HEDS-Net: A Hybrid State-Space Architecture with Axial Bridge and Progressive Weighting for Medical Image Segmentation

Jiaying Fan[1,4,5][0009−0006−4138−3417], Yongjie Liang[2][0009−0001−1059−088X], Junyue Cao[3,4][0000−0002−6334−3874], Peiyuan Wang[1][0009−0009−3742−672X], Bizhong Wei[1,4,5][0009−0000−1523−1125]*, and Yuexiang Li[2,4,5][0000−0001−8076−2619]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, 541004, China
[2] Life Sciences Institute, Guangxi Medical University, Nanning, 530021, China
[3] School of Information and Management, Guangxi Medical University, Nanning, 530021, China
[4] Guilin University of Electronic Technology Nanning Research Institute, Nanning, 530201, China
[5] Guangxi Academy of Artificial Intelligence, Nanning, 530201, China

Abstract. Medical image segmentation is fundamental to computer-aided diagnosis, requiring precise extraction of anatomical structures and pathological regions. While state-space models (SSMs) have demonstrated remarkable capabilities in capturing long-range dependencies with linear computational complexity, existing methods often struggle to effectively balance global context modeling and local detail preservation in medical images. To address this limitation, this paper proposes HEDS-Net, a Hybrid Vision State-space Transformer and Enhanced Deep Supervision Network. The approach introduces a Hybrid Vision State-space Transformer (HVST) module that integrates global sequential modeling with multi-scale local enhancement through a smooth progressive fusion mechanism, an Axial Cross-Information Synthesis Bridge (AXIS-Bridge) that enhances skip connections to refine cross-scale features, and a Progressively Weighted Deep Supervision Head (DS-Head) with dynamically adjusted weights for hierarchical auxiliary supervision. Extensive experimental evaluations on four medical image segmentation datasets, *i.e.,* ISIC17, ISIC18, Kvasir-SEG, and Synapse, demonstrate that HEDS-Net outperforms advanced baseline models on key evaluation metrics: in skin lesion and polyp segmentation tasks, it achieves precise lesion detection and background exclusion; in multi-organ segmentation, it enables stable segmentation of both small and large anatomical structures. Code is available at `https://github.com/Momentisbeused/HEDS-Net`.

Keywords: Medical image segmentation · Hybrid vision state-space transformer · Deep supervision.

* Corresponding author: Bizhong Wei (Email: wbz@guet.edu.cn)
Jiaying Fan, Yongjie Liang, and Junyue Cao contribute equally to this work.

1 Introduction

Medical image segmentation is fundamental to computer-aided diagnosis, requiring precise extraction of anatomical structures and pathological regions [25]. CNN-based models, represented by U-Net [16], are effective in capturing local spatial patterns but have limited capacity for long-range dependency modeling [17]. Transformer-based methods such as Swin-UNet [4] and TransUNet [6] improve global context modeling, while hybrid architectures such as TransFuse [29] and MEW-UNet [20] further explore complementary local-global representation learning through branch fusion or multi-axis modeling. SSM/Mamba-based methods provide another efficient route for long-range modeling: VM-UNet [18] and HC-Mamba [28] adapt Mamba-style blocks to medical segmentation, SegMamba [26] extends sequential modeling to volumetric segmentation, and LKM-UNet [23] and GLM-SFNet [5] further investigate large-kernel or global-local Vision-Mamba designs. These studies indicate that accurate segmentation depends not only on stronger global modeling, but also on how global semantics, local details, and decoder-side feature fusion are coordinated. This remains challenging in low-contrast lesions, scale-varying polyps, and multi-organ CT structures, where coarse semantic consistency and fine boundary preservation must be achieved simultaneously.

