MVDA-Net: Multi-View Dual-Alignment Network for 3D Carotid Artery Segmentation

Yang Wen[1], Xiaoyao Wang[1], Wenfeng Song[2], Wuzhen Shi[1(✉)], Huazhu Fu[3], Lei Zhu[4], and Bin Sheng[5]

[1] Guangdong Provincial Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, China
wzhshi@szu.edu.cn
[2] The Computer School, Beijing Information Science and Technology University, China
[3] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[4] The Hong Kong University of Science and Technology (Guangzhou), China
[5] School of Computer Science, Shanghai Jiao Tong University, China

Abstract. Accurate segmentation of the carotid artery from computed tomography angiography (CTA) images is essential to manage atherosclerotic vascular disease. However, studies specifically targeting carotid artery segmentation in CTA imaging remain relatively limited, and existing 3D segmentation methods face significant challenges when applied to this task, including discontinuous lesion distribution, extreme foreground scarcity, inadequate capture of the vessel wall's low-contrast subtle double-edge boundaries and irregular calcification contours. Moreover, these methods suffer from insufficient spatial and semantic complementarity modeling in encoder-decoder feature alignment. To address these limitations, we propose a novel Multi-View Dual-Alignment Network (MVDA-Net) that integrates three key innovations: (1) a Spatially-Aware Orthogonal Projection (SAOP) module that enhances global context modeling and representation of small vascular structures through multi-view projection across orthogonal anatomical planes with spatial attention; (2) a Hierarchical Inverted Attention Refinement (HIAR) module that propagates boundary-enhancing inverted attention from deep features to shallow layers for progressive boundary refinement; and (3) a Synergistic Feature Alignment (SFA) module that suppresses redundancy and enhances complementarity through feature differences while modeling semantic relationships via summation for effective spatial-semantic alignment. Comprehensive experiments on one in-house carotid artery CTA dataset (named CA) and three public benchmarks demonstrate that our proposed MVDA-Net achieves state-of-the-art performance. The code for MVDA-Net is publicly available at: https://github.com/Vi-cai/MVDA-Net.

Keywords: Carotid artery segmentation · 3D medical image segmentation · Multi-view projection · Feature alignment.

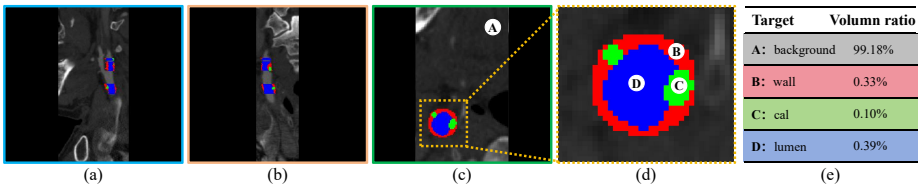

Fig. 1. Illustration of CA dataset. Anatomical views including (a) sagittal view, (b) coronal view, and (c) axial view. (d) Magnified view of the region indicated by the orange dashed box in (c). (e) Voxel-wise class distribution.

1 Introduction

Atherosclerosis underlies the majority of cardiovascular diseases and remains a major contributor to global morbidity and mortality [1]. Specifically, carotid atherosclerotic plaques pose severe clinical threats as major contributors to ischemic stroke [15]. Computed tomography angiography (CTA) offers high-resolution 3D imaging for comprehensive carotid artery evaluation. Therefore, accurate segmentation of the vessel wall, calcification, and lumen from these images is essential for disease assessment and treatment planning. Although several studies have investigated carotid artery segmentation [5, 22], dedicated approaches remain relatively limited, while generic 3D medical image segmentation frameworks [4, 6, 18] face significant challenges when directly applied to this task.

First, spatial discontinuity of lesions along the elongated vascular structure and severe foreground-background imbalance pose significant difficulties for accurate segmentation. Fig. 1 (a) and (b) illustrate the spatial discontinuity of lesions, while Fig. 1 (e) reveals extreme foreground scarcity. Current approaches such as APAUNet [8] utilizes axis projection to improve foreground-background ratios for small targets, but lacks mechanisms to emphasize critical anatomical regions; while LW-CTrans [10] applies multi-view pooling to capture global information, but inevitably loses critical spatial positional information essential for precise localization of small vascular structures.

