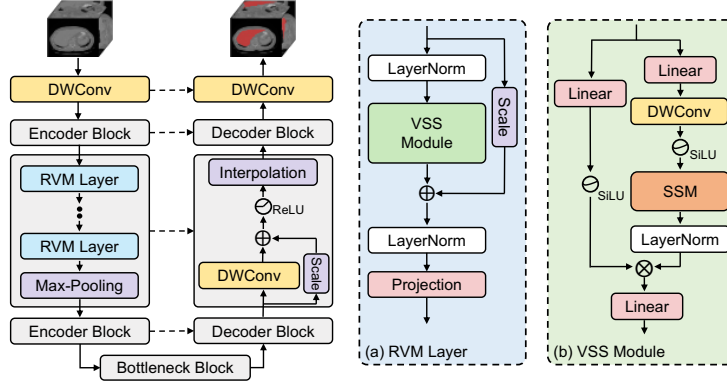


Fig. 2. The overall network architecture of **LightM-UNet** as well as the (a) *Residual Vision Mamba Layer (RVM Layer)*, the (b) *Vision State-Space Module (VSS Module)*.

l -th Encoder Block, where $l \in \{1, 2, 3\}$. Subsequent to this, **LightM-UNet** employs a Bottleneck Block to model long-range spatial dependencies while retaining the size of the feature maps unchanged. Following that, **LightM-UNet** integrates three consecutive Decoder Blocks for feature decoding and image resolution restoration. Following each Decoder Block, the number of channels in the feature maps is halved, and the resolution is doubled. Finally, the output of the last Decoder Block attains the same resolution as the original image, comprising 32 feature channels. **LightM-UNet** utilizes a DWConv layer to map the number of channels to the number of segmentation targets and applies a SoftMax activation function to generate the image mask. In alignment with the design of UNet, **LightM-UNet** also employs skip connections to furnish multi-level feature maps for the decoder.

2.2 Encoder Block

To minimize the number of parameters and computational cost, **LightM-UNet** employs Encoder Blocks comprising solely Mamba structures to extract deep features from the image. Specifically, given a feature map $F^l \in \mathbb{R}^{\tilde{C} \times \tilde{H} \times \tilde{W} \times \tilde{D}}$, where $\tilde{C} = 32 \times 2^l$, $\tilde{H} = H/2^l$, $\tilde{W} = W/2^l$, $\tilde{D} = D/2^l$, and $l \in \{1, 2, 3\}$, the Encoder Block initially flattens and transposes the feature map into a shape of (\tilde{L}, \tilde{C}) , where $\tilde{L} = \tilde{H} \times \tilde{W} \times \tilde{D}$. Subsequently, the Encoder Block utilizes N_l consecutive RVM Layers to capture global information, with the number of channels increased in the last RVM Layer. Following this, the Encoder Block reshapes and transposes the feature map into a shape of $(\tilde{C} \times 2, \tilde{H}, \tilde{W}, \tilde{D})$, succeeded by a Max-Pooling operation to reduce the resolution of the feature map. Ultimately, the l -th Encoder Block outputs the new feature map F^{l+1} with a shape of $(\tilde{C} \times 2, \tilde{H}/2, \tilde{W}/2, \tilde{D}/2)$.

Residual Vision Mamba Layer (RVM Layer) **LightM-UNet** proposes the RVM Layer to enhance the original SSM block for image deep semantic feature

extraction. Specifically, **LightM-UNet** utilizes advanced *residual connections* and *adjustment factors* to further enhance the long-range spatial modeling capability of SSM, with almost no introduction of new parameters and computational complexity. As depicted in Fig. 2 (a), given the input deep feature $M_{in}^l \in \mathbb{R}^{\tilde{L} \times \tilde{C}}$, the RVM Layer initially employs LayerNorm followed by the VSSM to capture spatial long-range dependencies. Subsequently, it utilizes an adjustment factor $s \in \mathbb{R}^{\tilde{C}}$ in the residual connection[3] for better performance. This process can be represented mathematically as follows:

