OPEN ACCESS
Journal of Innovative Optical Health Sciences
Vol. 16, No. 4 (2023) 2350009 (14 pages)
© The Author(s)
DOI: 10.1142/S1793545823500098

World Scientific
www.worldscientific.com

Cerebrovascular segmentation from mesoscopic optical images using Swin Transformer

Yuxin Li*, Qianlong Zhang*, Hang Zhou†, Junhuai Li*, Xiangning Li‡,§
and Anan Li‡,§,¶

*Shaanxi Key Laboratory for Network Computing and Security Technology
School of Computer Science and Engineering, Xi'an University of Technology
Xi'an 710048, P. R. China

†School of Computer Science, Chengdu University of Information Technology
Chengdu 610225, P. R. China

‡Britton Chance Center for Biomedical Photonics
Wuhan National Laboratory for Optoelectronics MoE Key
Laboratory for Biomedical Photonics
Huazhong University of Science and Technology
Wuhan 430074, P. R. China

§HUST-Suzhou Institute for Brainsmatics
Suzhou 215123, P. R. China
¶aali@hust.edu.cn

Received 11 January 2023
Accepted 30 March 2023
Published 11 May 2023

Vascular segmentation is a crucial task in biomedical image processing, which is significant for analyzing and modeling vascular networks under physiological and pathological states. With advances in fluorescent labeling and mesoscopic optical techniques, it has become possible to map the whole-mouse-brain vascular networks at capillary resolution. However, segmenting vessels from mesoscopic optical images is a challenging task. The problems, such as vascular signal discontinuities, vessel lumens, and background fluorescence signals in mesoscopic optical images, belong to global semantic information during vascular segmentation. Traditional vascular segmentation methods based on convolutional neural networks (CNNs) have been limited by their insufficient receptive fields, making it challenging to capture global semantic information of vessels and resulting in inaccurate segmentation results. Here, we propose SegVesseler, a vascular segmentation method based on Swin Transformer. SegVesseler adopts 3D Swin Transformer blocks to extract global contextual information in 3D images. This approach is able to maintain the connectivity and topology of blood vessels during segmentation. We evaluated the performance of our method on mouse cerebrovascular datasets generated from three different labeling and imaging modalities. The experimental results demonstrate that the segmentation

[¶] Corresponding author.

effect of our method is significantly better than traditional CNNs and achieves state-of-the-art performance.

Keywords: Vascular segmentation; Swin Transformer; mesoscopic optical imaging; fMOST.

1. Introduction

Brain health relies on a continuous supply of oxygen and energy, delivered via blood supplied by the cerebral vascular system. Abnormalities in the cerebral vascular network are often associated with neurological diseases.[1,2] Thus, accurately mapping the whole-brain vascular networks is critical for elucidating the pathogenesis of many diseases. In recent years, with the development of labeling and imaging techniques, mesoscopic optical imaging could acquire whole-mouse-brain vascular networks at capillary resolution.[3] Vascular segmentation, a fundamental step in vascular image processing, is a prerequisite for accurately analyzing vascular networks. However, segmenting mesoscopic optical whole-brain vascular images is challenging (see Fig. 1). First, perfusion-based labeling methods may result in incomplete vessel filling, leading to the absence of fluorescent signals in some vessels. This, in turn, makes it difficult to maintain vascular network connectivity in the image. Second, some labeling strategies only label vascular endothelial cells, resulting in lumens in the center of blood vessels during imaging. Furthermore, the signal-to-noise ratio can vary across brain regions, and non-vascular fluorescent signals may also influence segmentation.

Existing vascular segmentation algorithms are broadly classified into traditional vascular segmentation algorithms and deep learning-based methods. Traditional vascular segmentation algorithms include threshold-based methods, tracking-based methods, filter-based methods, active contour methods, etc.[4–6] For example, Shang *et al.*[7] and Cheng *et al.*[8] utilized an active contour model to segment vascular tree structures. Wang *et al.*[9]

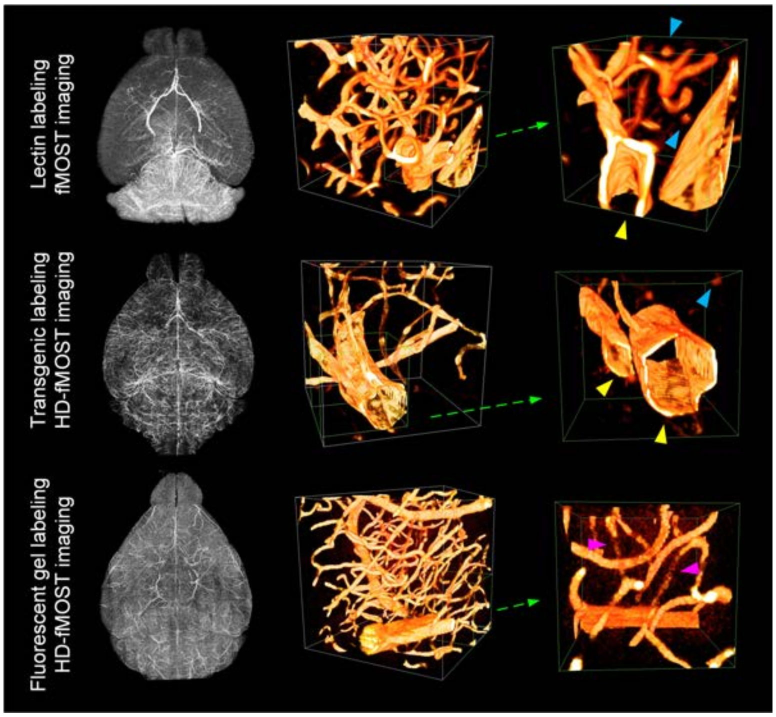


Fig. 1. Whole-mouse-brain vascular images from three different labeling and imaging modalities. Three representative volumes are randomly selected to illustrate the differences in vascular characteristics. Blue arrows indicate fluorescence interferences (nonvascular fluorescent signals) surrounding the blood vessels, yellow arrows highlight the lumens of thick blood vessels, and pink arrows mark vascular discontinuities at weak fluorescent signal regions.

