SMF-Net: Unlocking Multimodal Insights for Enhanced Stroke Lesion Segmentation

Meklit Atlaw[1], Geng Chen[1](✉), Haotian Jiang[1], Xuyun Wen[2], Hengfei Cui[1], and Yong Xia[1]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
meklitmaki29@gmail.com, geng.chen@ieee.org, jianghaotian100@163.com, hfcui@nwpu.edu.cn, yxia@nwpu.edu.cn

[2] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
wenxuyun@nuaa.edu.cn

Abstract. Stroke is a leading cause of death and disability worldwide, necessitating accurate lesion segmentation for effective diagnosis and treatment. Multimodal images provide complementary insights into stroke detection and progression. However, existing segmentation methods often struggle to fully leverage the distinct and dynamic sensitivities of these modalities. Current approaches, including encoder-decoder networks and SAM-based models, are either limited to single-modality data or rely on suboptimal fusion techniques, hindering their ability to adapt to the distinct nature of stroke lesions. To address these challenges, we propose SAM-driven Multimodal Fusion Network (SMF-Net) for enhanced stroke lesion segmentation. SMF-Net incorporates a multimodal Siamese image encoder based on the Swin Transformer to extract modality-specific features, alongside two novel fusion strategies: (1) Complementary dynamic fusion module, which uses pairwise co-attention and dynamic learnable weights to model interdependencies and adaptively combine multimodal features; and (2) Context-aware intermediate-layer fusion module, a lightweight, multi-layer fusion mechanism that captures multiscale features while preserving modality-specific information. Extensive experiments on an open benchmark dataset demonstrate that SMF-Net outperforms previous stroke lesion segmentation methods through effective multimodal integration.

Keywords: Stroke lesion segmentation, MRI, Multimodal fusion, SAM

1 Introduction

Stroke is a major neurological disorder, the second most common cause of death, and the third leading cause of disability worldwide. Its rising prevalence results in

Corresponding author: Geng Chen

millions of deaths annually, leaving many survivors with lasting impairments [3]. Accurate stroke lesion segmentation in medical imaging is crucial for assessing lesion extent and severity, enabling timely clinical decisions to optimize treatment. Various imaging modalities capture different aspects of stroke evolution and location [11]. MRI techniques, such as Diffusion-Weighted Imaging (DWI), Apparent Diffusion Coefficient (ADC), and Fluid-Attenuated Inversion Recovery (FLAIR) provide critical insight into the progression and location of the lesion. Although the use of multiple imaging modalities improves the accuracy of diagnosis and segmentation, the dynamic nature of brain tissue damage and the varying imaging sensitivities require effective integration strategies for accurate segmentation of stroke lesions.

Recent studies have explored various deep learning-based approaches for stroke lesion segmentation, particularly using encoder-decoder networks [1, 2, 9, 13, 16, 17]. D-UNet [25] combines 2D and 3D features to capture complex lesion patterns, while X-Net [14] focuses on capturing long-range dependencies to enhance segmentation accuracy. SAN-Net [21] is tailored for multi-site stroke lesion segmentation, while W-Net [19] addresses fuzzy stroke boundaries to improve precision. While these models have proven effective for single-modality data, efforts to address multi-modal imaging have led to methods such as [5, 6, 15], which attempt to integrate different modalities using image stacking or basic fusion techniques. More recently, the Segment Anything Model (SAM) [8] has inspired a wave of universal segmentation models, including MedSAM [12], Mobile-SAM [22], and EfficientSAM [20], which have shown promise in various medical imaging tasks.

