Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/145/2025.21898

SWD3-Net: A Spatiotemporal Attention-Based Network for Cerebrovascular Segmentation

Wenxin Hua[1,†], Guang Li[1,a,*], Guangxiang Si[1], Yan Li[1,b,†], Ning Gu[2]

[1]State Key Laboratory of Digital Medical Engineering, Jiangsu Key Laboratory for Biomaterials and Devices, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210009, China
[2]Nanjing Key Laboratory for Cardiovascular Information and Health Engineering Medicine, Institute of Clinical Medicine, Nanjing Drum Tower Hospital, Medical School, Nanjing University, Nanjing, 210093, China
a. 101012537@seu.edu.cn, b. liyan@seu.edu.cn
**corresponding authors*
[†]These authors contributed equally to this paper

Abstract: Cerebrovascular segmentation in Time-of-Flight magnetic resonance angiography (TOF-MRA) is a fundamental task for assisting diagnosis, aiding in the detection and analysis of vascular structures to assist clinicians in medical decision-making. Due to the complexity of vascular structures and low background contrast, automated segmentation of brain vessels remains highly challenging. Traditional U-Net architectures struggle to effectively capture long-range and multi-scale vascular features, which limits their ability to distinguish fine structural details across slices and may lead to voxel misclassification due to convolution's local receptive field. While CNN-Transformer hybrid models have shown promising results, we propose SWD3-Net, a novel deep learning framework incorporating a Spatiotemporal Feature Extraction module that leverages dilated convolutions to capture multi-scale contextual information across depth layers, enhancing inter-slice feature representation. In addition, we introduce a weighted binary cross-entropy loss function to address severe class imbalance. Results show SWD3-Net surpasses advanced models and the state-of-the-art deep learning medical image segmentation benchmark nnU-Net, achieving a 2% higher DSC. Our method effectively preserves vascular structures and maintains topological consistency.

Keywords: Cerebrovascular Segmentation, Spatiotemporal Feature Representation, Deep Learning, TOF-MRA

1. Introduction

Cerebrovascular disease significantly contributes to global mortality and disability. Time-of-flight magnetic resonance angiography (TOF-MRA) is a key non-invasive imaging technique for diagnosing these diseases[1]. However, noise and slow blood flow degrade image quality, complicating interpretation. Automated TOF-MRA segmentation is crucial for clinical diagnosis but faces challenges due to significant morphological variations. Current segmentation largely relies on manual annotation, which is time-consuming and inconsistent[2]. While deep learning has advanced organ segmentation, its performance in vascular segmentation is limited due to complex vessel

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/145/2025.21898

morphology and class imbalance. Most existing methods, based on U-Net[3] and V-Net[4], focus on local feature extraction but struggle to capture long-range dependencies making them less suitable for vascular segmentation. Therefore, an approach that integrates long-range and multi-scale dependencies and spatiotemporal feature fusion is needed.

2. Related work

Deep learning has advanced cerebrovascular segmentation, with most methods still relying on CNN-based architectures. Sanchesa et al.[5] enhanced U-Net by replacing convolutions with Inception modules[6] to improve feature extraction and prevent overfitting. Hilbert et al.[7] proposed a multi-scale model with a context path, designed to capture both high- and low-resolution volumes. Lee et al.[8] explored feature relationships across slices by stacking U-Nets to extract spatial features, using LSTM connections between the encoder and decoder.

However, CNN-based methods struggle to capture long-range dependencies. In contrast, the self-attention mechanism in Transformers, such as Vision Transformer (ViT)[9] and Swin Transformer[10], effectively captures both the dependencies and multi-scale features. Swin-Unet[11] integrates hierarchical Transformer blocks in both the encoder and decoder. Chen et al.[12] applied a Transformer model to cerebrovascular segmentation, where the encoder uses consecutive Transformer modules to extract features and passes them to a CNN-based decoder.

In 2.5D and 3D medical imaging, regions of interest often span multiple slices, and adjacent slices are visually similar and form connectivity along the z-axis. In this work, we propose SWD3-Net, an improved Swin-Unet model that enhances inter-slice connectivity by integrating the temporal attention mechanism into the Swin-Transformer block. Due to class imbalance in vascular imaging, where non-vascular regions dominate during training, we propose a class-imbalanced loss function. The results have illustrated the effectiveness and interpretability of our approach. Our work has two main contributions: (1) we design a spatiotemporal feature extraction (STFE) module and integrate it into the Swin-Transformer block, which simultaneously captures vascular continuity across consecutive slices and long-range spatial features, enhancing cerebrovascular segmentation, and (2) we introduce a class-imbalanced loss function to address the prediction bias, reducing the dominance of non-vascular regions during training.