To address these challenges, this paper proposes HEDS-Net, a Hybrid Vision State-space Transformer and Enhanced Deep Supervision Network. It follows a U-shaped encoder–decoder architecture with three coordinated innovations: (1) HVST couples a VSS branch [13] for global sequential modeling with an ADCA branch for multi-scale local enhancement through Smooth Progressive Fusion, where early VSS-dominant fusion stabilizes semantic organization and later ADCA-enhanced fusion refines boundary-sensitive details; (2) AXIS-Bridge improves skip connections through decoder-conditioned feature mixing and axis-wise spatial refinement; and (3) DS-Head provides progressively weighted auxiliary supervision during training while inference relies on the main output. Extensive experiments on four datasets demonstrate the superior performance of our HEDS-Net.

2 HEDS-Net

HEDS-Net is a U-shaped encoder–decoder architecture designed to integrate long-range semantic modeling, boundary-sensitive local enhancement, and decoder-aware skip refinement. As shown in Fig. 1, the encoder is built with HVST blocks that combine Visual State-Space modeling and ADCA-based local enhancement, while the decoder uses AXIS-Bridge to refine skip fusion with decoder context. DS-Head introduces progressively weighted auxiliary supervision during training to stabilize multi-scale optimization. The three proposed modules are detailed below.

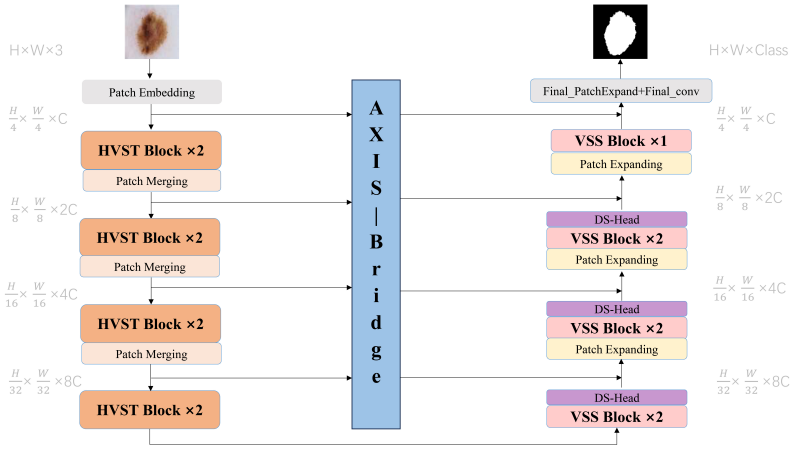

Fig. 1. HEDS-Net architecture. A U-shaped encoder-decoder with HVST, AXIS-Bridge, and DS-Head for global-local modeling, axis-aware skip fusion, and progressively weighted supervision.

2.1 Hybrid Vision State-space Transformer (HVST)

As summarized in Fig. 2, HVST is designed to couple efficient global sequence modeling with local detail enhancement within one encoder block. The Visual State-Space (VSS) branch, derived from VMamba [13], performs directional selective scanning for long-range dependency modeling with linear complexity. In parallel, the Atrous Dilated Channel Attention (ADCA) branch enriches boundary-sensitive local cues through multi-rate depthwise separable convolutions and channel attention. Smooth Progressive Fusion then makes the VSS stream dominant in early optimization and gradually increases the ADCA contribution for local refinement, followed by a residual DropPath connection.

Visual State-Space (VSS) treats spatial tokens as sequences and applies four-direction selective scanning to capture long-range dependencies with linear complexity [13]. **Atrous Dilated Channel Attention (ADCA)** expands local receptive fields with multi-rate depthwise separable convolutions and channel attention, enhancing scale-aware local responses.

Smooth Progressive Fusion blends the two streams with a scheduled local weight:

$$F_{\text{fused}} = (1 - \lambda_{\text{local}})F_{\text{vss}} + \lambda_{\text{local}}F_{\text{local}}, \tag{1}$$

where λ_{local} increases smoothly according to a cosine schedule during training. This schedule first emphasizes VSS-driven semantic organization and then strengthens ADCA-based local refinement, allowing the fused representation to evolve from stable global structure to boundary-sensitive details. The block outputs Input + DropPath(F_{fused}).