proposed a vascular segmentation method based on the sequential Monte Carlo tracking algorithm, which expressed the vascular segmentation on CT angiography images as a Bayesian tracking problem. Frangi *et al.*[10] adapted a Hessian matrix-based edge detection filter to detect tubular-like structures. Zhao *et al.*[11] constructed a weighted symmetric filter for segmenting 3D vascular images from different imaging modalities. Traditional vascular segmentation algorithms require designers with prior knowledge to manually extract image features, and these methods can only be used for specific imaging modalities. In contrast, deep learning approaches, especially convolutional neural networks (CNNs), could automatically extract image features and have been widely applied in biomedical image processing. Several CNN-based approaches have been developed for segmenting cerebrovascular images. For example, Wang *et al.*[12] proposed a cascaded CNN architecture for segmenting cerebral vasculature in confocal microscopic images. Tahir *et al.*[13] and Damseh *et al.*[14] utilized CNN-based methods for the segmentation of mouse cerebrovascular images generated by two-photon microscopy, while Haft-Javaherian *et al.*[15] segmented similar images generated by multi-photon microscopy. These CNN-based methods have achieved state-of-the-art results during vascular segmentation. However, insufficient receptive fields lead to some limitations when extracting deep semantic information. These limitations prevent the CNN-based approach from overcoming the defects in mesoscopic optical whole-brain vascular images and obtaining accurate segmentation results.

In recent years, the Transformer architecture has garnered attention for its impressive performance in natural language processing (NLP) and computer vision (CV).[16] Compared to CNN-based methods, the Transformer has a greater capacity to capture global semantic information when processing images, making it increasingly popular for medical image segmentation tasks.[17–20] Specifically, the Swin Transformer[21] has emerged as a promising approach for various CV tasks, serving as a backbone network for extracting image features. This method has achieved conspicuous achievements in image classification, object detection, and image segmentation.[22–24] Therefore, we hope to utilize the Swin Transformer to overcome the defects of existing CNN methods during the segmentation of mesoscopic optical vascular images.

This paper proposes SegVesseler, a Swin Transformer-based method to segment mesoscopic optical vascular images. SegVesseler utilizes the Swin Transformer block as a feature extractor, which enables the extraction of global contextual information with larger receptive fields and maintains the connectivity and topology of blood vessels during segmentation. Additionally, a four-layer U-shaped framework consisting of encoders and decoders is designed to capture and integrate vascular features. During the encoding process, a multi-resolution feature extraction structure is employed to extract vascular features at various scales. In the decoding process, the extracted vascular features are fused and restored to the original spatial resolution to achieve segmentation. We validated our method on three whole-mouse-brain vascular datasets generated by different labeling and imaging modalities. Experimental results showed that our approach has a better segmentation effect than CNN-based methods and achieved state-of-the-art.

2. Materials and Methods

2.1. *Whole-brain imaging and datasets*

In this study, we utilized three whole-mouse-brain vascular datasets obtained through different labeling and imaging modalities, namely the Lectin-Dylight594 labeled whole-mouse-brain vasculature combined with fluorescence Micro-Optical Sectioning Tomography (fMOST)[25] imaging (referred to as lectin dataset), the Tek::cre-Ai47 transgenic mouse labeled dataset combined with high-definition fMOST (HD-fMOST)[26] imaging (referred to as the tek dataset), and the Lectin-Dylight488 + gel-BSA-FITC labeled vasculature combined with HD-fMOST imaging (referred to as the gel dataset). All animal experiments were approved by the Institutional Animal Ethics Committee of Huazhong University of Science and Technology.

The lectin dataset is the cerebral vasculature of an adult C57BL/6 mouse. First, Lectin-DyLight594 was intravenously injected into the tail of the mouse to label blood vessels. Subsequently, the brain was extracted and continuously imaged using fMOST at a resolution of $0.35 \times 0.35 \times 2\,\mu$m.

The tek dataset is the cerebral vasculature of a hybrid mouse produced by crossing a Tek-Cre transgenic mouse with an Ai47 transgenic mouse. In this mouse, the endothelial cells of the vascular wall were labeled with green-fluorescent proteins (GFP). The brain was imaged using HD-fMOST at a resolution of $0.32 \times 0.32 \times 1\,\mu$m.

In the gel dataset, we labeled the cerebral vasculature of a C57BL/6 mouse by perfusing Lectin-Dylight488 and fluorescent gel (gel-BSA-FITC) simultaneously. The murine brain was imaged using HD-fMOST at the resolution of $0.32 \times 0.32 \times 1\,\mu$m.

To prepare the data for training and testing our method, we resampled the resolution of the three vascular datasets to $1 \times 1 \times 1\,\mu$m/pixel. Subsequently, we cropped 82 volumes of size $200 \times 200 \times 200$ pixels from the lectin dataset, 100 volumes of size $160 \times 160 \times 160$ pixels from the tek dataset, and 100 volumes of size $160 \times 160 \times 160$ pixels from the gel dataset. To construct the ground truth, we manually annotated the blood vessels in all volumes using the segmentation editor module of Amira. These original and annotated volumes were used to train and test our method.