Despite the advancements, these methods face significant challenges in stroke lesion segmentation. Most approaches are primarily single-modality focused, limiting their ability to capture the complementary information offered by different imaging techniques. When it comes to multimodal strategies, existing methods often rely on basic fusion techniques, such as stacking images from different modalities or using simple sub-networks for modality fusion. These methods often overlook the distinct sensitivities of each modality, leading to suboptimal utilization of modality-specific features and an inability to adapt to their changing relevance during stroke progression. Furthermore, SAM-based models and their adaptations, while effective in other tasks, are not yet optimized for the unique intensity patterns and lesion characteristics specific to stroke. In addition, these models are not tailored for multimodal integration, an essential aspect of accurate stroke diagnosis and clinical decision making. Therefore, addressing these gaps is crucial to improving stroke lesion segmentation.

To this end, we propose SAM-driven Multimodal Fusion Network (SMF-Net), a SAM-based framework leveraging Siamese networks and optimized fusion strategies to overcome challenges in multimodal stroke lesion segmentation. Particularly, we first use a multimodal Siamese image encoder based on the Swin Transformer to extract distinct features from each modality. We then introduce the Complementary Dynamic Fusion (CDF) with pairwise co-attention that models pairwise relationships between the modalities, capturing interdepen-

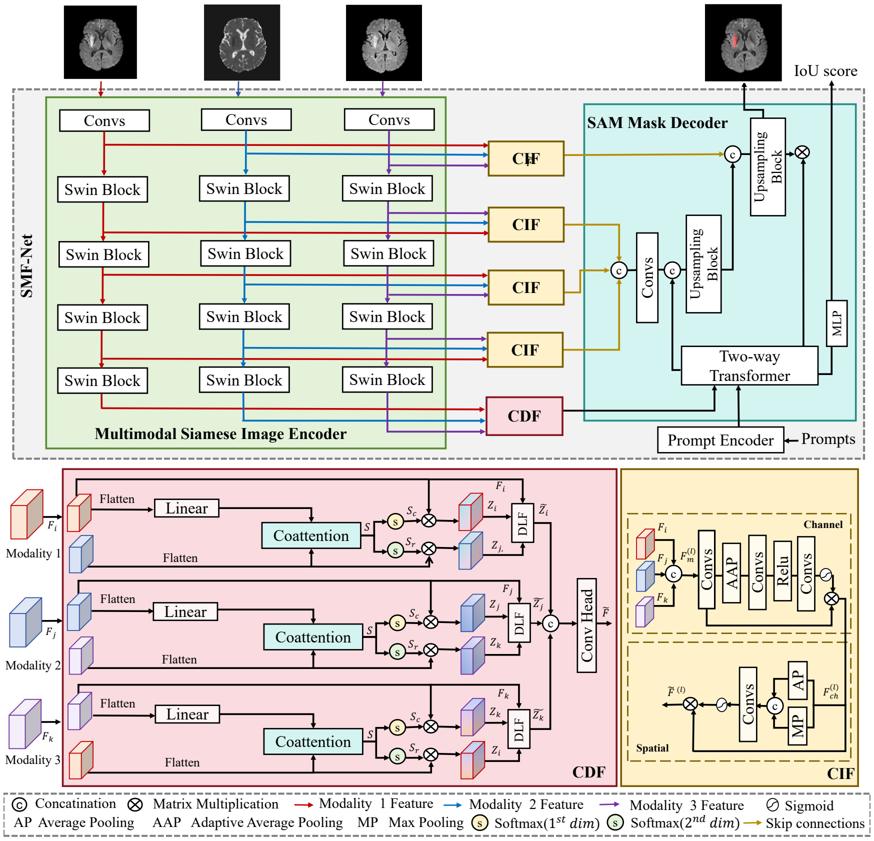

Fig. 1. An overview of the proposed SMF-Net and its components, CDF and CIF.

dencies and enhancing feature representation, and dynamic learnable fusion that combines features from all modalities, adapting based on learned weights to optimize integration. Finally, we introduce Intermediate-Layer Contextual Fusion (CIF), a lightweight, multi-layer attention-based fusion strategy that captures features at multiple scales while preserving modality-specific information. CIF integrates features through skip connections between the encoder and decoder, enhancing the integration of multiscale features. Extensive experiments on open benchmark dataset demonstrate that SMF-Net improves multimodal stroke lesion segmentation by effectively integrating modality-specific information, outperforming previous methods.

