GLM-SFNet: Global-Local Vision-Mamba with Semantic Fusion for Medical Image Segmentation

Jiahui Chen[1,†], Fei Qi[1,†,⋆,[0000−0002−2161−1551]], Chengyuan Chang[1], Qinjie Hu[1], Kaiwen Fu[1], Xiaotian Wang[1], and Kun Liu[2,[0000−0002−5034−9249]]

[1] Xidian University, Xi'an, Shaanxi 710126, China
[2] Hebei University of Technology, Tianjin, 300401, China

Abstract. Mamba-based architectures have shown promising performance in medical image segmentation. Accurate segmentation demands effective capture and integration of both global context and local details. However, existing methods often lack a balanced approach to extracting and fusing global and local information within the encoder and decoder. To address this issue, we introduce Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), which is designed for balanced global-local feature processing in medical image segmentation. In the encoder, GLM-SFNet employs a Local-Global Vision State Space block (LGVSS). LGVSS strategically integrates four-directional scanning Mamba to capture comprehensive global context while incorporating Learnable Descriptive Convolution (LDC) to ensure detailed local feature extraction. For the decoder, we propose a Semantic Fusion Decoder (SFD), which achieves enhanced information integration and boundary precision by strategically combining global and local semantic fusion modules. Extensive experiments on three benchmark datasets demonstrate that GLM-SFNet achieves state-of-the-art segmentation performance while maintaining a lightweight architecture.

Keywords: Medical Image Segmentation · Mamba · Attention · Semantic Fusion · Global Context · Local Features

1 Introduction

Segmentation is a fundamental task in medical image analysis [26], providing essential visual references for clinical diagnosis and improving both efficiency and accuracy. It has been widely applied in skin lesion analysis [2, 5] and organ segmentation in abdominal CT scans [12].

U-Net [20] represents a breakthrough in medical image segmentation, with its encoder-decoder structure and skip connections inspiring numerous improvements [16, 1]. While U-Net and its variants [1] effectively capture local features, convolutional kernels inherently struggle to model long-range dependencies. SwinUNet [3] incorporates Swin Transformer [15, 24] to enhance global

[⋆] Corresponding author: fred.qi@ieee.org.
[†] Equal contributions.

context modeling but at the cost of detail loss and increased computational overhead. TransUNet [4] and MISSFormer [10] leverage both convolutional and transformer-based approaches.The high computational complexity of models employing Transformer [3, 4, 10] limits their the clinical applications.

Recently, Mamba [7] has reduced the computational complexity to linear time, while preserving global modeling capabilities for long sequences, as compared to the quadratic complexity of the Transformer [24]. Mamba has been applied to computer vision [8, 31, 14] for its efficiency. Specifically, SliceMamba [6], VM-UNet [21], VM-UNet v2 [30], and UltraLight VM-UNet [28] have adopted Mamba architecture to efficiently learn visual representations in medical image segmentation.

However, focusing mainly on global context modeling, these Mamba-based segmentation models [6, 21, 30, 28] often have limitations in preserving spatial local details, hindering for accurate boundary segmentation. Firstly, in the encoder, employing Mamba for long-range dependency modeling, these methods may not adequately represent local spatial relationships and geometric properties. These local details are crucial for accurately identifying organ boundaries and lesion regions [29]. Although VMamba [14] applies a cross-scan module to mitigate this issue, it still fails to maintain spatial consistency between adjacent pixels. Secondly, concerning the decoder, these methods suffer from poor semantic alignment in feature fusion due to the reconstruction-oriented upsampling. Poor semantic alignment, caused by independent upsampling of high-level features and potential spatial distortion of low-level features, reduces the semantic integrity of the fused high-resolution outputs. Moreover, existing decoders do not effectively integrate global context and local details.

To overcome above issues, we introduce a Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), which has a Local-Global Vision State Space Block (LGVSS) for enhanced local-global representation and a Semantic Fusion Decoder (SFD) for improved semantic integration. LGVSS combines a four-directional scanning Mamba pathway for long-range dependency modeling with a Learnable Descriptive Convolution (LDC) [9] branch to capture local details. SFD, a plug-and-play decoder, employs Global Cross-scale Fusion (GCF) and Local Cross-scale Fusion (LCF) modules. GCF enhances semantic alignment in deep stages through global cross-attention, and LCF aligns high-level semantic features with low-level spatial features in shallow stages using local cross-attention. The collaborative combination of GCF and LCF within SFD ensures semantic coherence and spatial continuity throughout decoding.