$$\widetilde{M}^l = VSSM(LayerNorm(M_{in}^l)) + s \cdot M_{in}^l \quad (1)$$

Following this, the RVM Layer employs another LayerNorm to normalize \widetilde{M}^l , and subsequently utilizes a projection layer to convert \widetilde{M}^l into a deeper feature. The above process can be formulated as:

$$M_{out}^l = Projection(LayerNorm(\widetilde{M}^l)) \quad (2)$$

Vision State-Space Module (VSS Module) Following the approach outlined in [13], **LightM-UNet** introduces the VSS Module (depicted in Fig. 2 (b)) for long-range spatial modeling. The VSS Module takes the feature $W_{in}^l \in \mathbb{R}^{\tilde{L} \times \tilde{C}}$ as input and channels it into two parallel branches. In the first branch, the VSS Module expands the feature channels to $\lambda \times \tilde{C}$ using a linear layer, where λ represents a pre-defined channel expansion factor. Subsequently, it applies a DWConv, SiLU activation function [20], followed by the SSM and LayerNorm. In the second branch, the VSS Module also expands the feature channels to $\lambda \times \tilde{C}$ using a linear layer, followed by the SiLU activation function. Subsequently, the VSS Module aggregates features from the two branches using the Hadamard product and projects the channel number back to \tilde{C} to generate the output W_{out} with the same shape as the input W_{in} . The above process can be formulated as:

$$\begin{aligned} W_1 &= LayerNorm(SSM(SiLU(DWConv(Linear(W_{in})))))) \\ W_2 &= SiLU(Linear(W_{in})) \\ W_{out} &= Linear(W_1 \odot W_2) \end{aligned} \quad (3)$$

where \odot denotes the Hadamard product.

2.3 Bottleneck Block

Similar to Transformer, Mamba encounters convergence challenges when the network depth becomes excessive [21]. Consequently, **LightM-UNet** addresses this issue by incorporating four successive RVM Layers to construct bottlenecks for further modeling spatial long-term dependency. Within these bottleneck regions, the number of feature channels and the resolution remain unchanged.

2.4 Decoder Block

LightM-UNet employs Decoder Blocks to decode feature maps and restore image resolution. Specifically, given $F_D^l \in \mathbb{R}^{\tilde{C} \times \tilde{H} \times \tilde{W} \times \tilde{D}}$ from the skip connection and

$P_{in} \in \mathbb{R}^{\tilde{C} \times \tilde{H} \times \tilde{W} \times \tilde{D}}$ from the output of the previous block, the Decoder Block first performs feature fusion using an addition operation. Subsequently, it utilizes a DWConv, a residual connection, and a ReLU activation function to decode the feature map. Additionally, an *adjustment factors* s' is added to the residual connection to enhance the decoding capability. This process can be expressed mathematically as:

$$P_{out} = ReLU(DWConv(P_{in} + F_D^l) + s' \cdot (P_{in} + F_D^l)) \quad (4)$$

The Decoder Block ultimately employs bilinear interpolation to restore predictions to the original resolution.

3 Experiments

Datasets and Experimental Setups. To assess the performance of our model, we select two publicly available medical image datasets: the LiTs dataset [1], comprising 3D CT images, and the Montgomery&Shenzhen dataset [9], comprising 2D X-ray images. These datasets are extensively utilized in existing segmentation research [12,24] and are employed here to validate the performance of the 2D and 3D versions of LightM-UNet, respectively. The data were randomly partitioned into training, validation, and testing sets in a ratio of 7:1:2.

LightM-UNet was implemented using the PyTorch framework and the number of RVM Layers in the three Encoder Blocks is set as 1, 2, and 2, respectively. All experiments were conducted on a single Quadro RTX 8000 GPU. SGD was employed as the optimizer, initialized with a learning rate of 1e-4. The PolyLRScheduler was used as the scheduler, and a total of 100 epochs were trained. In addition, the loss function was designed as a simple combination of Cross Entropy loss and Dice loss. For the LiTs dataset, the images were normalized and resized to $128 \times 128 \times 128$, with a batch size of 2. For the Montgomery&Shenzhen dataset [9], the images were normalized and resized to 512×512 , with a batch size of 12.