2.2. *Proposed method*

2.2.1. *Network architecture*

The architecture of our method is illustrated in Fig. 2, which adopts a four-layer U-shaped feature extraction architecture to perform the segmentation of mesoscopic optical vascular images. This architecture comprises two main sections: the encoder and the decoder. In the encoder section, we utilize 3D Swin Transformer blocks to capture global features with larger receptive fields, which helps to model long-term dependencies and maintain the connectivity and topology of vascular networks during segmentation. The patch merging blocks are placed between two feature extractors to perform downsampling operations, which decrease the spatial resolution of input features while increasing the number of channels. This multi-resolution hierarchy could extract the feature information at different scales. In the decoder section, multiple patch-expanding blocks and skip connections are employed. The patch-expanding blocks perform upsampling operations to recover the spatial resolution of features. The skip connections combine the multi-scale feature information extracted by encoder and decoder sections to supplement fine-grained details and improve the segmentation accuracy. At the end of the decoder section, a patch-expanding final block is used to compress the channel dimension to one, followed by the normalization of pixel values between zero and one using the Sigmoid function.

2.2.2. *Patch embedding block*

The patch embedding block adopts a 3D convolutional kernel with a size of $2 \times 2 \times 2$ and a step of 2 to downsample the input volume and reshape it into

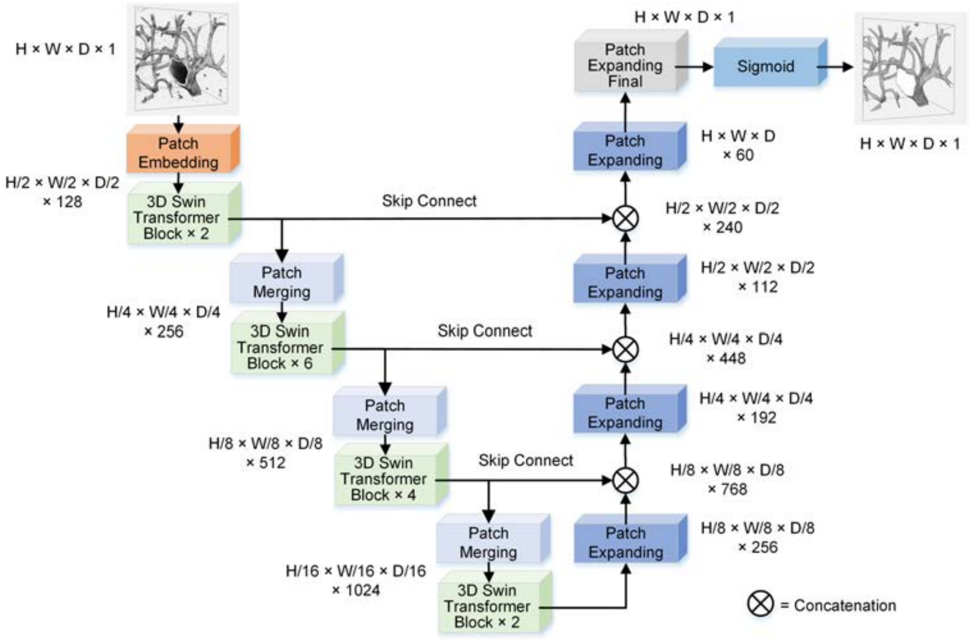


Fig. 2. The architecture of SegVesseler.

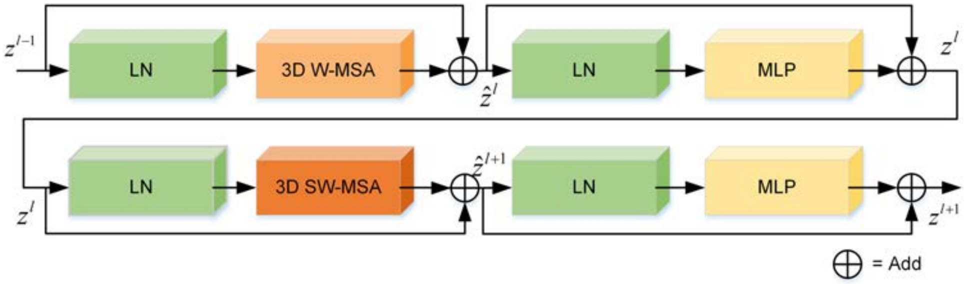


Fig. 3. 3D Swin Transformer block architecture.

a nonoverlapping patch (the size of 2 × 2 × 2 pixels) sequence. The spatial resolution of input volumes is reduced to half of their original size, while the channel dimension is mapped to 128. The number of patches is the length of the input sequence for Swin Transformer blocks.

2.2.3. *3D Swin transformer block*

The 3D Swin Transformer block[21] is used to extract global semantic information, and its architecture is shown in Fig. 3. Each block is composed of a window-based 3D multi-head self-attention (3D W-MSA) module, a shifted window-based 3D multi-head self-attention (3D SW-MSA) module, multi-layer perceptrons (MLP) layers, LayerNorm (LN) layers, and residual modules. The 3D W-MSA block calculates self-attention within each window, while the 3D SW-MSA block facilitates information interaction between different windows. Moreover, the MLP layers are used to fuse feature information and promote model convergence by stacking LN layers. The introduction of residual modules could facilitate the backpropagation of gradient flows and suppress the degradation of the network.

2.2.4. *Patch merging block*

The patch merging blocks are employed for downsampling operations. Specifically, we divided input patches into eight sub-blocks of the same size and concatenated them in the channel dimension. As a result, the spatial resolution of each patch is reduced to half of its original size, while the channel size is expanded by a factor of eight. Then the linear layer is used to reduce the number of channels to a quarter. With such processing, the spatial resolution of input patches reduces to half of the original size while the channel size expands by a factor of two.