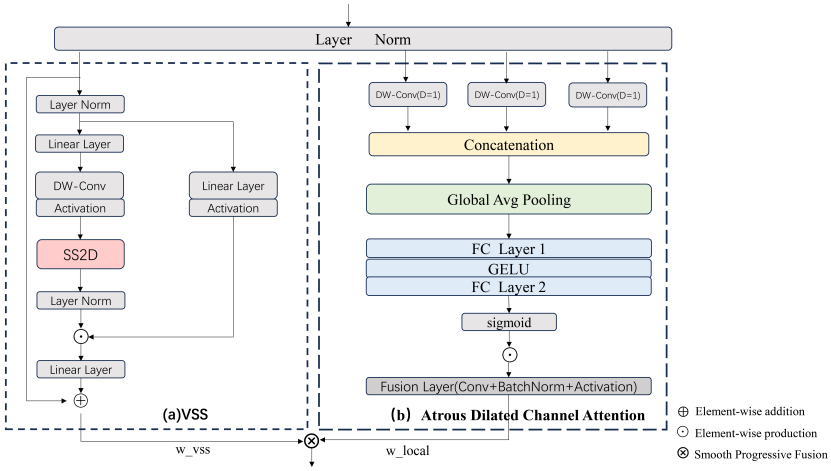

Fig. 2. Detailed Architecture of the Hybrid Vision State-space Transformer (HVST) module. (a) VSS; (b) Atrous Dilated Channel Attention.

2.2 Axial Cross-Information Synthesis Bridge (AXIS-Bridge)

The Axial Cross-Information Synthesis Bridge (AXIS-Bridge), visualized in Fig. 3, strengthens skip fusion by aligning semantics and handling spatial anisotropy common in vessels and organ boundaries [27]. It performs two steps: context-conditioned mixing between encoder and decoder features, followed by axis-wise refinement to retain fine structures. Given encoder features $X \in \mathbb{R}^{B\times H\times W\times C_{\text{enc}}}$ and decoder features $g \in \mathbb{R}^{B\times H\times W\times C_{\text{dec}}}$, both are first projected to a shared channel dimension via 1×1 convolution:

$$x_{\text{proj}} = \text{Conv}_x(X), \;\; g_{\text{proj}} \;\; = \text{Conv}_g(g). \tag{2}$$

AXIS-Bridge derives a context gate from the sum of projected features to modulate the encoder skip based on decoder semantics. Axis-wise averaging along horizontal and vertical directions produces direction-aware descriptors, which are fused to generate attention weights emphasizing elongated structures. A final 1×1 convolution aligns channels before residual addition to the decoder stream.

2.3 Progressively Weighted Deep-Supervision Head (DS-Head)

An overview of DS-Head is given in Fig. 4. DS-Head adds lightweight auxiliary heads to intermediate decoder stages for hierarchical guidance and easier optimization. Each head predicts a mask (sigmoid for binary, softmax for multi-class), is upsampled to label resolution, and is trained with BCE+Dice or CE+Dice. Epoch-dependent weights $w_\ell(e)$ gradually decay shallow heads so deeper semantics dominate later training. During inference, auxiliary supervision is disabled and the final prediction is produced by the main head.

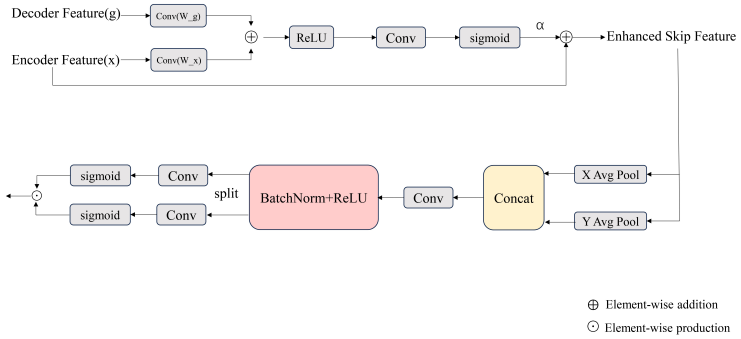

Fig. 3. AXIS-Bridge: context-conditioned mixing of encoder–decoder features, followed by axis-wise refinement to produce the enhanced skip.