Second, accurate boundary delineation is challenging due to low-contrast double-layer carotid vessel walls and irregular calcification contours. Current methods such as BSNet [7] employs Sobel filters to extract edge maps, while BMANet [21] applies boundary-guided multi-level attention across decoder layers. However, these approaches either depend on non-adaptive predefined operators or perform decoder-level guidance after spatial details have been lost, lacking hierarchical boundary refinement within the encoding stages.

Third, insufficient spatial and semantic complementarity modeling in encoder-decoder feature alignment hinders accurate segmentation of complex vascular structures. Current methods employ simple strategies such as UNETR++ [18] employs direct element-wise addition, D-Former [20] utilizes simple concatenation, and Swin UNETR [3] integrates residual blocks to refine encoder features.

However, these approaches fail to adaptively model their complementary characteristics, leading to suboptimal alignment.

To address these challenges, we propose a novel Multi-View Dual-Alignment Network (MVDA-Net), which consists of novel SAOP, HIAR, and SFA modules and is designed for multi-structure segmentation of the vessel wall, lumen, and calcification. Firstly, we introduce the Spatially-Aware Orthogonal Projection module to address discontinuous lesion distribution and extreme foreground scarcity by enhancing global context modeling through multi-view projection across orthogonal anatomical planes, propagating features between discontinuous regions, enhancing small target representation, and preserving spatial relationships via spatial attention. Subsequently, we design the Hierarchical Inverted Attention Refinement module, which inverts attention from deep features to amplify underemphasized boundary regions and propagates them to shallow layers for progressive boundary refinement. Furthermore, we develop the Synergistic Feature Alignment module that employs feature differences to suppress redundancy and enhance complementarity, while modeling semantic relationships via summation, synergistically achieving spatial-semantic alignment. Finally, we conduct comprehensive experiments on one in-house carotid artery CTA dataset (named CA) and three public 3D medical image segmentation benchmarks, which demonstrate the superiority of MVDA-Net across diverse anatomical structures and imaging modalities.

2 Method

As illustrated in Fig. 2, our MVDA-Net employs a four-stage encoder-decoder architecture designed for carotid artery segmentation. The 3D input is first fed through the SAOPFormer for encoding, then processed by HIAR modules for boundary refinement, subsequently aligned with decoder features via SFA modules, and finally generated the prediction result through the segmentation head.

2.1 Spatially-Aware Orthogonal Projection Transformer

To address the challenges of discontinuous lesion distribution and extreme foreground scarcity, we propose the Spatially-Aware Orthogonal Projection Transformer (SAOPFormer). As shown in Fig. 2, this architecture is built on a novel SAOPBlock, which consists of two modules: a Spatially-Aware Orthogonal Projection (SAOP) module and a convolution module.

Given an input feature map $\boldsymbol{X}$, it performs four distinct global average pooling operations in parallel to generate view-specific features. The process is defined as:

$$\hat{\boldsymbol{X}}_{\{\mathrm{S,A,C,G}\}} = \{\mathrm{SagittalGAP}, \mathrm{AxialGAP}, \mathrm{CoronalGAP}, \mathrm{GAP}\}(\boldsymbol{X}), \tag{1}$$

where SagittalGAP, AxialGAP, and CoronalGAP denote global average pooling operations along the sagittal, axial, and coronal anatomical planes, respectively, and GAP represents global average pooling across all spatial dimensions.