2.2.5. *Patch expanding block*

The patch expanding blocks upsample the feature maps extracted by the encoder section, which helps to recover the spatial resolution of the feature maps while reducing their channel size. Specifically, we first reduced the channel dimension of input features to a quarter of the original size using a linear layer, and then used the rearrange operation to reshape the spatial resolution to twice the original size.[17] The patch expanding final block at the end of the decoder section only compresses the channel dimension to one without adjusting the spatial resolution.

2.3. ***Loss function***

The Dice loss function was employed to correct the error between segmentation results and ground truth. This function can measure the similarity between sets and is widely adapted in medical image segmentation. During training, the Dice loss function prefers to excavate foreground regions, which could effectively deal with the imbalance between foreground and background pixels in vascular images.[27–29] Its calculation process is shown in the following equation:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i + \sum_{i}^{N} g_i}, \tag{1}$$

where p_i and g_i, respectively, represent the segmentation result and the ground truth of pixel i, and N is the number of pixels.

2.4. ***Implementation detail***

Our proposed network was implemented using Pytorch. All experiments were conducted on a server with 8 NVIDIA Tesla V100 GPUs. We partitioned each dataset into training, validation, and testing sets, with proportions of 70%, 15%, and 15%, respectively. We evaluated the effectiveness of our method on different datasets. During the training stage, we employed the Adam optimizer to facilitate the convergence of the model, with an initial learning rate set to 0.0001. The learning rate was then reduced to 99% of its previous value every 10 epochs. Due to the limited number of samples in the training set, we applied random cropping and transposition to augment the data. In each epoch of the training process, volumes of size 128 × 128 × 128 pixels were randomly cropped from the original images and randomly transposed the sequence of dimensions (axis).

2.5. *Evaluation metrics*

We utilized seven evaluation metrics to quantify the effectiveness of our segmentation method. The first four metrics, namely Accuracy (Acc), Precision (Prec), Recall (Rec), and $F1$-score ($F1$), are based on the intersection ratio of pixels. The calculation processes for these four metrics are presented in Eqs. (2)–(5). The remaining three metrics, namely centerline Dice (clDice),[30] Hausdorff Distance (HD),[31] and Average Hausdorff Distance (AVD),[32] are utilized to evaluate the similarity of the topological structure. The computation of these three metrics is presented in Eqs. (6)–(10).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{2}$$

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{5}$$

where TP represents the number of positive samples correctly predicted, TN represents the number of positive samples incorrectly predicted, FP represents the number of negative samples correctly predicted, and FN represents the number of negative samples incorrectly predicted.

$$\begin{aligned}&\mathrm{clDice}(V_P, V_L)\\ &\quad = 2 \times \frac{\mathrm{Tprec}(S_P, V_L) \times \mathrm{Tsens}(S_L, V_P)}{\mathrm{Tprec}(S_P, V_L) + \mathrm{Tsens}(S_L, V_P)},\end{aligned} \tag{6}$$

$$\begin{aligned}\mathrm{Tprec}(S_P, V_L) &= \frac{|S_P \bigcap V_L|}{|S_P|},\\ \mathrm{Tsens}(S_L, V_P) &= \frac{|S_L \bigcap V_P|}{|S_L|},\end{aligned} \tag{7}$$

where V_L represents the binary mask of the ground truth, V_P represents the binary mask of the segmentation results, S_L is the skeleton extracted from the ground truth, and S_P is the skeleton extracted from the segmentation results. $\mathrm{Tprec}(S_P, V_L)$ is defined as topological precision, $\mathrm{Tsens}(S_L, V_P)$ is defined as topological sensitivity.

$$H(A, B) = \max\{h(A, B), h(B, A)\}, \tag{8}$$

$$\begin{aligned}h(A, B) &= \max_{a \in A} \min_{b \in B} \|a - b\|,\\ h(B, A) &= \max_{b \in B} \min_{a \in A} \|b - a\|,\end{aligned} \tag{9}$$

where $|| \cdot ||$ represents the distance paradigm between finite point sets A and B. $h(A, B)$ and $h(B, A)$ are respectively called directed HD from point set A to B and directed HD from point set B to A, while $H(A, B)$ is called HD between finite point set A and B.

$$D_{\mathrm{AVD}}(XY) = \frac{\begin{pmatrix}\frac{1}{X}\sum_{x\in X} \min_{y\in Y} d(x, y)\\ +\frac{1}{Y}\sum_{y\in Y} \min_{x\in X} d(x, y)\end{pmatrix}}{2}, \tag{10}$$

where X represents the number of voxels in the ground truth and Y represents the number of voxels in the segmentation results. $\sum_{x\in X} \min_{y\in Y} d(x, y)$ represents the sum of the shortest distance of all points from point set X to Y, dividing the value by X is the directed AVD from the ground truth to the segmentation result. Meanwhile, $\sum_{y\in Y} \min_{x\in X} d(x, y)$ represents the sum of the shortest distance of all points from point set Y to X, dividing this value by Y is the directed AVD from the segmentation result to the ground truth.

3. Results

3.1. *Visualization results*

We conducted training and testing of our method on three datasets described in Sec. 2.1. The visualization results are given and compared with the ground truth.

