EFMS-Net: Efficient Frequency-Enhanced Multi-Scale Network for Ischemic Stroke Segmentation

Jie Yang[1], Shaowei Shen[2], Xuwei Fan[3], Ning Chen[4], Zhibin Gao[5], Lianfen Huang[2,(✉)], and Yihong Zhan[6,(✉)]

[1] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[2] School of Informatics, Xiamen University, Xiamen, China
lfhuang@xmu.edu.cn
[3] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, China
[4] College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China
[5] Navigation Institute, Jimei University, Xiamen, China
[6] The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China
yihongzhan31@163.com

Abstract. The combination of multi-modal medical imaging for ischemic stroke infarct segmentation is crucial for clinical treatment. However, existing methods often improve segmentation accuracy at the cost of efficiency, rendering them impractical for mobile health applications. To overcome this limitation, we integrate Mamba, a state-space model for long-sequence modeling, with convolutional operations to capture both global and local dependencies. To further enhance the feature representation, we incorporate multi-scale feature interaction and frequency-domain processing. As a result, we propose a novel Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net) to achieve an optimal trade-off between segmentation accuracy, inference speed, and parameter efficiency. Extensive experiments on four datasets demonstrate the effectiveness and efficiency of EFMS-Net. We release a new dataset to promote further research in ischemic stroke infarct segmentation. The dataset is available on GitHub.

Keywords: Ischemic Stroke Infarct Segmentation · Efficiency · Mamba · Feature Interaction · Frequency-domain Processing.

1 Introduction

Ischemic stroke occurs when blood supply to a specific area of the brain is partially or completely obstructed, resulting in high mortality and morbidity. Its

diagnosis and treatment heavily depend on timely intervention and neuroimaging techniques, particularly Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [1]. Employing image segmentation techniques to identify and segment the infarct core can assist clinicians in making timely and accurate treatment decisions.

In recent years, CNN- and Transformer-based models have dominated medical image segmentation. SwinUNETR [2] uses a Swin Transformer encoder [3] and a multi-scale decoder for efficient 3D segmentation, while MedNeXt [4] improves accuracy by adjusting kernel size, channels, and layers. However, challenges such as limited receptive fields and the high computational cost of long-range modeling remain, making these models impractical for resource-limited environments like mobile health applications. UNETR++ [5] addresses these with an Efficient Paired Attention (EPA) block to reduce computational burden. However, its performance on high-frequency features, such as infarct regions with blurred boundaries, remains suboptimal. Preserving high-frequency features in CT and MRI images is particularly challenging due to variations in signal intensity, texture, and shape, as well as the loss of details in deeper neural network layers. These challenges highlight the need for improved edge-feature representation. MEA-Net [6] enhances boundary information with Edge Feature Extraction (EFE), while DE-Net [7] improves boundary delineation using a specialized loss function. However, both models still struggle with the computational efficiency challenges addressed by UNETR++.

These limitations have driven researchers to explore alternative models that balance accuracy, inference speed, and computational efficiency. One promising approach is to employ Mamba [8], a State Space Sequence Model (SSM), to efficiently model long-range dependencies with linear complexity. Building on this idea, several studies [9,10,11] have integrated CNNs with Mamba, combining the former's ability to capture detailed local features with the latter's strength in modeling global dependencies. Another promising approach, instead, leverages the frequency domain to extract discriminative features and enhance semantic representation [12,13,14]. We propose an Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net) based on the typical U-Net structure [15], which is well-known for its excellent performance in medical image segmentation, and further design an efficient encoder-decoder architecture built upon it. Two novel blocks, Dual-branch Adaptive Frequency Fusion (DAFF) block and Laplacian-enhanced Multi-Scale Attention (LMSA) block, are incorporated into the encoder and skip connections of EFMS-Net.

Our primary contributions are as follows: (1) To improve segmentation in regions with fuzzy boundaries and low contrast, the DAFF block extracts global and local features via two parallel subnetworks. The block then applies the Discrete Cosine Transform (DCT) to obtain frequency representations of these features and utilizes the Adaptive Frequency Weighting (AFW) layer to adaptively adjust the weights of all frequency components. (2) The LMSA block employs direction-aware attention across multiple scales to enhance channel-spatial interactions and utilizes the parameter-free Laplacian of Gaussian operator to

improve edge representation. (3) A new dataset for 3D segmentation of ischemic stroke is proposed, comprising 120 MRI scans annotated by experts.