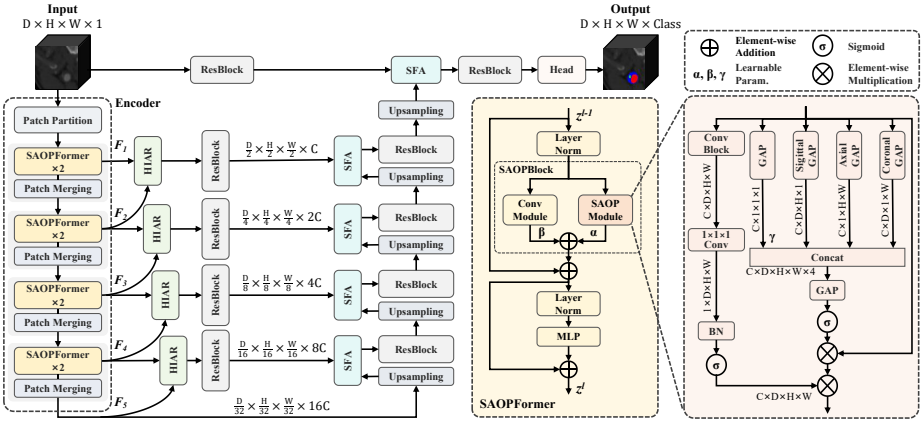

Fig. 2. The overview of the proposed MVDA-Net. MVDA-Net comprises three key modules: a Spatially-Aware Orthogonal Projection Transformer (SAOPFormer) for hierarchical feature extraction, Hierarchical Inverted Attention Refinement (HIAR) modules progressively refine boundaries, and Synergistic Feature Alignment (SFA) modules efficiently align encoder and decoder features.

Then, four view-specific features $\hat{\boldsymbol{X}}_{\{\mathrm{S,A,C,G}\}}$ are fused to recalibrate the input feature map $\boldsymbol{X}$, generating the global context representation $\boldsymbol{X}_{\mathrm{SAOP}}$. In parallel, the convolution module processes the input feature map $\boldsymbol{X}$ to produce the local feature representation $\boldsymbol{X}_{\mathrm{Conv}}$.

Finally, $\boldsymbol{X}_{\mathrm{SAOP}}$ and $\boldsymbol{X}_{\mathrm{Conv}}$ are aggregated using learnable weights α and β to dynamically balance global and local information. The output of the SAOPBlock is defined as $\boldsymbol{X}_{\mathrm{SAOPBlock}}$, computed by Eq.(2):

$$\boldsymbol{X}_{\mathrm{SAOPBlock}} = \alpha \boldsymbol{X}_{\mathrm{SAOP}} + \beta \boldsymbol{X}_{\mathrm{Conv}}. \tag{2}$$

2.2 Hierarchical Inverted Attention Refinement Module

To address the boundary delineation challenge, we propose the Hierarchical Inverted Attention Refinement (HIAR) module. This module progressively refines boundary regions by leveraging inverted spatial attention derived from deeper, semantically strong encoder features $\boldsymbol{F}_{i+1}$ to guide shallower, detail-rich features $\boldsymbol{F}_i$ $(1 \leq i \leq 4)$, where $\boldsymbol{F}_i$ denotes the output feature map of the i-th encoder stage. Adjacent encoder stages are employed to reduce semantic gaps between feature levels, enabling stable cross-level feature interaction for boundary refinement. As illustrated in Fig. 3 (a), the deeper feature $\boldsymbol{F}_{i+1}$ is first processed by a convolution and an upsampling operation. The resulting feature is then transformed into an inverted spatial attention map through sigmoid activation and subtractive inversion. Meanwhile, the shallower feature $\boldsymbol{F}_i$ is enhanced by SimAM [24] to suppress redundancy and highlight important details, yielding the enhanced feature $\boldsymbol{F}_e$. Finally, the inverted attention map is applied to $\boldsymbol{F}_e$,

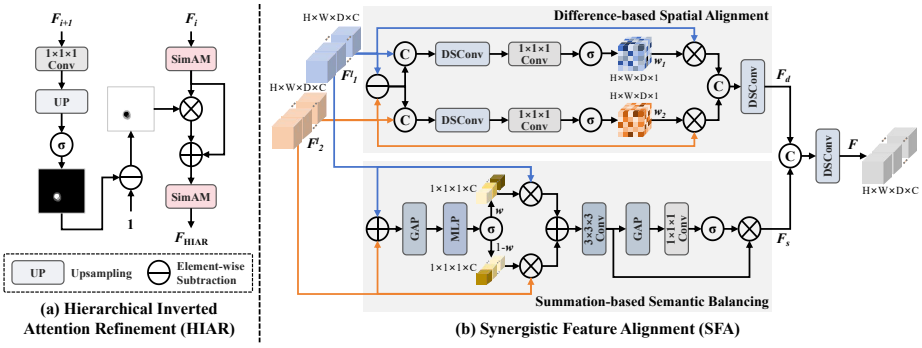

Fig. 3. (a) The Hierarchical Inverted Attention Refinement module. (b) The Synergistic Feature Alignment module.

followed by a residual connection and another SimAM module. The process is defined as follows:

$$\boldsymbol{F}_{\text{HIAR}} = \text{SimAM}(\boldsymbol{F}_e \odot (1 - \sigma(\text{UP}(\text{Conv}_1(\boldsymbol{F}_{i+1})))) + \boldsymbol{F}_e), \tag{3}$$

where Conv_1 denotes a $1 \times 1 \times 1$ convolution, UP is the upsampling operation, σ is the sigmoid function, and $\odot$ denotes the element-wise multiplication.