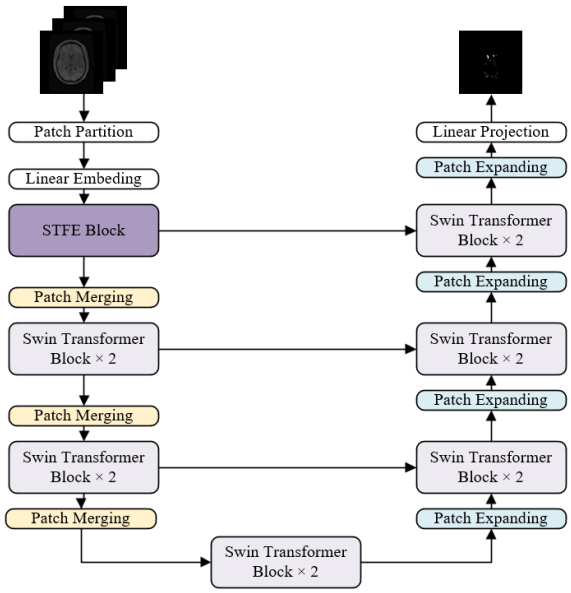

Figure 1: Overall architecture of the proposed SWD3-Net

DOI: 10.54254/2755-2721/145/2025.21898

3. Methodology

3.1. Overview of the architecture

Figure 1 illustrates our proposed SWD3-Net and its components. SWD3-Net builds on Swin-Unet, but the standard Swin-Transformer Block only captures spatial information within the same slice. For vessel segmentation, which involves strong inter-slice connectivity, learning relationships between consecutive slices is crucial. To address this, we design a temporal attention branch to capture inter-slice features using dilated convolutions with varying dilation rates across consecutive slices. This enables a more comprehensive representation of vascular characteristics, improving connectivity and segmentation. Since inter-frame features show small changes overall, we introduce the temporal attention branch only in Stage 1 to balance feature extraction and computational efficiency. Later stages use single-frame features, preventing overfitting and excessive parameter increase.

3.2. Spatiotemporal Feature Extraction Module

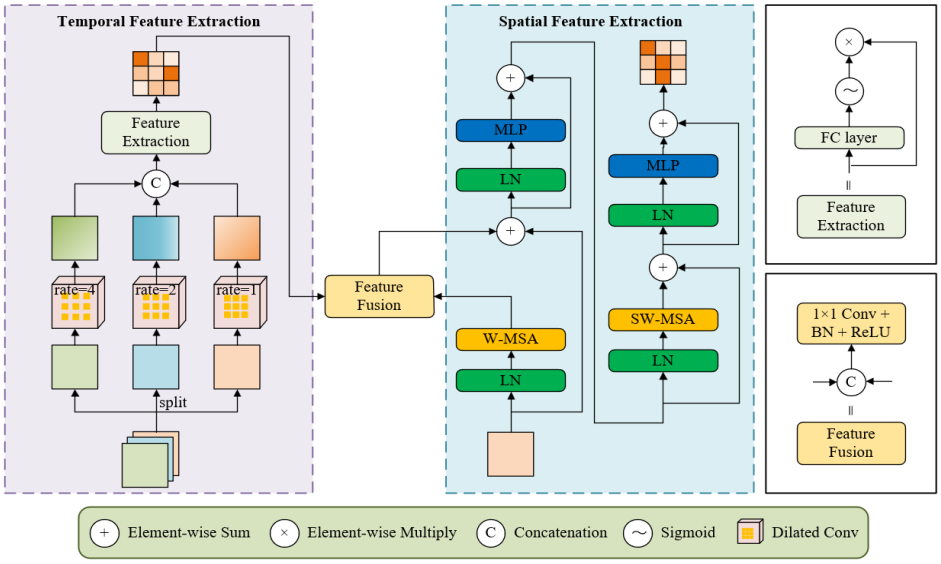

Figure 2: Structure of the STFE module, comprising the temporal and spatial feature extraction components. 1, 2, and 4 represent three different dilation rates.