3.1.1. *Lectin dataset*

The segmentation results of the lectin dataset are shown in Fig. 4. The vascular labeling approach

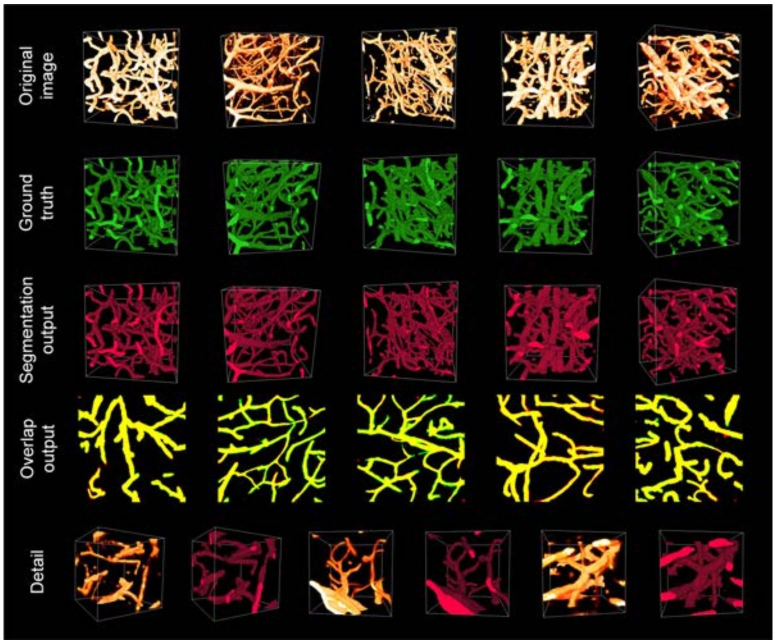


Fig. 4. Segmentation results of the lectin dataset. The overlap output in the fourth row is the overlap between the ground truth (green) and segmentation result (red) of 50 μm maximum intensity projections (MIPs). The yellow part indicates the correct segmentation regions, the red part indicates the false-positive regions, and the green part indicates the false-negative regions. The fifth row shows three detailed areas from the test set and the corresponding segmentation results generated by our proposed method. The size of areas is 100 × 100 × 100 pixels.

used in the lectin dataset was perfusion-based, resulting in lower signal intensities at capillaries and higher intensities at thick vessels. Furthermore, there are numerous fluorescent cell bodies surrounding the vessels. These factors will bring some difficulties to traditional segmentation methods. However, our proposed method effectively overcomes these challenges. The method could efficiently detect and segment capillaries at weak fluorescent signal regions. In addition, our approach could suppress the nonvascular fluorescence interferences and eliminate most of the fluorescent cell bodies in the background. There are few false-positive regions (isolated fluorescent cell bodies are segmented into the foreground) in the segmentation results. The connectivity and topology of segmented vascular networks are also preserved.

3.1.2. *Tek dataset*

The tek dataset utilized transgenic fluorescent proteins for vascular labeling, so the fluorescence signals are generally presented in the vascular wall (the epithelial cells). Vessels in these images are characterized by large vessel lumens. In addition, the background regions also contain some nonvascular fluorescence interferences caused by the expression of fluorescence signals in other regions. The segmentation results are shown in Fig. 5. It can be seen that our method could effectively identify and fill the central cavities as well as remove the fluorescence interferences around the blood vessels. There are few false-positive and false-negative regions, indicating that our method can effectively capture the detailed information in the blood vessels and segment them accurately.

3.1.3. *Gel dataset*

The gel dataset also used a perfusion-based method for vascular labeling. Based on lectin staining, the dataset used FITC-based vessel filling to fill the vessel lumen with a fluorescent gel. Compared with the tek dataset, the vascular lumen in the gel dataset is solid. Due to insufficient perfusion during sample preparation, some vessels had a weak fluorescent signal with a poor signal-to-noise ratio. The topology of vascular networks is difficult to be maintained. However, our method could still effectively identify weak fluorescent signals and segment them (Fig. 6). In addition, there are no significant false-negative regions in the segmentation results,

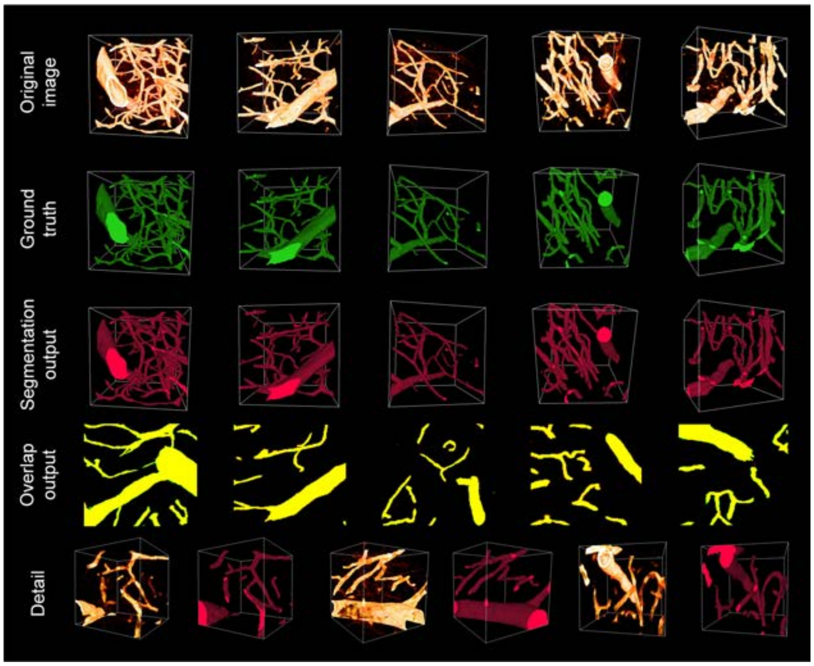


Fig. 5. Segmentation results of the tek dataset. The overlap output in the fourth row is the overlap between the ground truth (green) and segmentation result (red) of 50 μm MIPs. The yellow part indicates the correct segmentation regions, the red part indicates the false-positive regions, and the green part indicates the false-negative regions. The fifth row shows three detailed regions from the test set and the corresponding segmentation results generated by our proposed method. The size of regions is 100 × 100 × 100 pixels.