2 Method

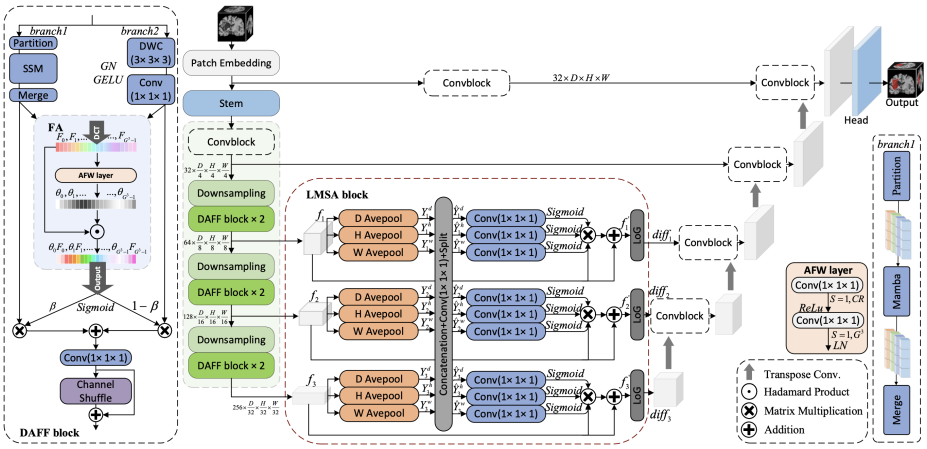

Fig. 1. The overview of the proposed EFMS-Net.

The overall architecture of EFMS-Net is illustrated in Fig. 1. The proposed model integrates DAFF and LMSA block to enhance multi-scale feature extraction and hierarchical information aggregation for stroke segmentation. The encoder begins with a stem layer that applies a $4 \times 4 \times 4$ convolution with a stride of 4 to obtain the initial feature map $f_0 \in \mathbb{R}^{C \times D/4 \times H/4 \times W/4}$ from the 3D input volume $I \in \mathbb{R}^{C \times D \times H \times W}$. This map is subsequently processed through three encoder stages, each consisting of two DAFF blocks and a downsampling layer. Additionally, the LMSA block are incorporated into the skip connections to enhance multi-scale feature interaction. The decoder gradually restores resolution through four stages by aggregating features from skip connections. Finally, transposed convolutions map the channels to segmentation targets, followed by a sigmoid activation to generate the binary segmentation mask.

Dual-branch Adaptive Frequency Fusion Block: The DAFF block consists of two parallel branches. In the first branch, inspired by [16], we use cascaded $3 \times 3 \times 3$ depthwise convolutions followed by $1 \times 1 \times 1$ standard convolutions, with channel-wise GroupNorm and GELU activation applied between the convolutional layers. The second branch employs the bidirectional Mamba model [17], where the 3D feature maps are flattened into a sequence of 1D patch embeddings. These embeddings are then partitioned into s subsets with a step size of s to capture dependencies between adjacent sampling tokens. Finally, outputs

of the two branches, Mamba and $Conv_{casc}$, are converted into their frequency domain representation F_i using modified DCT [18], which is given by:

$$F_i = \sum_{d=0}^{N-1} \sum_{h=0}^{N-1} \sum_{w=0}^{N-1} x_{d,h,w} \cos\left(\frac{(2d+1)\mu\pi}{2N}\right) \cos\left(\frac{(2h+1)\nu\pi}{2N}\right) \cos\left(\frac{(2w+1)k\pi}{2N}\right) \quad \text{s.t. } i \in \{0, 1, 2, \ldots, G^3 - 1\} \tag{1}$$

where $x_{d,h,w}$ denotes the pixel intensity at position (d, h, w) within the 3D image block $x \in \mathbb{R}^{C \times N \times N \times N}$. The channel C is split into G^3 groups, and $\mu, \nu, k \in \{0, 1, \cdots, G-1\}$ represent the frequency indices along the depth, height, and width dimensions of the block, respectively. When μ, ν and k in Formula 1 are zero, it's equivalent to Global Average Pooling (GAP), which captures only the lowest frequency component. The multi-spectral frequency vector F is given by:

$$F = cat([F_0, F_1, \cdots, F_{G^3-1}]) \tag{2}$$

The Frequency Attention (FA) mechanism is introduced to learn the importance of different frequency components. Its AFW layer is designed to generate learnable weights $\Theta \in \mathbb{R}^{G^3 \times 1 \times 1^3 \times 1}$ for each component of F. Specifically, the AFW block employs two $1 \times 1 \times 1$ convolutional layers with ReLU and LayerNorm, whose convolutional weights are trainable parameters that adaptively adjust each frequency component. The output weights Θ are normalized along the channel dimension using a SoftMax function. F is then reshaped into the shape of $\mathbb{R}^{G^3 \times \frac{C}{G^3} \times 1^3 \times 1}$, where the weight of each group corresponds to a frequency component $F_i \in \mathbb{R}^{\frac{C}{G^3} \times 1^3 \times 1}$. By performing element-wise multiplication between Θ and F, the layer re-weights frequency features and outputs a refined representation F', as follows:

$$\Theta = SoftMax(AFW(F)) \tag{3}$$

$$F' = \Theta F = cat(\theta_0 F_0, \theta_1 F_1, \ldots, \theta_{G^3-1} F_{G^3-1}) \tag{4}$$

where θ_i and F_i represent the i-th components of Θ and F, respectively. Instead of selecting only the top-k frequency components based on prior knowledge, as in FcaNet model [12], the proposed method captures a broader range of frequency components across the entire image block and dynamically reweights them.

The multi-frequency channel attention weights for global and local features are denoted as β and β', respectively. These weights adaptively modulate the features by multiplication, and the whole process can be formulated as follows:

$$\beta = Sigmoid\left(FA(Mamba(f)) + FA(Conv_{casc}(f))\right) \tag{5}$$

$$f' = \beta \times Mamba(f) + \beta' \times Conv_{casc}(f) \tag{6}$$

where $\beta, \beta' \in \mathbb{R}_{\geq 0}^{C\times 1\times 1\times 1}$ and $\beta + \beta' = 1$. The resulting feature maps f' are then fused through a $1 \times 1 \times 1$ convolution. Furthermore, a channel shuffle operation [19] is employed to enhance cross-channel communication among feature maps.

Laplacian-enhanced Multi-Scale Attention Block: To maintain the spatial consistency of multi-scale feature maps, LMSA block is applied to $\{f_l\}_{l=1}^{3}$, modeling the correlation between spatial and channel attention information. Specifically, LMSA block performs directional pooling along depth, height, and width dimensions to extract long-range dependencies while maintaining positional information in the orthogonal directions. The process is expressed as follows:

$$Y_l^d(d) = \frac{1}{H_l \times W_l} \sum_{0\leq i\leq H_l} \sum_{0\leq j\leq W_l} f_l(d_l, i, j) \tag{7}$$

$$Y_l^h(h) = \frac{1}{D_l \times W_l} \sum_{0\leq k\leq D_l} \sum_{0\leq j\leq W_l} f_l(k, h_l, j) \tag{8}$$

$$Y_l^w(w) = \frac{1}{D_l \times H_l} \sum_{0\leq k\leq D_l} \sum_{0\leq i\leq H_l} f_l(k, i, w_l) \tag{9}$$

where $Y_l^d \in \mathbb{R}^{C_l\times d_l\times 1\times 1}$, $Y_l^h \in \mathbb{R}^{C_l\times 1\times h_l\times 1}$ and $Y_l^w \in \mathbb{R}^{C_l\times 1\times 1\times w_l}$ represent the outputs of the l-th feature map along the depth, height, and width dimensions. Here, $d_l \in \{0, 1, ..., D_l - 1\}$, $h_l \in \{0, 1, ..., H_l - 1\}$, and $w_l \in \{0, 1, ..., W_l - 1\}$ are depth, height, and width indices. These outputs are adjusted to match the minimum channel dimension and then concatenated into $\mathbb{R}^{C_{min}\times(d_1+h_1+w_1+...+d_3+h_3+w_3)\times 1\times 1}$, followed by a shared $1 \times 1 \times 1$ convolution to facilitate layer-wise feature interaction. The resulting outputs are split along spatial dimensions back into $\hat{Y_l^d} \in \mathbb{R}^{C_l\times d_l\times 1\times 1}$, $\hat{Y_l^h} \in \mathbb{R}^{C_l\times 1\times h_l\times 1}$, and $\hat{Y_l^w} \in \mathbb{R}^{C_l\times 1\times 1\times w_l}$. Finally, a $1 \times 1 \times 1$ convolution followed by a sigmoid activation is applied to each $\hat{Y_l^d}$, $\hat{Y_l^h}$, and $\hat{Y_l^w}$ to get the corresponding attention weights. These weights are used to refine the original feature maps and a residual connection is introduced to preserve the original feature representations and mitigate overfitting. The process is expressed as follows:

$$f_l' = f_l + f_l \times \sigma(\mathrm{Conv}(\hat{Y_l^d})) \times \sigma(\mathrm{Conv}(\hat{Y_l^h})) \times \sigma(\mathrm{Conv}(\hat{Y_l^w})) \tag{10}$$