To evaluate LightM-UNet, we compared it with two CNN-based segmentation networks (nnU-Net [8] and SegResNet [15]), two Transformer-based networks (UNETR [7] and SwinUNETR [6]), and a Mamba-based network (U-Mamba [14]), which are commonly used in medical image segmentation competitions. Additionally, we employed Mean Intersection over Union (mIoU) and Dice similarity score (DSC) as evaluation metrics.

Comparative Results. The comparative experimental results presented in Table. 1 demonstrate that our LightM-UNet achieves comprehensive state-of-the-art performance on the LiTS dataset[11]. Notably, compared to larger models like nnU-Net, LightM-UNet not only exhibits superior performance but also significantly reduces the number of parameters and computational costs by $47.39\times$ and $15.82\times$, respectively. When compared to the contemporaneous U-Mamba [14], LightM-UNet shows a performance improvement of 2.11% in terms of average mIoU. Particularly for tumors, which are often too small to be easily detected,

Table 1. Comparative experimental results on the LiTS [1] dataset using various 3D segmentation models.

Models	Params(M)	GFLOPs	Liver		Tumor		Average	
			DSC	mIoU	DSC	mIoU	DSC	mIoU
nnU-Net [8]	88.62	7,240.26	95.77	91.94	68.45	56.34	82.11	74.13
SegResNet [15]	18.79	1,158.30	96.11	92.56	71.22	59.76	83.67	76.16
UNETR [7]	92.62	1,891.35	94.06	88.95	48.58	37.01	71.32	62.98
SwinUNETR [6]	61.99	1,570.32	95.24	91.07	66.67	55.09	80.95	73.08
U-Mamba [14]	173.53	18,057.20	95.94	92.33	70.05	58.42	83.00	75.37
LightM-UNet	1.87	457.62	96.31	92.92	72.86	62.05	84.58	77.48

Table 2. Comparative experimental results on the Montgomery&Shenzhen [9] dataset using various 2D segmentation models.

Models	Params(M)	GFLOPs	DSC	mIoU
nnU-Net [8]	126.56	5,594.98	96.13	92.66
SegResNet [15]	6.29	748.96	95.97	92.36
UNETR [7]	87.51	1,267.53	95.36	91.26
SwinUNETR [6]	25.12	909.26	95.37	91.31
U-Mamba [14]	244.10	12,521.27	95.89	92.23
LightM-UNet	1.09	267.19	96.17	92.74

LightM-UNet achieves a mIoU improvement of 3.63%. Importantly, as a method incorporating Mamba into the UNet architecture, LightM-UNet utilizes only 1.07% fewer parameters and 2.53% less computational resources compared to U-Mamba [14].

The experimental results for the Montgomery&Shenzhen datasets [9] are summarized in Table. 2. LightM-UNet once again achieves optimal performance and significantly surpassed other Transformer-based and Mamba-based literature. Besides, LightM-UNet stands out for its remarkably low parameter count, utilizing only 1.09M parameters. This represents a reduction in parameters by 99.14% and 99.55% compared to nnU-Net [8] and U-Mamba [14], respectively. For a more clear visualization of the experimental findings, please refer to Fig. 1. Fig. 3 illustrates segmentation result examples demonstrating that compared to other models, LightM-UNet exhibits smoother segmentation edges and does not produce erroneous identification for small objects (such as tumors).

Ablation Results. We conduct extensive ablation experiments to demonstrate the effectiveness of our proposed modules. We first analyze the performance of CNN, Transformer, and Mamba within the UNet framework. Specifically, we replace the VSS Module in LightM-UNet with a convolution operation with a 3×3 kernel for CNN and with the self-attention mechanism for the Transformer. Considering the memory constraints, for CNN, we replace all VSS Modules in LightM-UNet, while for the Transformer, we follow the design of TransUNet [2] and only replace the VSS Modules in the Bottleneck Block. The experimental results on LiTS dataset [1] are shown in Table. 3, indicating that replacing