Inspired by network architectures and vascular characteristics like sparse distribution and continuous extension, we designed a spatiotemporal feature extraction module that leverages W-MSA for spatial dependencies and dilated convolutions for temporal dependencies. The temporal branch runs parallel to the spatial branch, allowing each to capture distinct features.

As illustrated in Figure 2, the input features $X \in \mathbb{R}^{C\times H\times W\times D}$ are first divided into three parallel feature groups along the depth dimension, labeled as f_i ($i \in \{1, 2, 3\}$), each with the same shape as X. Dilated convolutions with varing dilation rates ri are applied to capture multi-scale temporal features, then the outputs are concatenated along the depth dimension:

$$f_i' = C_{ri}(f_i) \tag{1}$$

$$X' = \mathbb{C}\left(f_1^{'}, f_2^{'}, f_3^{'}\right) \tag{2}$$

DOI: 10.54254/2755-2721/145/2025.21898

This operation enhances temporal perception while preserving original feature information. The fused feature map $X' \in \mathbb{R}^{C\times H\times W\times D}$ is then processed by a fully connected (FC) layer for feature attention:

$$X'_{att} = X' \otimes \sigma\left(W_{FC} \cdot X' + b_{FC}\right) \tag{3}$$

where W_{FC} is the weight matrix, b_{FC} is the bias term, and σ denotes a nonlinear activation function. The attention map indicates the importance of temporal features at each position and is applied element-wise to X'. Finally, the first frame of X'_{att} is fused with the spatial feature X_{MSA} from W-MSA to produce the final representation:

$$X_{fuse} = ReLU\left(BN\left(Conv(\mathbb{C}(X'_{att}[:,:,:,0], X_{MSA}])\right)\right) \tag{4}$$

3.3. Weighted Loss Function

Cerebrovascular structures occupy only a minor portion of the brain, creating a significant class imbalance between the foreground and background regions. This imbalance can lead to prediction bias in the network. The traditional Binary Cross-Entropy Loss (BCE Loss) is calculated as:

$$BCE(p,y) = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i\, log(p_i) + \, (1-y_i)\, \text{alog}(1-p_i)] \tag{5}$$

where p_i is the predicted foreground probability, y_i is the target label (1 for the foreground, 0 for the background), and N is the total sample count. BCE Loss optimizes pixel-wise predictions but struggles with vascular segmentation due to class imbalance—background voxels dominate, causing vascular regions to be overlooked. To address this issue, we modify the loss function as:

$$WeightedBCE(p,y) = -\frac{1}{N}\sum_{i=1}^{N}[\, \omega_1 y_i\, log(p_i) + \, \omega_2(1-y_i)\, log(1-p_i)] \tag{6}$$

where ω_1 and ω_2 are the weights for the foreground and background classes, respectively.

4. Experiments and results

4.1. Dataset and Implementation Details

We used the public dataset of healthy volunteers (https://public.kitware.com/Wiki/TubeTK/Data) annotated by Hilbert et al. [5] for cerebrovascular arteries, consisting of 20 cases. The images have a resolution of 448x448x128 with 0.51x0.51x0.8 mm^3. To manage the size of the images and hardware limitations, the input was divided into patches of size 448x448x3. The data were saved in. npy format, ordered by slice images, label images, and slice labels, to enable 3D image reconstruction.

The experiments were conducted using the Pytorch framework, with training performed on one NVIDIA RTX 4090D GPU. The network underwent 250 epochs with a batch size of 2. During training, dropout rate was set to 0.1 and the initial learning rate was set to 0.001. The results were evaluated using three metrics: precision (PRE), recall (REC), and Dice Similarity Coefficient (DSC).

4.2. Experiments

In this subsection, we evaluate the segmentation performance of our spatiotemporal attention model against several recent methods on the aforementioned dataset.

Table 1 shows that our method achieves the highest DSC and recall, improving by 2% and 5%, over nnU-Net v2. However, it slightly lowers precision due to a focus on foreground detection, increasing false positives, thereby reducing precision. Despite this, SWD3-Net balances recall and

DOI: 10.54254/2755-2721/145/2025.21898

precision effectively, achieving the best DSC, validating its effectiveness in intracranial artery segmentation. Fig. 3 presents the 3D visualization results.

Table 1: Quantitative comparison of our method and other networks. The best results are in red, while the second-best are in green. Our method notably outperforms nnU-Net in terms of DSC and REC.