2.3 Synergistic Feature Alignment Module

To address insufficient spatial and semantic complementarity modeling in encoder-decoder feature alignment, we propose the Synergistic Feature Alignment (SFA) module. As shown in Fig. 3 (b), the SFA module takes encoder features $\boldsymbol{F}_1^l$ and decoder features $\boldsymbol{F}_2^l$ as inputs, processing them via two parallel branches to achieve efficient alignment.

The **Difference-based Spatial Alignment** branch first computes a feature difference map $\hat{\boldsymbol{F}}_d = \boldsymbol{F}_1^l - \boldsymbol{F}_2^l$ to explicitly model spatial complementarity, which is then concatenated with the input features to generate spatial attention weights $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ via a depthwise separable convolution, followed by a $1 \times 1 \times 1$ convolution and a sigmoid activation. The output of the Difference-based Spatial Alignment branch is defined as $\boldsymbol{F}_d$, computed by Eq.(4):

$$\boldsymbol{F}_d = \text{DSConv}((\boldsymbol{w}_1 \odot \boldsymbol{F}_1^l) \oplus (\boldsymbol{w}_2 \odot \boldsymbol{F}_2^l)), \tag{4}$$

where $\oplus$ represents the concatenation operation along the channel dimension, and DSConv denotes the depthwise separable convolution.

The **Summation-based Semantic Balancing** branch captures channel-wise semantic relationships by first aggregating encoder–decoder features via element-wise summation, followed by GAP and an multilayer perceptron (MLP) to generate a dynamic channel weight $\boldsymbol{w}$. The output of the Summation-based Semantic Balancing branch is defined as $\boldsymbol{F}_s$, computed by Eq.(5):

$$\hat{\boldsymbol{F}} = \text{Conv}_3((\boldsymbol{w} \odot \boldsymbol{F}_1^l) + ((1 - \boldsymbol{w}) \odot \boldsymbol{F}_2^l)), \quad \boldsymbol{F}_s = \sigma(\text{Conv}_1(\text{GAP}(\hat{\boldsymbol{F}}))) \odot \hat{\boldsymbol{F}}, \tag{5}$$

Table 1. Quantitative comparison with state-of-the-art methods on CA, Lung [19], Spleen [19], and BraTS [14]. Dice is adopted as our evaluation metrics (**bold** indicates the best and underline the second best results).

Methods	CA				Lung	Spleen	BraTS			
	Wall	Cal	Lumen	Avg	Tumor	Spleen	WT	ET	TC	Avg
CoTr [23]	54.55	69.29	67.91	63.92	76.85	92.46	91.58	78.89	84.03	84.83
UNETR [4]	44.44	67.90	52.88	55.07	70.92	84.23	91.47	79.90	84.76	85.38
CTN [16]	55.96	70.62	69.66	65.42	–	88.21	90.51	80.44	84.73	85.23
3D UX-NET [11]	57.47	71.06	66.94	64.16	72.79	90.32	91.36	79.18	84.18	84.91
nnFormer [25]	52.91	68.08	67.29	62.76	76.61	88.69	91.04	79.25	85.32	85.20
UNETR++ [18]	51.91	70.37	65.56	62.62	80.68	–	91.06	80.69	84.56	85.44
VSmTrans [13]	57.00	72.03	69.88	66.30	72.93	90.79	**91.86**	80.50	83.47	85.28
SegFormer3D [17]	49.18	64.54	63.25	58.99	78.88	62.26	90.76	76.70	83.52	83.66
PHNet [12]	56.26	70.46	68.93	65.22	63.72	80.53	91.54	80.60	84.54	85.56
VSNet [9]	55.31	71.01	68.02	64.78	70.40	90.31	90.36	80.42	83.70	84.83
Ours	**57.70**	**72.34**	**71.26**	**67.10**	**82.04**	**94.39**	91.71	**81.89**	**85.74**	**86.45**

where Conv_3 denotes a $3 \times 3 \times 3$ convolution.