The main contributions of this paper are as follows: (1) We propose LGVSS, which effectively captures long-range dependencies and local details in images. (2) We propose SFD, a plug-and-play medical image segmentation decoder. It performs synchronized upsampling and feature fusion to enhance the hierarchical flow of semantic information from high-level to low-level features. (3) GLM-SFNet achieves state-of-the-art performance on three public datasets. On the Synapse dataset, it reaches an average DSC of 84.82% and an HD95 of 11.87mm.

2 Method

In this section, we introduce the overall architecture of GLM-SFNet and present details of each component.

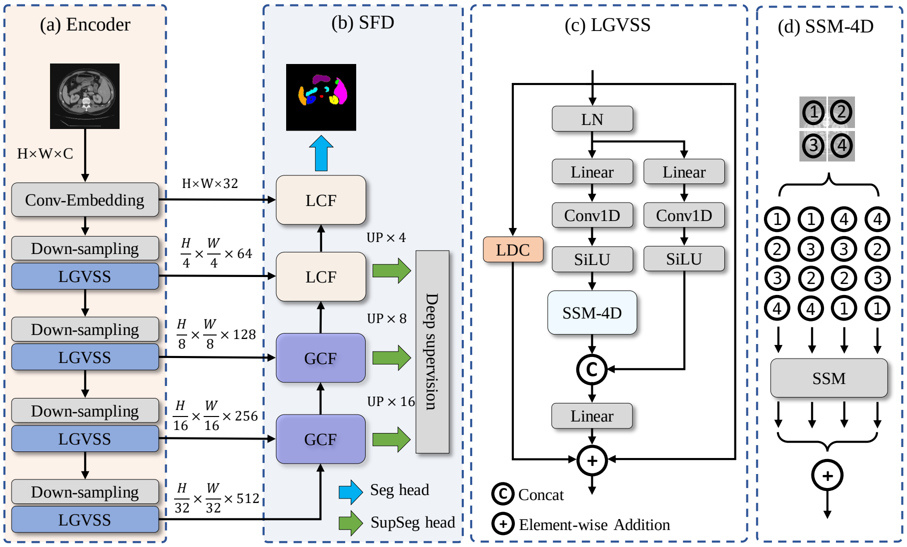

Fig. 1. The overall architecture of GLM-SFNet. (a) A multi-scale encoder with four hierarchical stages. (b) Semantic Fusion Decoder (SFD). (c) Local-Global Vision State Space Block (LGVSS). (d) Four-directional scan State Space Model (SSM-4D).

2.1 GLM-SFNet

As shown in Fig. 1, following the mainstream encoder-decoder design, the proposed GLM-SFNet consists of a Mamba-based encoder and a Semantic Fusion Decoder (SFD). According to Fig. 1(a), the encoder takes an input image $F \in \mathbb{R}^{H\times W\times C}$ and processes it through a Conv-Embedding layer, adjusting the channel dimension to produce a feature map $F' \in \mathbb{R}^{H\times W\times 32}$. This feature map F' is then passed through four consecutive stages. Each stage has a Down-Sampling layer and an LGVSS block, for multi-scale feature extraction. As illustrated in Fig. 1(b), the SFD has four stages symmetric to the encoder. In its two deeper stages, the SFD applies Global Cross-scale Fusion (GCF) to fuse high-level features with fine semantic alignment. In contrast, Local Cross-scale Fusion (LCF) integrates high-level semantics with low-level spatial features in the two shallower layers. In addition, SFD includes a Seg head to generate the final segmentation masks and three auxiliary SupSeg heads for deep supervision over intermediate outputs.

For reproducibility, we detail the implementation of the aforementioned simple layers. The Conv-Embedding comprises two 3×3 convolutional layers. Each Down-Sampling layer contains a max pooling layer followed by two 3×3 convolutional layers. The SupSeg head comprises a bilinear interpolation layer and a 1×1 convolutional layer, while the Seg head includes a 3×3 convolutional layer, a bilinear interpolation layer, and a 1×1 convolutional layer. The details of LGVSS, GCF, and LCF are provided in subsequent subsections.