2 Methodology

In this section, we first provide a comprehensive overview of SMF-Net in subsection 2.1. We then present a detailed explanation of its core components in subsections 2.2 and 2.3, outlining their structure and functional contributions.

2.1 Network Architecture

The overview of SMF-Net is shown in Fig. 1. Given input image modalities $I_i \in \mathbb{R}^{h \times w \times c}$, where h, w and c denote the height, width, and channel dimensions of the image and i denote different modalities, the multimodal Siamese encoder, based on a modified Swin Transformer [4], extracts modality-specific features. Using shared weights across modalities, the encoder ensures consistent feature extraction while maintaining computational efficiency. For each layer l, the encoder generates modality-specific feature maps $F_i^{(l)} \in \mathbb{R}^{h_l \times w_l \times c_l}$, where h_l, w_l, and c_l denote the spatial height, width, and channel dimensions of the feature in that layer. The CDF module simultaneously models the cross-modal relationships between modalities, refines, and fuse feature representations extracted from the final layer of the encoder. First, the CDF module captures the interactions between modality-specific features through pairwise co-attention, facilitating the integration of complementary information while preserving modality-specific characteristics. Then, adaptive weighting dynamically refines and fuses representations, ensuring that the most relevant modality contributions are emphasized while suppressing less informative features. Finally, the resulting features are concatenated and passed through a convolution operation to produce a unified encoded feature representation $F_{\text{CDF}}^{(l)} \in \mathbb{R}^{h_l \times w_l \times c_l}$. To preserve multiscale information, the CIF module dynamically fuses features from each intermediate layer of the encoder. This strategy captures modality-specific information in multiple layers, producing a fused representation $F_{\text{CIF}}^{(l)} \in \mathbb{R}^{h_l \times w_l \times c_l}$ in each layer. These fused multilayer features are then used as skip connections, enabling better multiscale feature propagation between the encoder and decoder. Skip connections are formed by concatenating the deeper $F_{\text{CIF}}^{(l)}$ outputs for early upsampling, while the shallowest output is integrated later to refine spatial details. At the final scale, the $F_{\text{CDF}}^{(l)}$ module fuses features and serves as a deep skip connection by providing a global image embedding to the decoder. The final segmentation output is generated by processing these fused features (i.e., $F_{\text{CDF}}^{(l)}$ and $F_{\text{CIF}}^{(l)}$) through a SAM mask decoder [8]. In addition, a prompt mechanism is kept the same as [4] to further refine the segmentation results.

2.2 Complementary Dynamic Fusion Module

To effectively model modality relationships while preserving modality-specific characteristics, we use a multimodal Siamese image encoder with shared weights, ensuring consistent feature extraction. The final-layer features F_i that contain deeper features are processed and fused by the CDF module, which integrates complementary information among modalities. CDF consists of pairwise co-attention to model interactions and dynamic learnable fusion to adaptively combine co-attended features. Inspired by [10], we apply co-attention to model modality interactions, enabling the network to capture modality relationships and dynamically emphasize complementary features.

Pairwise co-attention. Given two modality-specific feature maps $F_i \in \mathbb{R}^{h \times w \times c}$ and $F_j \in \mathbb{R}^{h \times w \times c}$, extracted from the final layer of the encoder, the co-attention mechanism captures pairwise modality relationships. In particular, we first flatten the feature maps F_i and F_j. Then, F_i is projected using a learnable weight matrix W. The co-attention mechanism computes the attention matrix S by evaluating the pairwise interaction between F_i and F_j as:

$$S = F_i^\top W F_j, \tag{1}$$

where $W \in \mathbb{R}^{c\times c}$ is a learned weight matrix, and $F_i, F_j \in \mathbb{R}^{c\times hw}$ are the flattened feature maps of modalities i and j, respectively. We then compute the softmax along both the column and row dimensions of S, resulting in attention weights:

$$S_{\mathrm{c}} = \mathrm{softmax}(S, \mathrm{dim} = 1), \quad S_{\mathrm{r}} = \mathrm{softmax}(S, \mathrm{dim} = 2), \tag{2}$$

where S_{c} and S_{r} represent the column and row-wise attention weights, respectively. Using these attention maps, we generate co-attended feature representations by applying attention-weighted matrix multiplication:

$$Z_i = F_j S_{\mathrm{c}}, \quad Z_j = F_i S_{\mathrm{r}}^\top, \tag{3}$$

where Z_i and Z_j are the refined co-attended feature maps. Pairwise co-attention emphasizes informative regions across modalities, enhancing shared feature representation.

Dynamic Learnable Fusion. To integrate complementary information from multiple modalities, we introduce a dynamic learnable fusion mechanism that adaptively assigns spatial attention weights to the original and co-attended features via element-wise multiplication, producing the refined modality-specific feature $\tilde{Z}_i$ as:

$$S_\alpha, S_\lambda, S_\gamma = \mathrm{softmax}(\mathrm{Conv}(\mathrm{Concat}(F_i, Z_i, Z_j)), \tag{4}$$

$$\tilde{Z}_i = S_\alpha \circ F_i + S_\lambda \circ Z_i + S_\gamma \circ Z_j, \tag{5}$$

where Concat$(\cdot)$ represents a concatenation operation followed by a convolution operation Conv$(\cdot)$ with kernel size 1 and softmax function. The weights S_α , S_λ, and S_γ represent the learned importance of F_i, Z_i and Z_j, respectively. Once modality-specific features are refined, the final multimodal feature representation $\tilde{F}$ is calculated as:

$$\tilde{F} = \mathrm{Conv}(\mathrm{Concat}(\tilde{Z}_i, \tilde{Z}_j, \tilde{Z}_k)), \tag{6}$$

where $\tilde{Z}_i, \tilde{Z}_j$, and $\tilde{Z}_k$ denote the refined modality-specific feature corresponding to different imaging modalities.

2.3 Context-aware Intermediate-layer Fusion Module

We integrate multimodal features across layers using the CIF module, acknowledging the importance of multilayer fusion as highlighted in [23, 24]. Modality-specific features from the intermediate layer undergo lightweight fusion followed by attention-based refinement. Given intermediate-layer modality features $F^{(l)}$, we concatenate them along the channel dimension and apply a convolution block to learn a weighted combination $F_{\mathrm{m}}^{(l)}$. We further apply sequential channel and spatial attention, following [18], which employs lightweight attention for adaptive refinement. Channel attention assigns modality-wise weights and refines $F_{\mathrm{m}}^{(l)}$, yielding channel-refined feature map $F_{\mathrm{ch}}^{(l)}$ via element-wise multiplication, as:

$$F_{\mathrm{ch}}^{(l)} = \mathcal{F}_{\mathrm{ch}}(F_{\mathrm{m}}^{(l)}) \circ F_{\mathrm{m}}^{(l)}, \tag{7}$$

where $\mathcal{F}_{\mathrm{ch}}(\cdot)$ denotes channel attention. $\mathcal{F}_{\mathrm{ch}}(\cdot)$ comprises of adaptive average pooling, two convolution operations with kernel size 1 and ReLU activation in between, and a final sigmoid function. To capture spatially significant features, we apply spatial attention. $F_{\mathrm{ch}}^{(l)}$ undergoes average and max pooling, followed by concatenation and convolution operation with kernel size 7 yielding the spatial attention map. This map refines $F_{\mathrm{ch}}^{(l)}$, producing layerwise multimodal feature representation $\tilde{F}^{(l)}$ via element-wise multiplication, as:

$$\tilde{F}^{(l)} = \mathcal{F}_{\mathrm{sp}}(F_{\mathrm{ch}}^{(l)}) \circ F_{\mathrm{ch}}^{(l)}, \tag{8}$$

where $\mathcal{F}_{\mathrm{sp}}(\cdot)$ represents spatial attention refinement. After modality-specific fusion at each layer, the fused features propagate through skip connections, preserving high-resolution details while deeper layers retain contextual information.