The output f_l' is processed through Laplacian of Gaussian (LoG) [20], defined as $\mathrm{LoG}(f) = \nabla^2(f * G)$, which smooths the features using a Gaussian filter to suppress high-frequency noise and then calculates the second-order derivative to capture edge information. However, directly computing the second-order derivative can be computationally expensive. To improve efficiency, we approximate the LoG by combining a Gaussian filter with $2\times$ upsampling and $2\times$ downsampling to extract high-frequency information. The formula is as follows:

$$diff_l = f_l - upsample(downsample(f_l * G)). \tag{11}$$

Where G is a normalized 3D Gaussian kernel of size $5 \times 5 \times 5$, and $*$ denotes the convolution operation. Downsampling is performed by slicing the feature map to

reduce its resolution, while upsampling restores the resolution by zero-padding and applying Gaussian convolution to enhance the feature details. Finally, the difference between the original and upsampled feature maps approximates the high-frequency components.

3 Experiments

3.1 Experimental Setup

We use three public datasets (ISLES'22 [21],ISLES'18 [22], ATLAS v2.0 [23]) and a self-constructed dataset for ischemic stroke lesion segmentation. The ISLES'22 dataset includes 400 multi-modal MRI scans (FLAIR, DWI, ADC) for pre- and post-operative cases. ISLES'18 has 156 records from 103 patients, featuring perfusion maps (CBF, CBV, MTT, Tmax) and CT images (excluding 4D CTP). ATLAS v2.0, one of the largest stroke imaging datasets, contains T1-weighted images from 1,271 patients. Our self-constructed dataset comprises DWI images from 120 patients, with in-plane dimensions ranging from 180×140 to 200×160, and 36 to 44 along the z-axis. All images were resampled to (3.0, 1.0, 1.0) mm voxel spacing and manually annotated by three clinicians using ITK-SNAP [24].

Table 1. Performance comparison of methods on the ISLES'18, ISLES'22, and ATLAS v2.0 datasets. The evaluation metrics are DSC (%), HD95 (mm), P (M), and FLOPS (G). Best values are in bold, and second-best values are underlined.

Methods	ISLES'18				ISLES'22				ATLAS v2.0			
	DSC	HD95	P	FLOPs	DSC	HD95	P	FLOPs	DSC	HD95	P	FLOPs
CNN-based												
nnUNet	60.5	10.1	29.9	84.9	79.4	10.9	30.8	566.0	61.1	23.0	31.2	479.0
UX-Net	55.6	15.0	51.1	300.5	80.0	9.9	51.1	3407.2	60.7	22.3	51.1	2884.5
MedNeXt	59.3	10.8	31.0	132.7	79.1	10.2	31.7	627.5	61.2	<u>20.1</u>	31.7	535.3
Transformer-based												
nnFormer	54.1	14.1	62.4	71.0	79.5	10.1	140.7	281.3	58.8	24.5	131.2	261.1
UNETR^{++}	58.6	10.5	31.1	<u>23.8</u>	80.3	9.9	46.5	**85.0**	59.1	23.1	42.6	**70.5**
Mamba-based												
U-Mamba	<u>61.4</u>	<u>8.3</u>	42.4	148.0	<u>80.4</u>	<u>8.7</u>	42.9	1165.7	<u>62.3</u>	23.3	42.8	990.6
LKM-UNet	56.5	11.1	62.7	187.8	78.4	11.1	64.0	1232.2	59.7	22.9	64.4	1047.0
LightM	53.6	12.4	<u>5.1</u>	40.3	78.7	9.1	<u>5.1</u>	<u>117.3</u>	59.3	23.6	<u>5.1</u>	<u>99.9</u>
Ours	**62.7**	**8.1**	**2.5**	**9.5**	**81.8**	**7.3**	**3.8**	224.6	**62.6**	**18.8**	**3.8**	187.8

For a fair comparison, all models were implemented in `nnUNetv2` [25] and trained with default settings for 1,000 epochs on the same preprocessed data. Performance was evaluated using 5-fold cross-validation. To reduce overfitting,

Table 2. Memory and efficiency comparison of methods on ISLES'18, ISLES'22, and ATLAS v2.0 datasets. Evaluation metrics are TM (MiB), IM (MiB) and IT (case/s).