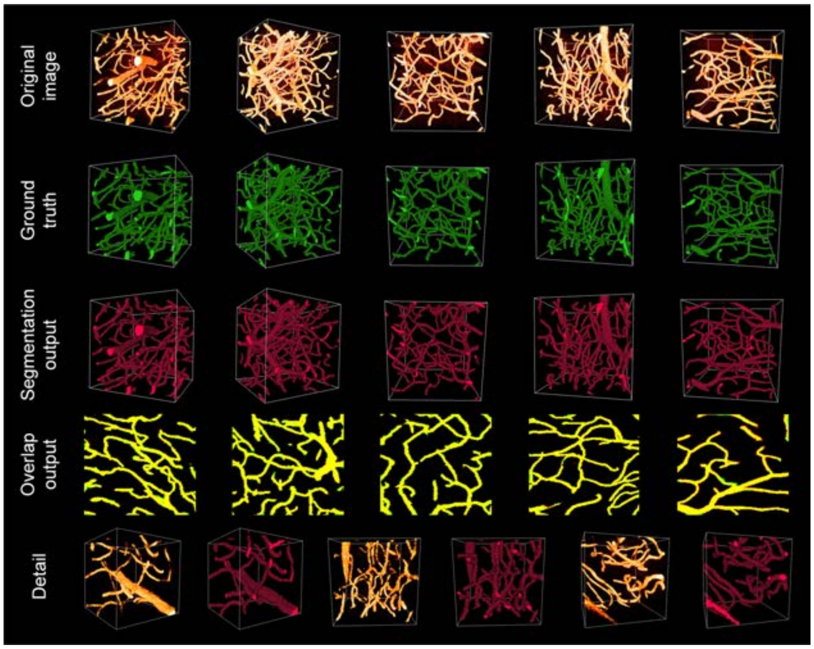


Fig. 6. Segmentation results of the gel dataset. The overlap output in the fourth row is the overlap between the ground truth (green) and segmentation result (red) of 50 μm MIPs. The yellow part indicates the correct segmentation regions, the red part indicates the false-positive regions, and the green part indicates the false-negative regions. The fifth row shows three detailed regions from the test set and the corresponding segmentation results generated by our proposed method. The size of regions is 100 × 100 × 100 pixels.

further representing the powerful segmentation capability of our method in detecting weak fluorescence signal regions.

3.2. *Quantitative results and comparison*

In this subsection, we compare the segmentation results of our method with those of other methods, including V-Net,[27] 3D U-Net,[33] FCN,[34] and VoxResNet.[35] The comparisons are CNN-based methods. Comparison results are shown in Fig. 7. It can be seen that 3D U-Net and VoxResNet have difficulty in handling the cavities in the center of thick vessels and the discontinuities at capillaries. Although V-Net and FCN can fill some lumens in the thick vessels, they still cannot deal with the weak fluorescence signal regions, leading to difficulty in maintaining the connectivity of vascular networks. Furthermore, due to a large number of nonvascular fluorescence signals in the lectin dataset, the comparison methods, especially the FCN model, generated numerous over-segmentation regions around the thick vessels. In contrast, the segmentation results of the SegVesseler model are significantly better than the comparisons. Our method accurately identifies discontinuities in some weak fluorescence signal regions and effectively fills the thick vessel lumens. The over-segmentation regions caused by fluorescence interferences are minimized.

To further validate segmentation effectiveness, we present quantitative evaluation results of the three datasets in Tables 1–3. The quantitative results are consistent with the visualization results. Our approach achieved the best result for almost all evaluation metrics.

On the one hand, our method can identify and fill vessel lumens in large vessels and effectively segment capillaries with weak fluorescence signals.

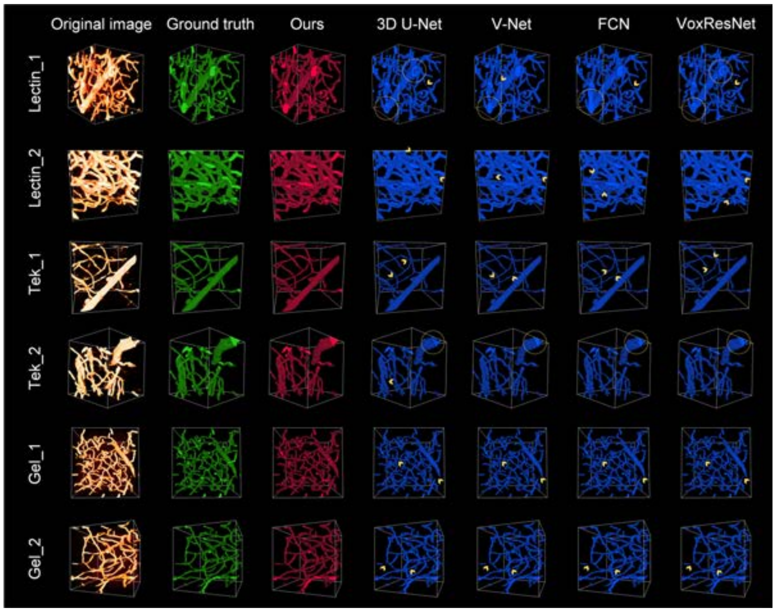


Fig. 7. Comparative segmentation results between the proposed method and other methods. Some missing or false segmentations are highlighted with yellow arrows and circles.