3 Experiments

3.1 Dataset

We evaluate SMF-Net on the Ischemic Stroke Lesion Segmentation (ISLES) 2022 dataset [7]. This dataset consists of 250 multimodal and multicenter MRI modalities for the segmentation of acute to subacute stroke lesions. Each scan includes DWI, ADC, and FLAIR sequences.

3.2 Implementation Details

Proper alignment between modalities is crucial in multimodal learning, and since FLAIR images were misaligned with DWI and ADC, we applied image registration to ensure alignment. Voxel dimensions were standardized to $2 \times 2 \times 2\,\mathrm{mm}^3$, and images were resized to 256×256 pixels using cropping and padding. We followed the preprocessing pipeline in [12]. Then, we included a margin of three axial slices before and after the lesion to provide additional context and ensure a balanced distribution of training samples. DWI consistently yielded the highest segmentation performance and was selected as our baseline modality due to its clinical relevance in acute stroke detection. The dataset was split 80%-20% for training and testing. Training was conducted on an NVIDIA GeForce RTX 3090 (24GB RAM) with a batch size of 4 for 70 epochs. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4}, adjusted via the ReducedLROnPlateau scheduler. We use dice loss to optimize spatial overlap, focal loss to

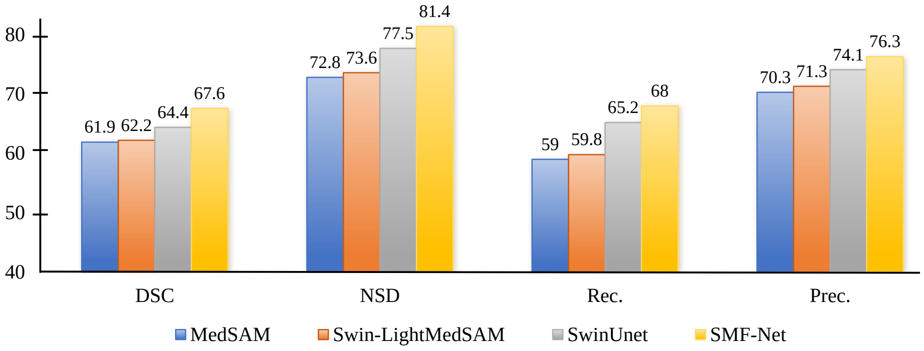

Fig. 2. Comparison of SMF-Net with previous medical image segmentation methods on the ISLES 2022 dataset.

Table 1. Ablation study results evaluating the effectiveness of different components in the SMF-Net and evaluating the contribution of different modalities.

Modality			Fusion			Evaluation Performance			
ADC	FLAIR	DWI	CDF	DLF	CIF	DSC	NSD	Rec.	Prec.
✓	✗	✗	✗	✗	✗	42.2	46.7	50.0	43.0
✗	✓	✗	✗	✗	✗	48.2	56.1	50.9	52.7
✗	✗	✓	✗	✗	✗	62.2	73.6	59.8	71.3
✓	✓	✓	✓	✗	✗	66.0	77.7	60.4	75.3
✓	✓	✓	✓	✗	✓	67.2	80.5	67.7	74.0
✓	✓	✓	✓	✓	✗	67.3	79.6	65.2	72.7
✓	✓	✓	✓	✓	✓	67.6	81.4	68.0	76.3

handle class imbalance by focusing on hard examples, and MSE loss to supervise the IoU prediction head. We followed the data augmentation strategy in [4], additionally we applied random affine transformations, including rotations (±90°), scaling (0.8–1.2), and translations (±10 pixels) with 50% probability. The network performance was evaluated using widely used metrics: Dice Score (DSC), Dice(NSD), Recall, and Precision.