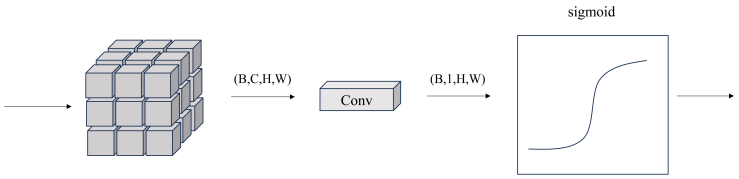

Fig. 4. Design of the Progressively Weighted DS-Head.

3 Experiments

Datasets. HEDS-Net is evaluated on four public datasets: ISIC17 [8] and ISIC18 [7] for skin lesion segmentation (256×256, 1500/650 and 1886/808 train/test), Kvasir-SEG [10] for polyp segmentation (256×256, 657/165 train/test), and Synapse [12] for multi-organ CT segmentation (224×224, 18/12 train/test).

Evaluation Metrics. For ISIC17, ISIC18 and Kvasir-SEG datasets, mIoU, Dice similarity coefficient (DSC), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe) are reported. For Synapse, mean DSC and HD95 are evaluated.

Implementation Details. HEDS-Net is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU. Inputs are resized to 256×256 for ISIC17, ISIC18 and Kvasir-SEG, and 224×224 for Synapse, normalized to $[0, 1]$, and augmented with random horizontal/vertical flips and rotations. AdamW is used with an initial learning rate of 1e−3, cosine decay, and standard weight decay; all datasets share the same 300-epoch schedule. For binary datasets, a hybrid BCE+Dice loss is employed; for Synapse, multi-class cross-entropy is combined with mean Dice over the eight abdominal organs.

Efficiency Analysis. Computational complexity is profiled at task-specific input resolutions. HEDS-Net has approximately 44.6M parameters and requires

Table 1. Comparative experimental results on ISIC17, ISIC18, and Kvasir-SEG binary segmentation datasets (Bold indicates the best).

Dataset	Model	mIoU (%)	DSC (%)	Acc (%)	Spe (%)	Sen (%)
ISIC17	UNet [16]	76.98	86.99	95.65	97.43	86.82
	UTNetV2 [9]	77.35	87.23	95.84	98.05	84.85
	TransFuse [29]	79.21	88.40	96.17	97.98	87.14
	MALUNet [19]	78.78	88.13	96.18	**98.47**	84.78
	HC-Mamba [28]	79.27	88.18	95.17	97.47	86.99
	VM-UNet [18]	80.23	89.03	96.29	97.58	89.90
	HEDS-Net (ours)	**81.72**	**89.94**	**96.59**	97.70	**91.08**
ISIC18	UNet [16]	77.86	87.55	94.05	96.69	85.86
	UNet++ [30]	78.31	87.83	94.02	95.75	88.65
	Att-UNet [15]	78.43	87.91	94.13	96.23	87.60
	UTNetV2 [9]	78.97	88.25	94.32	96.48	87.60
	SANet [24]	79.52	88.59	94.39	95.97	89.46
	TransFuse [29]	80.63	89.27	94.66	95.74	**91.28**
	MALUNet [19]	80.25	89.04	94.62	96.19	89.74
	VM-UNet [18]	81.35	89.71	94.91	96.13	91.12
	HC-Mamba [28]	80.60	89.25	94.84	**97.08**	87.90
	HEDS-Net (ours)	**81.84**	**90.01**	**95.09**	96.41	90.98
Kvasir-SEG	U-Net [16]	74.60	81.82	90.01	93.27	**82.58**
	ResUnet++ [11]	79.37	81.98	91.70	93.79	75.67
	V-Net [14]	68.97	82.52	93.29	95.42	78.25
	Swin-UNet [4]	75.58	86.27	91.26	92.16	76.38
	TransUNet [6]	76.05	86.41	92.35	97.26	77.17
	VM-UNet [18]	70.84	82.93	94.87	97.69	79.66
	HEDS-Net (ours)	**81.43**	**87.97**	**96.20**	**98.49**	80.25