2.2 Local-Global Vision State Space Block (LGVSS)

The structure of an LGVSS block is depicted in Fig. 1(c). Compared against the MambaVision Mixer [8], an improved State Space Model (SSM), the LGVSS block has a distinct main pathway, a left-side branch, and a residual connection on the right. The main pathway comprises two structural similar parallel branches. One branch utilizes the four-directional scanning SSM (SSM-4D), as illustrated in Fig. 1(d), to enhance multi-directional long-range dependency modeling. The other symmetric branch compensates for potential information loss and local forgetting caused by sequential scanning by omitting SSM-4D. The outputs of both branches are concatenated and projected through a linear layer. At the left side, the Learnable Descriptive Convolution (LDC) [9] branch captures local spatial structures in the 2D feature map by introducing learnable local descriptors to extract fine-grained texture features. The residual connection stabilizes gradient propagation. Finally, the outputs from all three branches are summed to generate the final output of LGVSS.

2.3 Global Cross-scale Fusion (GCF)

GCF integrates high-level semantic information between encoder and decoder features within deeper stages of the network. It facilitates effective feature fusion with the Multi-Head Attention (MHA) mechanism [24], as depicted in Fig. 2(a). The inputs are derived from consecutive stages of the encoder and decoder, denoted as $F_{\text{enc}} \in \mathbb{R}^{H \times W \times C/2}$ and $F_{\text{dec}} \in \mathbb{R}^{H/2 \times W/2 \times C}$, respectively.

Initially, Batch Normalization (BN) is applied to normalize features as F'_{enc} and F'_{dec}. These normalized features are subsequently fed into MHA, where F'_{enc} serves as queries (Q), while F'_{dec} provides keys (K) and values (V), each after a global linear embedding. MHA is then applied to obtain fused features, effectively integrating semantic content across scales. Since keys and values span the entire spatial space of feature maps, MHA operates globally. Finally, the fused features are further refined through channel attention and spatial attention [27] to enhance feature representation capability.

2.4 Local Cross-scale Fusion (LCF)

LCF aims to enhance semantic fusion in features of shallow stages while achieving fine-grained details. To address challenges of locality and sparsity of attention

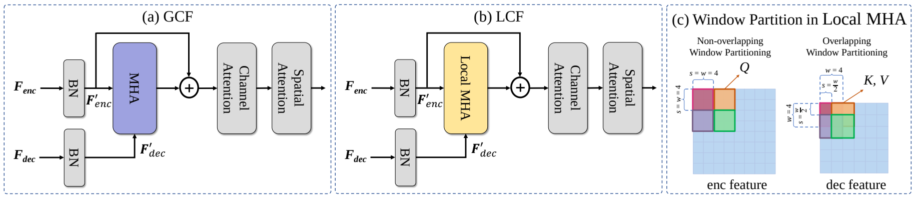

Fig. 2. (a) Global Cross-scale Fusion (GCF). (b) Local Cross-scale Fusion (LCF). (c) Non-overlapping (left) and overlapping (right) window partitioning in Local MHA.

in shallow-stage features [11], Local MHA is adopted. Local MHA can reduce redundancy and improve fusion efficiency by computing cross-attention only between corresponding local windows of encoder and decoder features.

As shown in Fig. 2(b), LCF adopts a computational process similar to GCF, but replaces MHA with Local MHA as its core component. To apply Local MHA, encoder features F'_{enc} are partitioned into non-overlapping windows of size $w \times w$ with a stride of $s = w$. The window size w is set to 4 when the spatial sizes ratio between F'_{enc} and F'_{dec} is two, whereas if the ratio is four, w is adjusted to 8. These windows are then linearly embedded to produce queries (Q). In parallel, decoder features F'_{dec} are partitioned into overlapping windows of a fixed size $w = 4$ with a stride $s = w/2$, ensuring alignment between encoder and decoder windows. These decoder windows are then projected through a Linear layer to generate keys (K) and values (V). The processes of non-overlapping and overlapping window partitioning are depicted in Fig. 2(c). This strategic alignment of window partitioning enables low-level spatial features to effectively query and integrate high-level semantic information, thereby capturing the fine-grained details essential for boundary segmentation.

In addition, LCF applies dynamic embedding on normalized decoder features F''_{dec} to derive F'_{dec} through two stacked dynamic convolutions. This process enhances local feature representation and represents another difference from GCF.