3.3 Quantitative and Qualitative Results

To assess the performance of SMF-Net, we compared it with widely adopted segmentation approaches. These include Swin-Unet [1], MedSAM [12], and Swin-LightMedSAM [4]. The results are shown in Fig. 2. SMF-Net demonstrates superior performance among SAM-based methods, achieving the highest DSC (67.6%), outperforming both MedSAM (DSC: 61.9%) and Swin-LightMedSAM (DSC: 62.2%). Furthermore, SMF-Net achieves a significantly greater performance gain over SwinUnet. A paired t-test between SMF-Net and the baseline on DSC scores confirms the improvement is statistically significant ($p < 0.005$).

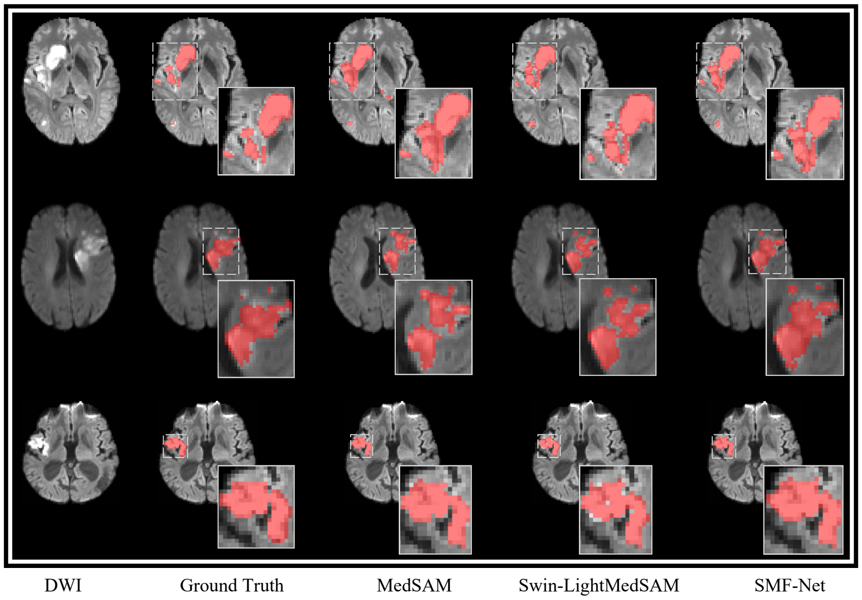

Fig. 3. Qualitative results of stroke segmentation on ISLES 2022 datasets, with the corresponding lesion area in red.

Furthermore, the qualitative assessment of the SMF-Net and SAM-based methods on ISLES 2022 is visually presented in Fig. 3. Swin-LightMedSAM shows improved performance over MedSAM but still leaves some gaps in the segmenting lesion regions. However, SMF-Net enhances lesion segmentation by reducing missing regions compared to other methods.

3.4 Ablation Study

we conducted an ablation study to evaluate the contribution of modalities and fusion strategies in SMF-Net. In Table 1, we summarize the ablation study results that assess the contribution of each imaging modality (ADC, FLAIR, DWI) and the fusion strategies (CDF, Dynamic Learnable Fusion (DLF), and CIF) in SMF-Net. Among the single-modality experiments, using DWI achieves the highest performance (DSC: 62.2%, NSD: 73.6%), likely because of its high sensitivity to ischemic regions. In contrast, using ADC produces the lowest scores (DSC: 42.2%, NSD: 46.7%) due to limited contrast, while using FLAIR falls in between (DSC: 48.2%, NSD: 56.1%), highlighting its complementary role in stroke imaging. Building on the DWI baseline, introducing the CDF module yields a +3.8% increase in DSC and a +4.1% increase in NSD, emphasizing the benefits of multimodal fusion based on cross-modal interactions. Adding the CIF module further boosts performance by enabling multiscale feature fusion, whereas the DLF refines feature integration but does not replace CIFs benefits. The SMF-

Net, integrating CDF, CIF, and DLF achieves the best overall performance in all evaluation metrics. These improvements underscore the importance of leveraging multimodal data and complementary fusion strategies for accurate stroke lesion segmentation.