6.06/4.65 GFLOPs for 256×256 binary and 224×224 Synapse segmentation. On a single NVIDIA RTX 3090 GPU in FP32 with batch size 1, latency/peak memory averaged over 100 runs after 50 warm-up iterations are 19.85 ms/329.7 MB and 19.70 ms/327.0 MB for the two settings, indicating manageable cost for global–local modeling.

3.1 Comparative Results

HEDS-Net is compared with representative CNN-, Transformer-, and SSM-based segmentation models, including UNet [16], Att-UNet [15], TransUNet [6], Swin-UNet [4], and VM-UNet [18]. As shown in Table 1, HEDS-Net delivers consistently strong performance across ISIC17, ISIC18, and Kvasir-SEG, outperforming representative CNN-, Transformer-, and SSM-based baselines on most key metrics. The hybrid HVST encoder and AXIS-Bridge decoder jointly support complete lesion coverage and precise boundary localization, with improved mIoU and DSC over baseline models. As shown in Table 2, HEDS-Net achieves the highest overall DSC (81.41%) and the smallest HD95 (15.79 mm) on Synapse, with consistent improvements for both small and large organs, indicating effective mitigation of encoder–decoder semantic gaps.

Across heterogeneous datasets, HEDS-Net improves mIoU/DSC while maintaining a balanced sensitivity-specificity trade-off. The larger Kvasir-SEG margin

Table 2. Comparative experimental results on the Synapse multi-class abdominal organ segmentation dataset (Bold indicates the best in each column).

Model	DSC (%)	HD95 (mm)	Aor.	Gal.	Kid.(L)	Kid.(R)	Liv.	Pan.	Spl.	Sto.
V-Net [14]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	59.98
DARR [1]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [6]	74.68	36.87	87.47	66.36	80.60	78.19	93.74	56.90	85.87	74.16
UNet [16]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-Unet [15]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUnet [6]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransNorm [2]	78.40	30.25	86.23	65.10	82.18	78.63	94.22	55.34	89.50	76.01
Swin-UNet [4]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
TransDeepLab [3]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	**61.19**	89.00	78.40
UCTransNet [21]	78.23	26.75	-	-	-	-	-	-	-	-
MT-Unet [22]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MEW-UNet [20]	78.92	16.44	86.68	65.32	82.87	80.02	93.63	58.36	90.19	74.26
VM-UNet [18]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	**81.40**
HEDS-Net (ours)	**81.41**	**15.79**	88.40	**69.79**	**87.20**	**84.17**	**95.36**	60.62	86.71	79.01

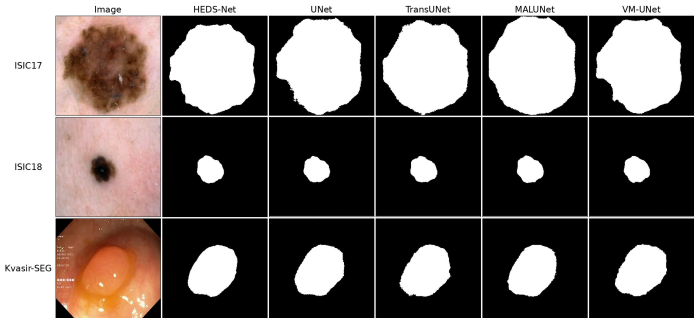

Fig. 5. Visualization of segmentation results obtained by different models on ISIC17, ISIC18, and Kvasir-SEG datasets.