Model	DSC (%)	PRE (%)	REC (%)
3D U-Net[13]	73.19	78.03	68.98
Attention U-Net[14]	74.73	66.28	86.59
V-Net[15]	78.64	78.34	79.65
nnU-Net v2[16]	87.88	91.35	84.70
SWD3-Net (ours)	89.82	89.50	90.44

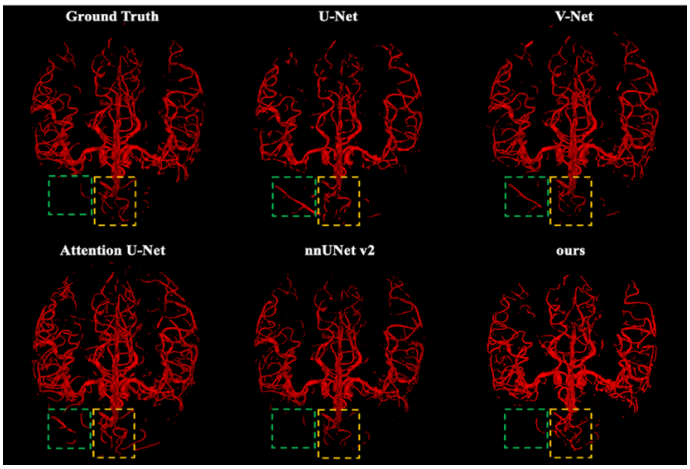

Figure 3: 3D visualization shows that general networks predict excess non-arterial vascular regions with numerous disconnections and noise. Our approach presents a more refined vascular tree.

5. Ablation studies

5.1. Dilated Convolution Rate Comparison

Table 2 shows the results of ablation study on the STFE module. Using Swin-Unet with traditional BCE loss as the baseline, the baseline+D1 model improved DSC by 7%, highlighting the significance of temporal feature extraction and the baseline+D3 model achieved the best performance across all evaluation metrics.

Table 2: Ablation study result on the STFE module. D1–D4 represent the different dilated convolution rates in the parallel branches, where D1 = [1, 1, 1], D2 = [1, 2, 3], D3 = [1, 2, 4], and D4 = [1, 3, 5].

Model	DSC (%)	PRE (%)	REC (%)
Baseline	79.97	85.79	74.56
Baseline+D1	87.01	87.69	86.83
Baseline+D2	88.13	88.73	87.79
Baseline+D3	88.19	89.23	87.82
Baseline+D4	88.08	88.93	87.74

DOI: 10.54254/2755-2721/145/2025.21898

5.2. Loss Function Comparison

Table 3 presents the optimal weight ratio (ω_1: ω_2) for our proposed loss function, indicating that the model's performance peaked when the weight ratio was set to 1:8.

We also conducted comparative experiments with other loss functions. As shown in Table 4, wBL improved DSC by 1.6% compared to the baseline. While it emphasizes foreground regions, it also increases false positives, limiting precision gains. However, wBL strikes a better balance between precision and recall, since it improves recall while maintains good precision.

These results show that wBL effectively reduces background class dominance and enhances segmentation performance.

Table 3: Ablation analysis of the loss parameter settings.

Method (ω_1: ω_2)	DSC (%)	PRE (%)	REC (%)
1:1	88.19	89.23	87.82
1:6	88.93	89.53	89.16
1:8	89.82	89.50	90.44
1:10	89.24	89.07	89.79
1:12	88.55	89.32	88.23

Table 4: Comparison of several different loss functions combined with BCE loss (BL) at a 4:6 ratio. Supervised by the Dice loss (DL), weighted Dice loss (Huang et al.[17]) (wDL), inverse distance-weighted Dice loss (Wang et al.[18]) (IDDL) and our WeightedBCE loss (wBL), respectively.

Method	DSC (%)	PRE (%)	REC (%)
BL	88.19	89.23	87.82
BL+DL	87.61	88.83	87.12
BL+wDL	83.36	83.44	84.09
BL+IDDL	86.11	84.35	87.87
wBL	89.82	89.50	90.44

6. Discussion

This study examines the impact of different dilation rates on the performance by evaluating four sets of dilation rates, as shown in Table 2. Due to image size and hardware limitations, the input image block size is fixed at 3. While D3 performed best in the overall results, the difference is minor. Studying a more lightweight network could handle larger image patches and parallel branches, potentially improving performance. In order to investigate the impact of hyperparameters in the loss function, we perform experiments with five parameter sets. Table 3 shows that the 1:8 ratio achieve the best DSC and recall but rank second in precision. Thus, perhaps by adaptively learning these hyperparameters may optimize model performance while ensuring adaptability across different datasets. Therefore, future work will focus on (1) collecting more data to enhance model robustness, (2) developing an adaptive strategy for hyperparameter tuning in the loss function, and (3) attempting to develop a lightweight network with stronger feature extraction ability.