4 Conclusion

In this work, we proposed SMF-Net, a multimodal stroke lesion segmentation framework that effectively integrates complementary information from multiple modalities. The model utilizes a multimodal Siamese image encoder for modality-specific feature extraction while maintaining computational efficiency. We introduced CDF with pairwise co-attention to enhance cross-modal interactions, dynamic learnable fusion for adaptive weighted fusion, and CIF to capture multi-scale information. Each component of SMF-Net was evaluated, demonstrating its contribution to the overall performance. Experimental results show that SMF-Net outperforms existing methods, emphasizing the importance of structured multimodal feature integration in stroke lesion segmentation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62201465 and 62271405.

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

References

1. 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)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Feigin, V.L., Norrving, B., Mensah, G.A.: Global burden of stroke. Circulation research **120**(3), 439–448 (2017)
4. Gao, R., Lyu, D., Staring, M.: Swin-LiteMedSAM: A lightweight box-based segment anything model for large-scale medical image datasets. In: Medical Image Segmentation Challenge, pp. 70–82. Springer (2024)
5. Garcia-Salgado, B.P., Almaraz-Damian, J.A., Cervantes-Chavarria, O., Ponomaryov, V., Reyes-Reyes, R., Cruz-Ramos, C., Sadovnychiy, S.: Enhanced ischemic stroke lesion segmentation in MRI using attention U-Net with generalized Dice focal loss. Applied Sciences **14**(18), 8183 (2024)
6. Gheibi, Y., Shirini, K., Razavi, S.N., Farhoudi, M., Samad-Soltani, T.: CNN-Res: deep learning framework for segmentation of acute ischemic stroke lesions on multimodal MRI images. BMC Medical Informatics and Decision Making **23**(1), 192 (2023)

7. 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., et al.: ISLES 2022: A multicenter magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
9. Liu, N., Li, H.: CMU-Net: A Cascaded Mini U-Network for Retinal Vessel Segmentation. SSRN Electronic Journal (2022)
10. Lu, X., Wang, W., Shen, J., Crandall, D., Luo, J.: Zero-shot video object segmentation with co-attention siamese networks. IEEE transactions on pattern analysis and machine intelligence **44**(4), 2228–2242 (2020)
11. Luo, J., Dai, P., He, Z., Huang, Z., Liao, S., Liu, K.: Deep learning models for ischemic stroke lesion segmentation in medical images: a survey. Computers in biology and medicine p. 108509 (2024)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. 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)
14. Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., Wang, S.: X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 247–255. Springer (2019)
15. Raju, C.S.P., Neelapu, B.C., Laskar, R.H., Muhammad, G.: Analysis of multimodal fusion strategies in deep learning for ischemic stroke lesion segmentation on computed tomography perfusion data. Multimedia Tools and Applications pp. 1–26 (2024)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
17. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24. pp. 36–46. Springer (2021)
18. 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)
19. Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., Li, J.: W-Net: A boundary-enhanced segmentation network for stroke lesions. Expert Systems with Applications **230**, 120637 (2023)
20. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16111–16121 (2024)
21. Yu, W., Huang, Z., Zhang, J., Shan, H.: SAN-Net: Learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization. Computers in Biology and Medicine **156**, 106717 (2023)

22. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
23. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE transactions on medical imaging **39**(9), 2772–2781 (2020)
24. Zhou, T., Zhou, Y., Li, G., Chen, G., Shen, J.: Uncertainty-aware hierarchical aggregation network for medical image segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2024)
25. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-UNet: a dimension-fusion U shape network for chronic stroke lesion segmentation. IEEE/ACM transactions on computational biology and bioinformatics **18**(3), 940–950 (2019)