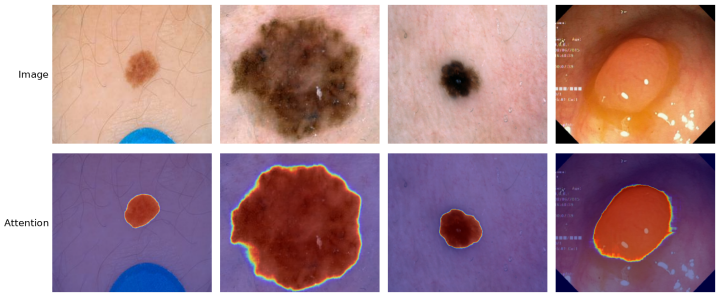

Fig. 6. Visualization of attention responses generated by HEDS-Net.

suggests better handling of variable polyp shapes, and the Synapse DSC/HD95 gains indicate robust generalization to multi-organ CT segmentation.

Table 3. Main ablation study of HEDS-Net components on ISIC17, ISIC18, and Kvasir-SEG datasets.

HVST	AXIS-Bridge	DS-Head	ISIC17		ISIC18		Kvasir-SEG	
			mIoU (%)	DSC (%)	mIoU (%)	DSC (%)	mIoU (%)	DSC (%)
–	–	–	79.93	88.71	80.47	88.87	79.68	86.72
✓	–	–	80.56	89.08	81.02	89.33	80.42	87.18
–	✓	–	80.73	89.02	81.12	89.27	80.41	87.23
–	–	✓	80.24	89.14	80.69	89.03	79.92	86.83
✓	✓	–	81.38	89.47	81.54	89.83	81.19	87.83
✓	–	✓	80.94	89.63	81.28	89.63	80.72	87.39
–	✓	✓	81.13	89.58	81.40	89.51	80.70	87.45
✓	✓	✓	**81.72**	**89.94**	**81.84**	**90.01**	**81.43**	**87.97**

3.2 Visualization Results

Qualitative evaluation on ISIC17, ISIC18, and Kvasir-SEG (Fig. 5) reveals that HEDS-Net generates segmentation masks with superior boundary precision compared to baseline models, particularly for lesions with irregular morphology or low contrast-to-background ratios. The hybrid global–local architecture in HVST facilitates accurate boundary delineation by integrating long-range contextual information with local discriminative features, which is consistent with the quantitative improvements in mIoU and DSC metrics (Table 1). On Kvasir-SEG, where polyps exhibit substantial shape and scale heterogeneity, the AXIS-Bridge mechanism preserves structural details during feature fusion, enabling robust segmentation across varying lesion sizes. The attention visualization in Fig. 6 demonstrates that HEDS-Net's progressive fusion strategy effectively concentrates attention on lesion regions while attenuating background interference, corroborating the balanced sensitivity-specificity trade-off observed in quantitative assessments.

3.3 Ablation Studies

Ablation experiments are conducted on ISIC17, ISIC18, and Kvasir-SEG. Table 3 summarizes the results when HVST, AXIS-Bridge, and DS-Head are incrementally added to the baseline. The combination of HVST and AXIS-Bridge yields substantial improvements (mIoU: 81.38%, 81.54%, 81.19%). The complete model achieves the best performance across all datasets, with mIoU improvements of 1.79%, 1.37%, and 1.75% over the baseline, confirming the synergistic effect of the three modules.

4 Conclusion and Future Work

This paper presented HEDS-Net, an encoder–decoder segmentation network with three proposed modules: HVST for Smooth Progressive Fusion of global sequential modeling and local detail enhancement, AXIS-Bridge for decoder-conditioned skip refinement, and DS-Head for progressively weighted auxiliary

supervision. Experiments on ISIC17, ISIC18, Kvasir-SEG, and Synapse show consistent improvements over representative CNN-, Transformer-, and SSM-based baselines, and ablation studies verify the complementary effects of the proposed modules. Although HEDS-Net achieves manageable inference cost, further studies are needed on more lightweight designs, finer-grained fusion variants, and broader validation in volumetric or multi-modal segmentation settings. Future work will investigate sample-adaptive fusion schedules, more fine-grained skip-refinement variants, broader external validation, and efficient extensions to multi-modal or volumetric medical image segmentation.