Finally, $\boldsymbol{F}_d$ and $\boldsymbol{F}_s$ are fused to produce the output of the SFA module:

$$\boldsymbol{F} = \text{DSConv}(\boldsymbol{F}_d \oplus \boldsymbol{F}_s). \tag{6}$$

3 Experiments

3.1 Datasets

In-house carotid artery segmentation dataset (CA). Carotid artery segmentation dataset (named CA) consists of CTA volumes from 150 patients. Each patient scan was divided into left and right carotid artery regions, yielding 300 individual volumes. The segmentation targets include three structures: vessel wall, calcification, and lumen. After excluding volumes with incomplete annotations, 266 volumes were retained. We conducted five-fold cross-validation experiments. All segmentation annotations were manually delineated by experienced radiologists and validated by a senior cardiovascular imaging specialist to ensure annotation accuracy and consistency. Informed consent was waived due to the retrospective study design and use of anonymized data.

Public benchmark datasets. We validated our method on three public benchmark datasets: MSD Lung (Lung) [19], MSD Spleen (Spleen) [19], and the Brain Tumor Segmentation (BraTS) [14]. For Lung (63 CT volumes), we used a 50/13 train/test split following UNETR++ [18]. For Spleen (41 CT volumes), we used an 80:20 train/test split. For BraTS (484 MRI scans), we used an 80:15:5 train/validation/test split following nnFormer [25], and evaluated on whole tumor (WT), enhancing tumor (ET), and tumor core (TC).

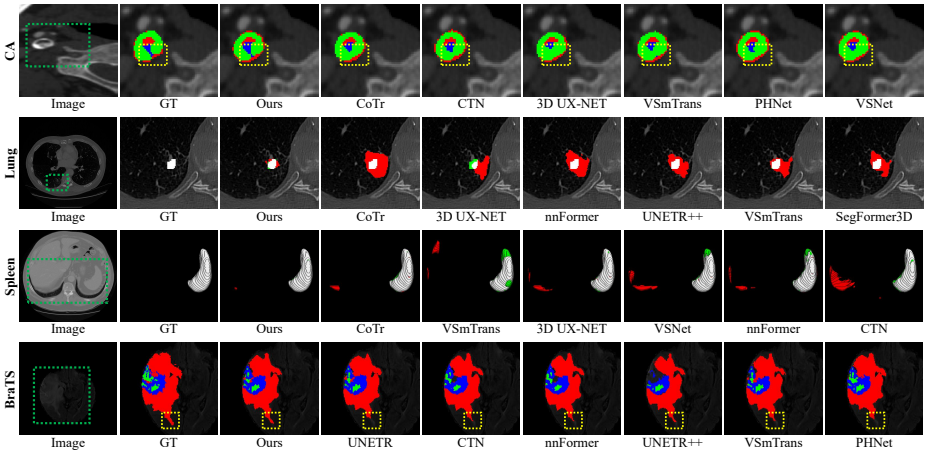

Fig. 4. Qualitative comparison of the proposed MVDA-Net (Ours) versus others on CA, Lung [19], Spleen [19], and BraTS [14].

3.2 Implementation Details

Our implementation is based on the Pytorch platform with MONAI libraries [2]. All models are trained on an NVIDIA RTX 4090 GPU with a weight decay coefficient of $3e^{-5}$. The learning rate is set to 0.001 for CA, 0.0001 for Spleen [19], and 0.01 for Lung [19] and BraTS [14] to accommodate their distinct characteristics. The input sizes are $96 \times 96 \times 96$ for both CA and Spleen [19], $192 \times 192 \times 32$ for Lung [19], and $128 \times 128 \times 128$ for BraTS [14]. The training epochs are set to 1k for CA, Lung [19], and Spleen [19], and 200 epochs for BraTS [14]. All datasets follow the nnU-Net [6] preprocessing and inference protocol.