3 Experiments and Results

3.1 Datasets and implementation

Datasets and Metrics To validate the effectiveness of the proposed network, we conducted experiments on three publicly available datasets: Synapse [12], ISIC2017 [2], and ISIC2018 [5]. For Synapse, we followed the configurations of TransUNet [4] and evaluated performance with the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). For ISIC2017 and ISIC2018, we used the UltraLight VM-UNet [28] protocol, evaluating performance with DSC, Sensitivity (SE), Specificity (SP), and Accuracy (ACC). HD95 is measured in millimeters (mm), and other metrics are in percentages (%).

Table 1. Comparative experimental results on the Synapse dataset. **Bold** indicates the best performance, and underline denotes the second-best.

Model	Average		Aorta	GB	KL	KR	Liver	PC	Spleen	SM
	DSC↑	HD95↓								
U-Net [20]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
Att-UNet [17]	71.70	34.47	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
TransUNet [4]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet [3]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MT-UNet [25]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MissFormers [10]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE [18]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
TransCASCADE [18]	82.68	17.34	86.63	68.48	87.66	**84.56**	94.43	65.33	90.79	83.52
VM-UNet [21]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
SliceMamba [6]	81.95	16.04	87.78	68.77	**88.30**	84.26	95.25	64.49	86.91	79.82
PVT-EMCAD-B2 [19]	83.63	15.68	88.14	68.87	88.08	84.10	**95.26**	68.51	**92.17**	83.92
GLM-SFNet	**84.82**	**11.87**	**88.32**	**74.78**	87.49	84.35	95.14	**71.24**	91.98	**85.31**

Implementation Details GLM-SFNet was implemented using Python 3.11, PyTorch 2.3.1, and CUDA 12.1. All experiments were conducted on an NVIDIA RTX 4090 GPU. During training, we applied random rotation and flipping for data augmentation and used the AdamW optimizer. For Synapse, we set the input resolution to 224×224, the batch size to 24, and the maximum number of training epochs to 600. The initial learning rate was 0.001, and the loss function was a weighted combination of Dice loss and cross-entropy loss, with both weights set to 1. For ISIC2017 and ISIC2018, the input resolution was set to 256×256, the batch size to 8, and the maximum number of training epochs to 250. The initial learning rate was 0.01, and the loss function was a weighted combination of BCE loss and Dice loss, with weights of 1 for both.

3.2 Quantitative and qualitative segmentation results

As shown in Table 1, our GLM-SFNet achieved superior performance on the Synapse dataset, with an average DSC of 84.82% and HD95 of 11.87 mm, outperforming all baseline models. Compared with PVT-EMCAD-B2 [19], it achieved a 1.19% DSC improvement and a 3.81 mm HD95 reduction. Compared with VM-UNet [21], it achieved a 3.74% DSC improvement and a 7.34 mm HD95 reduction. For small organs, GLM-SFNet achieved a 4.19% DSC improvement for the gallbladder (GB) and a 2.73% DSC improvement for the pancreas (PC). For organs with complex boundaries, it achieved a 1.39% DSC improvement for the stomach (SM). As shown in Table 2, GLM-SFNet also performed well on the ISIC2017 and ISIC2018 datasets, achieving DSC of 92.18% and 90.64%, which are 1.27% and 1.24% higher than UltraLight VM-UNet [28]. These results confirm that GLM-SFNet can effectively model long-range dependencies while capturing fine-grained local textures, enabling precise segmentation of organ boundaries and lesion areas.

As shown in Fig. 3, we present visual comparisons between GLM-SFNet and other methods on the Synapse dataset. Our method achieves precise bound-

Table 2. Comparative experimental results on the ISIC2017 and ISIC2018 datasets. **Bold** indicates the best, and underline denotes the second-best.