7. Conclusion

We propose a novel cerebrovascular segmentation network for TOF-MRA images, incorporating a spatiotemporal attention mechanism to enhance feature extraction. The STFE module captures temporal dependencies across layers, improving contextual representation. In addition, we introduce a weighted BCE loss to mitigate class imbalance by balancing the focus between foreground and

DOI: 10.54254/2755-2721/145/2025.21898

background regions. The experimental results show that our model surpasses multiple peer methods in performance, highlighting its effectiveness in cerebrovascular segmentation. These findings suggest that our approach has promising potential for broader applications in cerebrovascular imaging and medical image analysis.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62394312), the Natural Science Foundation of Jiangsu Province (BK20222002), and the Nanjing Science and Technology Development Foundation (202205066).

References

[1] *Chen, Z., Xie, L., Chen, Y., Zeng, Q., ZhuGe, Q., Shen, J., Wen, C., & Feng, Y. (2022). Generative adversarial network based cerebrovascular segmentation for time-of-flight magnetic resonance angiography image. Neurocomputing (Amsterdam), 488, 657–668.*

[2] *Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F., & Kong, D. (2017). Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. International Journal for Computer Assisted Radiology and Surgery, 12(3), 399–411.*

[3] *Ronneberger, O., Fischer, P., Brox, T., Wells, W. M., Frangi, A. F., Navab, N., & Hornegger, J. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015 (Vol. 9351, pp. 234–241). Springer International Publishing AG.*

[4] *Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 565–571.*

[5] *Sanchesa, P., Meyer, C., Vigon, V., & Naegel, B. (2019). Cerebrovascular Network Segmentation of MRA Images With Deep Learning. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 768–771.*

[6] *Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition. Proceedings, 1–9.*

[7] *Hilbert, A., Madai, V. I., Akay, E. M., Aydin, O. U., Behland, J., Sobesky, J., Galinovic, I., Khalil, A. A., Taha, A. A., Wuerfel, J., Dusek, P., Niendorf, T., Fiebach, J. B., Frey, D., & Livne, M. (2020). BRAVE-NET: Fully Automated Arterial Brain Vessel Segmentation in Patients With Cerebrovascular Disease. Frontiers in Artificial Intelligence, 3, 552258–552258.*

[8] *Lee, K., Sunwoo, L., Kim, T., & Lee, K. J. (2021). Spider U-Net: Incorporating Inter-Slice Connectivity Using LSTM for 3D Blood Vessel Segmentation. Applied Sciences, 11(5), 2014-.*

[9] *Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv.Org.*

[10] *Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002.*

[11] *Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., Michaeli, T., Karlinsky, L., & Nishino, K. (2023). Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In Computer Vision - ECCV 2022 Workshops (Vol. 13803, pp. 205–218). Springer.*

[12] *Chen, C., Zhou, K., Wang, Z., & Xiao, R. (2023). Generative Consistency for Semi-Supervised Cerebrovascular Segmentation From TOF-MRA. IEEE Transactions on Medical Imaging, 42(2), 346–353.*

[13] *Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., Joskowicz, L., Unal, G., Sabuncu, M. R., Ourselin, S., & Wells, W. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (pp. 424–432). Springer International Publishing.*

[14] *Oktay, O., Schlemper, J., Loic Le Folgoc, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning Where to Look for the Pancreas. arXiv.Org.*

[15] *Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 565–571.*

[16] *Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.*

DOI: 10.54254/2755-2721/145/2025.21898

[17] *Huang, Q., Sun, J., Ding, H., Wang, X., & Wang, G. (2018). Robust liver vessel extraction using 3D U-Net with variant dice loss function. Computers in Biology and Medicine, 101, 153–162.*

[18] *Wang, W., Xia, Q., Yan, Z., Hu, Z., Chen, Y., Zheng, W., Wang, X., Nie, S., Metaxas, D., & Zhang, S. (2024). AVDNet: Joint coronary artery and vein segmentation with topological consistency. Medical Image Analysis, 91, 102999–102999.*