3.3 Comparison with Other Methods

We compare our proposed MVDA-Net with 3D state-of-the-art (SOTA) CNN-based, Transformer-based, and hybrid methods, including CoTr [23], UNETR [4], CTN [16], 3D UX-NET [11], nnFormer [25], UNETR++ [18], VSmTrans [13], SegFormer3D [17], PHNet [12], and VSNet [9].

Results on CA dataset. Table 1 presents the comparative results on CA dataset. Our MVDA-Net achieves the best overall performance with an average DSC of 67.10%, surpassing VSmTrans [13] by 0.80%. Notably, our model demonstrates superior performance on challenging structures: 57.70% DSC for vessel wall segmentation and 72.34% for calcification segmentation. These results confirm that our model effectively handles complex boundary delineation in vessel walls and irregular calcification boundaries. As shown in Fig. 4, visual comparisons with six top-performing baselines demonstrate that our model achieves superior segmentation with notably finer edge details.

Results on public benchmark datasets. To further validate the generalization capability of our method, we conduct experiments on Lung [19], Spleen [19],

Table 2. Ablation study of our MVDA-Net design on CA.

Modules			CA			
SAOP	HIAR	SFA	Dice↑	Jaccard↑	Precision↑	Recall↑
–	–	–	66.30	51.91	66.35	71.56
✓	–	–	66.40	51.96	67.27	70.46
–	✓	–	66.46	52.08	66.61	71.48
–	–	✓	66.76	52.57	68.54	70.55
✓	✓	–	66.54	52.04	67.21	70.67
✓	–	✓	66.84	52.70	**69.58**	69.54
–	✓	✓	66.83	52.71	67.73	71.26
✓	✓	✓	**67.10**	**52.78**	67.38	**71.79**

Table 3. Influence of different initial weights on performance.

	Learnable	α	β	DSC↑
VSmTrans	✓	0.5	0.5	66.30
Ours	×	1.0	–	65.79
	×	0.5	0.5	66.66
	✓	0.8	0.2	66.97
	✓	0.5	0.5	**67.10**
	✓	0.2	0.8	66.86

and BraTS [14], with results presented in Table 1. Our method achieves DSC of 82.04% on Lung [19], 94.39% on Spleen [19], and 86.45% on BraTS [14] as the highest among all methods, demonstrating robust generalization across different medical image segmentation tasks. Fig. 4 presents qualitative comparisons, where the superior segmentation of our model is visually evident, particularly for irregular lung tumor boundaries.

3.4 Ablation Study

Quantitative analysis of proposed modules. We perform extensive ablation studies on CA to evaluate the contribution of each proposed module. As shown in Table 2, the baseline achieves a DSC of 66.30%, and each module individually brings clear gains. The complete model integrating all three components achieves the best overall performance, confirming synergistic operation for superior segmentation performance.

Influence of path weight. To determine the optimal fusion strategy between the SAOP module and the convolution module within SAOPBlock, we ablate different weight configurations on CA, as shown in Table 3. Our method with learnable weights initialized at 0.5 achieves the highest Dice of 67.10%, validating that the learnable weight mechanism enables adaptive balancing between global context modeling from the SAOP module and local feature extraction from the convolution module for superior segmentation performance.

4 Conclusion

In this paper, we present MVDA-Net, a U-shaped architecture for 3D carotid artery segmentation in CTA imaging, addressing the limitations of insufficient feature modeling for discontinuous lesions and extreme foreground scarcity, inadequate boundary refinement, and suboptimal encoder-decoder feature alignment. MVDA-Net achieves accurate segmentation through three synergistic contributions: the SAOP module captures comprehensive contextual information

through multi-view projection across orthogonal anatomical planes while extracting spatial positional information to address discontinuous lesion distribution and foreground scarcity; the HIAR module propagates boundary-enhancing inverted spatial attention from deep to shallow encoder layers to refine boundaries; and the SFA module enables efficient encoder-decoder feature alignment through difference-based spatial alignment and summation-based semantic balancing. Comprehensive experiments on CA and three public 3D medical image segmentation benchmarks demonstrate that MVDA-Net achieves best performance. Future work will focus on extending MVDA-Net to other vascular structures and optimizing the model for efficient clinical deployment.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 62301330), the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2024A1515010496 and 2022A1515110101), the Shenzhen Medical Research Fund (Grant Nos. E250200611 and E250200614), the Shenzhen Science and Technology Program (Grant No. 20231121103807001), the Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076), and the Shenzhen University AI4S Incubation Project (Grant No. VRLAB2026B03).