Methods	ISIC2017				ISIC2018			
	DSC ↑	SE ↑	SP ↑	ACC ↑	DSC ↑	SE ↑	SP ↑	ACC ↑
U-Net [20]	89.89	87.93	98.12	96.13	88.51	87.35	97.44	95.39
VM-UNet [21]	90.70	88.37	98.42	96.45	88.91	88.09	97.43	95.54
VM-UNet v2 [30]	90.45	87.68	98.49	96.37	89.02	89.59	97.02	95.51
MALUNet [22]	88.96	88.24	97.62	95.83	89.31	88.90	97.25	95.48
LightM-UNet [13]	90.80	88.39	98.46	96.49	88.98	88.29	97.65	95.55
EGE-UNet [23]	90.73	89.31	98.16	96.42	88.19	90.09	96.38	95.10
UltraLight VM-UNet [28]	90.91	90.53	97.90	96.46	89.40	86.80	**97.81**	95.58
GLM-SFNet	**92.18**	**90.87**	**98.53**	**97.08**	**90.64**	**90.59**	97.50	**96.04**

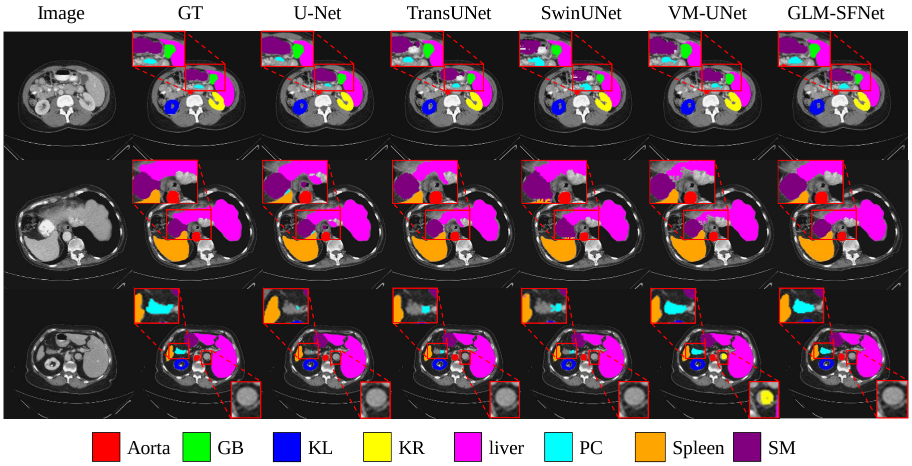

Fig. 3. Qualitative results of different methods.

ary delineation for complex-shaped organs, as demonstrated in stomach (SM) segmentation (first row) and liver segmentation (second row). For small organs like pancreas (PC) and kidney right (KR) (third row), GLM-SFNet shows superior performance while effectively addressing over-segmentation and under-segmentation issues. This stems from our encoder's effective modeling of long-range dependencies and accurate capture of local textural details, which provide comprehensive multi-scale features. The decoder further improves performance through cross-scale semantic fusion of hierarchical encoder features, achieving refined boundary accuracy.

3.3 Ablation Study

Table 3 presents the ablation experiments conducted on the Synapse dataset to validate the effectiveness of SFD. We replaced the decoders of PVTv2-EMCAD-B0, PVTv2-EMCAD-B2 [19], and SwinUNet [3] with SFD and compared their

Table 3. Effectiveness of SFD (ours) on the Synapse dataset.

Encoders	PVTv2-B0		PVTv2-B2		SwinUNet Encoder	
Decoders	EMCAD	SFD	EMCAD	SFD	SwinUNet Decoder	SFD
DSC	81.97	**83.14**	83.63	**84.42**	79.13	**82.60**
HD95	17.39	**17.10**	15.68	**13.58**	21.55	**14.52**

Table 4. Effect of different components in LGVSS.

LGVSS		DSC	Params(M)	GFLOPs
LDC	#SD			
No	1	83.18	7.74	4.16
No	2	83.53	7.74	4.17
No	4	84.05	7.74	4.18
Yes	4	**84.82**	14.35	4.64

Table 5. Impact of the global and local component combination in SFD.

SFD stages				DSC	Params(M)	GFLOPs
L	L	L	L	75.77	17.64	4.64
L	L	L	G	77.67	15.04	4.62
L	L	G	G	**84.82**	14.35	4.64
L	G	G	G	83.27	14.16	4.90
G	G	G	G	80.35	14.10	14.81

performances. The results demonstrated that SFD improved DSC of the three models by 1.2%, 0.79%, and 3.47%, and reduced the HD95 by 0.29 mm, 2.1 mm, and 7.03 mm, respectively. These results confirmed the advantages of SFD in cross-scale semantic fusion and delineate complex boundary shapes.