Acknowledgments. This study was funded by the Guangxi Natural Science Foundation (Grant No. 2025GXNSFAA069680), the "Yongjiang Plan" Leading Talent Team Program of Nanning City (Grant No. 2024020), the Basic Ability Enhancement Program for Young and Middle-aged Teachers of Guangxi (Grant No. 2025KY0157), the Innovation Project of GUET Graduate Education (Grant No. C25YJH01JZ01, GDZZ03250014), and the Innovation Project of Guangxi Academy of Artificial Intelligence Graduate Education (Grant No. SA2600000501).

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

References

1. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. Journal of Medical Imaging **6**(1), 014006–014006 (2019)
2. Azad, R., Al-Antary, M.T., Heidari, M., Merhof, D.: TransNorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. IEEE Access **10**, 108205–108215 (2022)
3. Azad, R., Heidari, M., Shariatnia, M., Aghdam, E.K., Karimijafarbigloo, S., Adeli, E., Merhof, D.: TransDeepLab: Convolution-free Transformer-based DeepLab v3+ for medical image segmentation. In: International Workshop on PRedictive Intelligence In MEdicine. pp. 91–102. Springer (2022)
4. 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: European Conference on Computer Vision. pp. 205–218. Springer (2022)
5. Chen, J., Qi, F., Chang, C., Hu, Q., Fu, K., Wang, X., Liu, K.: GLM-SFNet: Global-local vision-Mamba with semantic fusion for medical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 224–234. Springer (2025)
6. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of Transformers. Medical Image Analysis **97**, 103280 (2024)
7. 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)

8. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: International Symposium on Biomedical Imaging. pp. 168–172. IEEE (2018)
9. Gao, Y., Zhou, M., Liu, D., Metaxas, D.: A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv preprint arXiv:2203.00131 (2022)
10. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. pp. 451–462. Springer (2019)
11. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An advanced architecture for medical image segmentation. In: IEEE International Symposium on Multimedia. pp. 225–2255. IEEE (2019)
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: MICCAI Multi-atlas Labeling beyond Cranial Vault—workshop Challenge. vol. 5, p. 12. Munich, Germany (2015)
13. 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 (2024)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571. IEEE (2016)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
17. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 405–415. Springer (2023)
18. Ruan, J., Li, J., Xiang, S.: Vm-UNet: Vision Mamba UNet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
19. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: MALUNet: A multi-attention and light-weight unet for skin lesion segmentation. In: International Conference on Bioinformatics and Biomedicine. pp. 1150–1156. IEEE (2022)
20. Ruan, J., Xie, M., Xiang, S., Liu, T., Fu, Y.: MEW-UNet: Multi-axis representation learning in frequency domain for medical image segmentation. arXiv preprint arXiv:2210.14007 (2022)
21. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: UcTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
22. 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: International Conference on Acoustics, Speech and Signal processing. pp. 2390–2394. IEEE (2022)

23. Wang, J., Chen, J., Chen, D., Wu, J.: LKM-UNet: Large kernel vision Mamba UNet for medical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 360–370. Springer (2024)
24. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 699–708. Springer (2021)
25. Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L., Jiang, B.: A comprehensive review of deep learning for medical image segmentation. Neurocomputing **613**, 128740 (2025)
26. Xing, Z., Ye, T., Yang, Y., Cai, D., Gai, B., Wu, X.J., Gao, F., Zhu, L.: SegMamba-v2: Long-range sequential modeling Mamba for general 3D medical image segmentation. IEEE Transactions on Medical Imaging (2025)
27. Xu, G., Wang, X., Wu, X., Leng, X., Xu, Y.: Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey. Engineering Applications of Artificial Intelligence **142**, 109890 (2025)
28. Xu, J.: HC-Mamba: Vision Mamba with hybrid convolutional techniques for medical image segmentation. arXiv preprint arXiv:2405.05007 (2024)
29. Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing Transformers and CNNs for medical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 14–24. Springer (2021)
30. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11. Springer (2018)