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Björkegren, J.L., Lusis, A.J.: Atherosclerosis: recent developments. Cell **185**(10), 1630–1645 (2022)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. 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)
4. 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)
5. Huang, X., Wang, J., Li, Z.: 3d carotid artery segmentation using shape-constrained active contours. Computers in Biology and Medicine **153**, 106530 (2023)
6. 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)
7. Jiang, H., Li, L.F., Yang, X., Wang, X., Luo, M.X.: Bsnet: a boundary-aware medical image segmentation network. The European Physical Journal Plus **140**(1), 53 (2025)
8. Jiang, Y., Zhang, Z., Qin, S., Guo, Y., Li, Z., Cui, S.: Apaunet: axis projection attention unet for small target in 3d medical segmentation. In: Proceedings of the Asian Conference on Computer Vision. pp. 283–298 (2022)
9. Jichen Xu, Anqi Dong, Y.Y.e.a.: Vsnet: Vessel structure-aware network for hepatic and portal vein segmentation. Medical Image Analysis (2025). https://doi.org/10.1016/j.media.2025.103458

10. Kuang, H., Wang, Y., Tana, X., Yang, J., Sun, J., Liu, J., Qiu, W., Zhang, J., Zhang, J., Yang, C.: Lw-ctrans: A lightweight hybrid network of cnn and transformer for 3d medical image segmentation. Medical image analysis p. 102 (2025)
11. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: The Eleventh International Conference on Learning Representations
12. Lin, Y., Fang, X., Zhang, D., Cheng, K.T., Chen, H.: Boosting convolution with efficient mlp-permutation for volumetric medical image segmentation. IEEE Transactions on Medical Imaging (2025)
13. Liu, T., Bai, Q., Torigian, D.A., Tong, Y., Udupa, J.K.: Vsmtrans: A hybrid paradigm integrating self-attention and convolution for 3d medical image segmentation. Medical image analysis **98**, 103295 (2024)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
15. Nagai, Y., Kitagawa, K., Sakaguchi, M., Shimizu, Y., Hashimoto, H., Yamagami, H., Narita, M., Ohtsuki, T., Hori, M., Matsumoto, M.: Significance of earlier carotid atherosclerosis for stroke subtypes. Stroke **32**(8), 1780–1785 (2001)
16. Pan, C., Qi, B., Zhao, G., Liu, J., Fang, C., Zhang, D., Li, J.: Deep 3d vessel segmentation based on cross transformer network. In: 2022 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 1115–1120. IEEE (2022)
17. Perera, S., Navard, P., Yilmaz, A.: Segformer3d: an efficient transformer for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4981–4988 (2024)
18. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Shahbaz Khan, F.: Unetr++: Delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging **43**(9), 3377–3390 (2024). https://doi.org/10.1109/TMI.2024.3398728
19. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms (2019)
20. Wu, Y., Liao, K., Chen, J., Wang, J., Chen, D.Z., Gao, H., Wu, J.: D-former: A u-shaped dilated transformer for 3d medical image segmentation. Neural Computing and Applications **35**(2), 1931–1944 (2023)
21. Wu, Z., Chen, H., Xiong, X., Wu, S., Li, H., Zhou, X.: Bmanet: Boundary-guided multi-level attention network for polyp segmentation in colonoscopy images. Biomedical Signal Processing and Control **105**, 107524 (2025)
22. Xie, H., Gu, H., Li, M., Zhu, L., Wang, T., Li, Z., Wu, H.: Carotid artery segmentation in computed tomography angiography (cta) using multi-scale deep supervision with swin-unet and advanced data augmentation. Quantitative Imaging in Medicine and Surgery **15**(4), 3161–3175 (2025)
23. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation (2021)
24. Yang, L., Zhang, R.Y., Li, L., Xie, X.: Simam: A simple, parameter-free attention module for convolutional neural networks. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:235825945

25. 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). https://doi.org/10.1109/TIP.2023.3293771