Table 4 evaluates the impact of the LDC branch and the number of Scanning Directions (#SD) in LGVSS. Firstly, without LDC, unidirectional, bidirectional, and four directional (4D) scanning were applied to analyze the effect of the number of SDs. The results showed that increasing the number of SDs effectively enhanced the performance of LGVSS, while the number of parameters remained unchanged due to shared scanning components, and GFLOPs slightly increased. Adopting LDC with the 4D scanning scheme, DSC was further improved by 0.77%, with moderate increase in parameters and minor rise in GFLOPs.

Table 5 further analyzes the different combinations of LCF (L) and GCF (G) in SFD to determine the optimal decoding configuration. The table lists the four stages of the decoder from shallow to deep layers from left to right. The experimental results revealed that deep-layer features contain rich semantic information, while shallow-layer features exhibit abundant local details, both are crucial for segmentation. Ultimately, we found that the LLGG configuration achieved the best balance between performance and computational efficiency.

4 Conclusions

In this study, we introduced the Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), designed to achieve a balanced integration of global contexts and local details in medical image segmentation. Our innovation lies in both encoder and decoder. In encoder, the Local-Global Vision State Space (LGVSS) block enables comprehensive global context capture and detailed local feature extraction. The Semantic Fusion Decoder (SFD) enhances information integration and boundary precision. Extensive experiments on three public datasets demonstrate that this balanced design improves segmentation performance over existing state-of-the-art methods. Our work demonstrates that a balanced encoder-decoder architecture is critical in achieving better segmentation results.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers. 62176201 and 62173124), and the Beijing-Tianjin-Hebei Collaborative Basic Research Special Project (grant number F2024202102).

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

References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical Image Segmentation Review: The success of U-Net (Nov 2022)
2. Berseth, M.: ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection (Mar 2017)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In: ECCV 2022 Medical Computer Vision Workshop. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2022)
4. 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 (Feb 2021)
5. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
6. Fan, C., Yu, H., Huang, Y., Wang, L., Yang, Z., Jia, X.: Slicemamba with neural architecture search for medical image segmentation. arXiv preprint arXiv:2407.08481 (2024)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25261–25270 (2025)
9. Huang, P.K., Ni, H.Y., Ni, Y.Q., Hsu, C.T.: Learnable Descriptive Convolutional Network for Face Anti-Spoofing. In: 33rd British Machine Vision Conference (BMVC). BMVA Press, London, UK (2022-11-21/2022-11-24)
10. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: MISSFormer: An Effective Transformer for 2D Medical Image Segmentation. IEEE Transactions on Medical Imaging **42**(5), 1484–1494 (May 2023)
11. Jiao, J., Tang, Y.M., Lin, K.Y., Gao, Y., Ma, A.J., Wang, Y., Zheng, W.S.: Dilateformer: Multi-scale dilated transformer for visual recognition. IEEE Transactions on Multimedia **25**, 8906–8919 (2023)

12. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge. vol. 5, p. 12. Munich, Germany (2015)
13. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. arXiv preprint arXiv:2403.05246 (2024)
14. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: VMamba: Visual State Space Model. Advances in Neural Information Processing Systems **37**, 103031–103063 (Jan 2025)
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. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Mar 2021)
16. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3523–3542 (Jul 2022)
17. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. In: Medical Imaging with Deep Learning. Amsterdam (4-6 July 2018)
18. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6222–6231 (2023)
19. Rahman, M.M., Munir, M., Marculescu, R.: EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer, Cham (Oct 2015)
21. Ruan, J., Li, J., Xiang, S.: VM-UNet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
22. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: MALUNet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1150–1156. IEEE (2022)
23. 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)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
25. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 2390–2394. IEEE (2022)
26. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET Image Processing **16**(5), 1243–1267 (2022)

27. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
28. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight VM-UNet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arXiv preprint arXiv:2403.20035 (2024)
29. Yang, Z., Farsiu, S.: Directional connectivity-based segmentation of medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11525–11535 (2023)
30. Zhang, M., Yu, Y., Jin, S., Gu, L., Ling, T., Tao, X.: VM-UNET-V2: Rethinking vision mamba unet for medical image segmentation. In: International Symposium on Bioinformatics Research and Applications. pp. 335–346. Springer (2024)
31. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In: Proceedings of the 41st International Conference on Machine Learning. pp. 62429–62442. PMLR, Vienna, Austria (Jul 2024)