Methods	ISLES'18			ISLES'22			ATLAS v2.0		
	TM	IM	IT	TM	IM	IT	TM	IM	IT
nnUNet	1739	1169	**0.23**	8507	2279	1.62	6531	2085	2.92
UX-Net	3537	2499	0.44	32107	3667	11.73	23775	3105	21.74
MedNeXt	3001	2411	0.29	17341	2669	2.81	14761	2565	5.21
nnFormer	2764	2348	0.28	16645	2542	2.62	14217	2456	4.81
UNETR^{++}	2143	965	0.78	5629	1471	1.75	4699	**1359**	3.12
U-Mamba	3777	2333	0.46	30432	6283	7.34	22529	5459	13.65
LKM-UNet	8981	4841	0.86	40680	6973	10.13	35035	6573	18.72
LightM	7079	5289	0.54	33503	3817	12.95	29347	3584	24.45
Ours	**1401**	**917**	0.25	**5405**	**1403**	**1.59**	**4669**	1475	**2.91**

Table 3. Ablation study for different modules of EFMS-Net on the four datasets.

Modules	ISLES'18			ISLES'22			ATLAS v2.0			self-constructed		
	DSC	P	FLOPs	DSC	P	FLOPs	DSC	P	FLOPs	DSC	P	FLOPs
Baseline	58.1	4.7	9.9	78.5	12.2	227.4	58.5	12.1	190.1	37.3	4.6	55.7
DAFF only	60.6	2.4	9.3	80.8	3.7	224.4	61.1	3.7	187.7	40.4	2.4	52.4
LMSA w/o LoG	61.1	4.7	10.0	80.6	12.2	227.5	61.3	12.2	190.2	39.6	4.7	55.7
LMSA	61.5	4.7	10.1	80.9	12.2	227.6	61.8	12.2	190.3	40.2	4.7	55.8
DAFF + LMSA	62.7	2.5	9.5	81.8	3.8	224.6	62.6	3.8	187.8	40.9	2.5	52.5

data augmentation was applied. The proposed model, built with Monai, used a batch size of 2, a learning rate of 0.001, and the SGD optimizer, with a loss function combining soft dice and cross-entropy loss. Input patch sizes were $96 \times 160 \times 160$ (ISLES'22), $5 \times 224 \times 192$ (ISLES'18), $128 \times 128 \times 128$ (ATLAS v2.0), and $40 \times 192 \times 160$ (self-constructed dataset). All experiments ran on a single A100 80GB GPU. Metrics included Dice Similarity Coefficient (DSC), 95th Percentile Hausdorff Distance (HD95), Number of Parameters (P), FLOPs, Training Memory (TM), Inference Memory (IM), and Inference Time (IT).

3.2 Experimental Results

Comparison with SOTA Methods: In this section, we compare our method with SOTA approaches on three public datasets. Specifically, we evaluate CNN-based methods [4,25,26], Transformer-based methods [5,27] and Mamba-based methods [9,10,11], The experimental results are summarized in Table 1 and Table 2. EFMS-Net demonstrates the best segmentation accuracy, with an optimal balance across the number of parameters, memory usage, FLOPs, and inference speed on the three datasets. Although the segmentation accuracy of U-Mamba is only slightly behind EFMS-Net, its parameters, FLOPs, memory usage, and in-

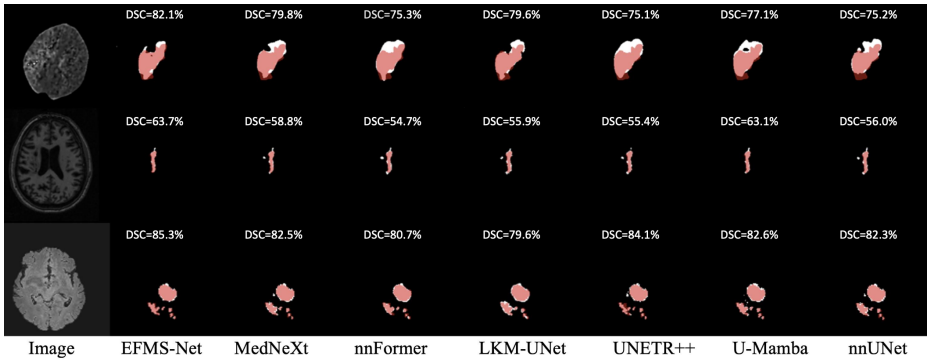

Fig. 2. Visualization of ischemic stroke infarct segmentation results. Row 1 shows ISLES'18 sample, Row 2 shows ATLAS v2.0 sample, and Row 3 shows ISLES'22 sample. Red regions indicate the ground-truth labels.