Table 1. Quantitative results of the lectin dataset.

	Acc	Prec	Rec	*F*1	HD	AVD	clDice
3D U-Net	98.44	85.49	87.13	85.49	37.05	0.2887	96.37
V-Net	98.30	85.37	86.26	84.83	**35.89**	0.4455	94.46
FCN	98.27	84.49	87.13	84.75	37.69	0.4338	91.79
VoxResNet	98.58	**89.60**	84.19	86.02	36.17	0.2779	96.64
Ours	**98.65**	87.19	**88.63**	**87.20**	37.42	**0.2426**	**97.33**

Table 2. Quantitative results of the tek dataset.

	Acc	Prec	Rec	$F1$	HD	AVD	clDice
3D U-Net	99.41	92.42	95.12	93.45	33.71	0.2188	94.91
V-Net	99.40	**96.37**	92.17	93.85	34.35	0.1802	96.91
FCN	99.26	93.88	93.11	92.90	40.31	0.3068	94.76
VoxResNet	99.54	92.87	95.62	94.04	36.90	0.2024	95.96
Ours	**99.68**	95.07	**97.59**	**96.23**	**33.54**	**0.0722**	**97.47**

Table 3. Quantitative results on the gel dataset.

	Acc	Prec	Rec	$F1$	HD	AVD	clDice
3D U-Net	99.32	90.06	89.50	88.70	19.67	0.1359	94.28
V-Net	99.34	**92.67**	86.10	88.36	**16.32**	0.1328	94.54
FCN	99.29	90.13	87.80	88.07	20.00	0.1374	95.19
VoxResNet	99.35	92.37	86.74	88.53	17.17	0.1348	94.54
Ours	**99.51**	92.09	**91.34**	**91.34**	19.91	**0.0974**	**96.31**

Compared with the highest score of comparison methods, our method improved the $F1$-score by 1.18%, 2.19%, and 2.64%, and the Recall by 1.50%, 1.97%, and 1.84%, respectively, which indicates that the similarity between the segmentation results of our method and the ground truth is preferable. On the other hand, since the powerful global information extraction capability of the 3D Swin Transformer block can effectively maintain the connectivity and topology of vascular networks, our method also achieves optimal evaluation results on the clDice metric (improved by 1.18%, 2.19%, and 2.64% respectively). Additionally, our method achieves the highest HD score on the tek dataset, while V-Net achieves the highest on the lectin and gel datasets. This is due to isolated noisy points in the lectin and gel segmentation results, which caused some unreasonable distances. However, although our method cannot achieve the highest HD score on all datasets, it achieves optimal evaluation results on the AVD metric, as it reduces the unreasonable distances caused by outlier points during calculation and pays more attention to the overall segmentation effect.

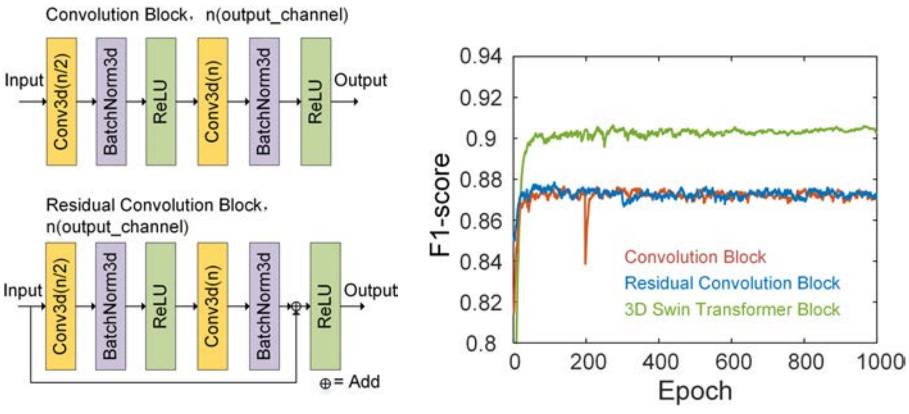
3.3. *Ablation studies*

In this subsection, two ablation experiments were designed to validate the effectiveness of the SegVesseler architecture for vascular segmentation.

In ablation experiment I, we replaced 3D Swin Transformer blocks with convolution blocks and residual convolution blocks respectively, and compared them with the original network. The structures of the two convolution blocks are shown in Fig. 8(a). The experiment was done on the gel dataset, and the results are shown in Fig. 8(b). The $F1$-score of the comparison methods decreased by approximately 2–3% compared to our method. This outcome indicates that the 3D Swin Transformer block has advantages in extracting deep semantic information compared with the traditional 3D convolution operation. Its powerful global modeling capability helps it maintain the topology of vascular networks during segmentation. Therefore, our method could improve the effectiveness of vascular segmentation effectively.

In ablation experiment II, we employed cross-validations to examine the generalizability of our method in cross-dataset tasks. Specifically, We selected two of the lectin, tek, and gel datasets as training sets and the remaining one as the test set. We compared the experimental results of our method with those of VoxResNet, which was the best-performing among the comparison models. The experimental results are shown in Table 4.

The experimental results demonstrate the strong generalizability of our method, achieving optimal values in nearly all evaluation metrics. We attribute this success to the Swin Transformer's ability to extract deep semantic features, while the traditional convolution operation focuses more on shallow pixel features. The SegVesseler architecture efficiently learns the tubular structure and connectivity



(a)

(b)

Fig. 8. Ablation study of different feature extraction modules. (a) The structures of the two convolution blocks. (b) The $F1$-score of each epoch on the gel validation set.

Table 4. Quantitative results of the cross-validation tasks, where L, T, and G represent lectin, tek, and gel datasets, respectively.