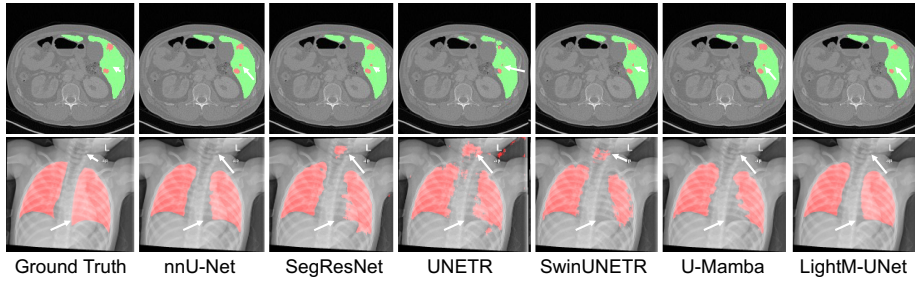


Fig. 3. Visualized segmentation examples of LiTS [1] (1st row, **red** parts indicate tumor and **green** parts indicate liver) and Montgomery&Shenzhen [9] (2nd row, **red** parts indicate lung) datasets. The white arrows point to the parts where significant differences exist in various segmentation results.

Table 3. Ablation studies in Residual Vision Mamba Layer on the LiTS [1] dataset.

Models	Params(M)	GFLOPs	Liver		Tumor		Average	
			DSC	mIoU	DSC	mIoU	DSC	mIoU
VSSM→Conv3	18.79	1,513.44	96.11	92.56	71.22	59.76	83.67	76.16
VSSM→Self-Attention	3.44	470.50	96.09	92.54	72.53	61.06	84.31	76.80
w/o Adjustment factors	1.87	457.62	96.28	92.86	71.54	60.73	83.91	76.79
w/o Residual connections	1.87	457.62	96.22	92.76	72.53	61.32	84.38	77.04
LightM-UNet	1.87	457.62	96.31	92.92	72.86	62.05	84.58	77.48

VSSM with either Convolution or Self-Attention leads to performance sacrifices. Additionally, Convolution and Self-Attention introduces a significant number of parameters and computational overhead. Furthermore, we observe that both Transformer-based and VSSM-based results outperform Convolution-based results, demonstrating the benefits of modeling long-range dependencies.

We further remove the *Adjustment factors* and *Residual connections* in the RVM Layer. Experimental results show that after removing these two components, the model’s parameter count and computational overhead hardly decrease, but the model’s performance significantly declines (0.44%↓ and 0.69%↓ mIoU). This validates our basic principle of enhancing model performance without introducing additional parameters and computational overhead. The additional ablation analysis on the Montgomery&Shenzhen datasets [9] can be found in the supplementary material.

4 Conclusion

In this study, we introduce **LightM-UNet**, a lightweight network based on Mamba, which achieves state-of-the-art performance in both 2D and 3D segmentation tasks while comprising only **1M** parameters, over 99% lower parameters and significant lower GFLOPS against latest Transformer-based architectures. We validate our approach through rigorous ablation studies within a unified framework, marking the initial attempt to employ Mamba as a lightweight strategy for

UNet. Our future work includes designing a more lightweight network and validate on more datasets of multiple organs, fostering their applications in mobile health and beyond.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). *Medical Image Analysis* **84**, 102680 (2023)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)
3. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X.: Recursive generalization transformer for image super-resolution. In: *Proceedings of the International conference on learning representations* (2024)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023)
5. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging* **38**(10), 2281–2292 (2019)
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 (2021)
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. 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)
9. Jaeger, S., Candemir, S., Antani, S., Wang, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery* **4**(6), 475 (2014)
10. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
11. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18278–18289 (2023)
12. Liao, W., Xiong, H., Wang, Q., Mo, Y., Li, X., Liu, Y., Chen, Z., Huang, S., Dou, D.: Muscle: Multi-task self-supervised continual learning to pre-train deep models for x-ray images of multiple body parts. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 151–161. Springer (2022)
13. 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)
14. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024)
15. 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, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. pp. 311–320. Springer (2019)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
 17. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
 18. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–490. Springer (2023)
 19. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* **53**, 197–207 (2019)
 20. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
 21. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 32–42 (2021)
 22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
 23. 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)
 24. Zhang, Y., Peng, C., Peng, L., Huang, H., Tong, R., Lin, L., Li, J., Chen, Y.W., Chen, Q., Hu, H., et al.: Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 68–77. Springer (2021)
 25. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)