ference time are significantly worse. In the ISLES'18 dataset, EFMS-Net reduces parameters and computational complexity by 17.0× and 15.6×, respectively. In the ISLES'22 dataset, EFMS-Net uses 25,027 MiB and 4,880 MiB less memory in the training and inference stages and is about 4.6× faster than U-Mamba. While the FLOPs and inference memory of EFMS-Net are slightly higher than those of the lightweight model UNETR++ on ISLES'22 and ATLAS v2.0, EFMS-Net consistently demonstrates superior performance across all other metrics.

Visualization Results: We validate the effectiveness of the proposed method by visualizing the segmentation results on three datasets, comparing them with SOTA methods, as illustrated in Fig. 2. EFMS-Net demonstrates better alignment with ground-truth labels compared to SOTA methods. For cases with low contrast (Row 1) and small infarct cores (Row 3), it achieves more accurate delineation of infarct areas and shapes through effective integration of global and local dependencies and edge feature enhancement.

Ablation Study: We perform ablation studies on four datasets, including a self-constructed one, to explore the effects of different EFMS-Net modules, as shown in Table 3. Specifically, replacing the DAFF module with an equivalent number of Convblock modules shows that both DAFF and LMSA improve EFMS-Net's performance over the baseline, highlighting their effectiveness in capturing multi-scale features and enhancing edge details. The dual-branch design of depthwise separable convolution and Mamba greatly reduces computational cost compared to traditional convolutional operations, while FA further improves the accuracy by adaptively weighting frequency components. LMSA outperforms DAFF by leveraging multi-scale channel-spatial interactions, and its LoG operator effectively preserves high-frequency details. Integrating these components, EFMS-Net achieves the best performance.

4 Conclusion

In this paper, we propose the Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net), a novel approach to ischemic stroke infarct segmentation that strikes a balance between accuracy, inference speed, and computational efficiency. Specifically, we introduce the Dual-branch Adaptive Frequency Fusion (DAFF) block, which captures both global and local dependencies from frequency and spatial perspectives. Additionally, the Laplacian-enhanced Multi-Scale Attention (LMSA) block enables multi-scale interactive attention and employs the Laplacian of Gaussian (LoG) operator to enhance high-frequency details. Comparative experiments and ablation studies on four datasets demonstrate the superior performance of the proposed model in terms of segmentation accuracy and efficiency, especially in resource-constrained environments.

Acknowledgments. The work presented in this paper was partially supported by the 2024 National Natural Science Foundation of China (Grant No. 62371406) and Natural Science Foundation of Xiamen, China (Grant number 3502Z202473053).

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

References

1. Hui, C., Tadi, P., Suheb, M.Z., Patti, L.: Ischemic stroke. StatPearls [Internet] (2024)
2. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
4. 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)
5. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: Delving into efficient and accurate 3D medical image segmentation. IEEE Transactions on Medical Imaging (2024)
6. Liu, H., Feng, Y., Xu, H., Liang, S., Liang, H., Li, S., Zhu, J., Yang, S., Li, F.: MEA-Net: Multilayer edge attention network for medical image segmentation. Scientific Reports 12(1), 7868 (2022)
7. Gu, R., Wang, L., Zhang, L.: DE-Net: A deep edge network with boundary information for automatic skin lesion segmentation. Neurocomputing 468, 71–84 (2022)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
9. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)

10. 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)
11. 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)
12. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 783–792 (2021)
13. She, D., Zhang, Y., Zhang, Z., Li, H., Yan, Z., Sun, X.: EoFormer: Edge-Oriented Transformer for Brain Tumor Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 333–343. Springer (2023)
14. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence 38(6), 6030–6038 (2024)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer (2015)
16. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258 (2017)
17. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
18. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Transactions on Computers 100(1), 90–93 (2006)
19. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856 (2018)
20. Marr, D., Hildreth, E.: Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences 207(1167), 187–217 (1980)
21. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., Robben, D.: ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data 9(1), 762 (2022)
22. The ISLES Challenge 2018 website. Available: https://www.isles-challenge.org/ISLES2018/
23. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., Wang, Z.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific Data 9(1), 320 (2022)
24. Yushkevich, P.A., Gao, Y., Gerig, G.: ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3342–3345 (2016)
25. 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)

26. 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. arXiv preprint arXiv:2209.15076 (2022)
27. 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)