	Training			Testing						
Net	L	T	G	L	T	G	$F1$	clDice	HD	AVD
Ours	√	√				√	**87.91**	**84.43**	32.05	**0.1763**
VoxResNet							81.62	81.24	**28.89**	0.5948
Ours	√		√		√		**77.71**	**86.19**	44.70	**0.8998**
VoxResNet							55.59	67.83	**38.38**	1.6434
Ours		√	√	√			**82.66**	**88.97**	**39.07**	**0.4553**
VoxResNet							80.47	87.59	40.02	0.6825

information in vascular images, enabling it to cope with differences arising from diverse imaging and labeling modalities. However, it is difficult for our method to achieve the highest score among all HD metrics, mainly caused by isolated fluorescent cells in the segmentation results (the reasons for this result are detailed in Sec. 3.2).

4. Discussion

Vessels in mesoscopic optical vascular images are foreground regions. However, vascular discontinuities generated by weak fluorescence signals or vessel lumens are represented as low grayscale values in images, which may be incorrectly classified as background regions at the pixel level during segmentation. Conversely, the grayscale value of nonvascular fluorescence signals is similar to the foreground signals in images, so there may be some false positives in the segmentation results. These issues indicate that segmenting vascular networks can not only consider vascular features at the pixel level but also need to consider deep semantic information. Therefore, we proposed SegVesseler to capture deep semantic information in vascular images with larger receptive fields and maintain connectivity and topology of the vascular networks during segmentation. In contrast to other methods, our approach efficiently identifies foreground and background regions in images, achieving state-of-the-art results.

Different labeling and imaging modalities can result in significant differences in the characteristics of mesoscopic optical vascular images. For the tek dataset, transgenic fluorescent proteins were used to label vascular endothelial cells with high fluorescent signal intensity, resulting in higher image signal-to-noise ratios at both large vessels and capillaries compared to perfusion-based labeling methods. In lectin and gel datasets, vessels are labeled by perfusion-based methods. The difference is that lectin marks the vessel wall, whereas the fluorescent gel could fill the entire vessel lumens. Due to insufficient perfusion during sample preparation, some vessels, especially in capillaries, had a weak fluorescent signal and poor signal-to-noise ratio. The number of these "discontinuity" features is so small that it is difficult to capture them when network training. Thus, the segmentation results of the perfusion-based labeled datasets (lectin, gel) are slightly less effective compared to the transgenic fluorescent protein-labeled dataset (tek).

Although our method has achieved exciting segmentation results, it has some weaknesses. Specifically, due to the weak fine-grained feature extraction capability of the Transformer (compared with CNN), our method has an ineffective segmentation at the edges of the vessels, and there are some isolated noise points in the segmentation results. In contrast, traditional CNN-based methods have some advantages in processing image details. The quantitative results in Sec. 3.2 reflect this viewpoint well, and our method achieves optimal values in almost all metrics except HD. Therefore, although our method suffers from inaccuracy in segmenting the individual voxels at the edge of blood vessels, its capability to accurately maintain the connectivity and topology of vascular networks is significant in reflecting the actual physiological structure of the vasculatures.

5. Conclusions

In this study, we have proposed a vascular segmentation method based on Swin Transformer. The proposed method utilized the 3D Swin Transformer block as a feature extractor, which could capture deep semantic information in vascular images with a larger receptive field, and overcome the limitations of traditional CNN-based methods. We evaluated the approach on three different mesoscopic optical cerebrovascular images and achieved state-of-the-art results on all datasets. In future work, we hope to utilize this method to complete the reconstruction of whole-brain vessels and apply it to various vessel analysis tasks, such as vascular-related brain disease analysis, mapping of the vascular atlas, and studying vascular morphogenesis.

Acknowledgments

This work was supported by the STI2030-Major Projects (2021ZD0201002); the National Natural Science Foundation of China (82102137, T2122015); Natural Science Foundation of Shaanxi Provincial Department of Education (21JK0796); the Open Project Program of Wuhan National Laboratory for Optoelectronics (2021WNL OKF006) and the Natural Science Foundation of Sichuan Province (2022NSFSC0964).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. V. Muoio, P. B. Persson, M. M. Sendeski, "The neurovascular unit — concept review," *Acta Physiol.* **210**(4), 790–798 (2014).
2. B. J. Andreone, B. Lacoste, C. Gu, "Neuronal and vascular interactions," *Annu. Rev. Neurosci.* **38**, 25–46 (2015).
3. J. Wu, Y. He, Z. Yang, C. Guo, Q. Luo, W. Zhou, S. Chen, A. Li, B. Xiong, T. Jiang, "3D BrainCV: Simultaneous visualization and analysis of cells and capillaries in a whole mouse brain with one-micron voxel resolution," *Neuroimage* **87**, 199–208 (2014).
4. S. Moccia, E. De Momi, S. El Hadji, L. S. Mattos, "Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics," *Comput. Methods Prog. Biomed.* **158**, 71–91 (2018).
5. D. Lesage, E. D. Angelini, I. Bloch, G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Med. Image Anal.* **13**(6), 819–845 (2009).
6. D. Q. Jia, X. H. Zhuang, "Learning-based algorithms for vessel tracking: A review," *Comput. Med. Imag. Graph.* **89**, 101840 (2021).
7. Y. F. Shang, R. Deklerck, E. Nyssen, A. Markova, J. de Mey, X. Yang, K. Sun, "Vascular active contour for vessel tree segmentation," *IEEE Trans. Biomed. Eng.* **58**(4), 1023–1032 (2011).
8. Y. Z. Cheng, X. Hu, J. Wang, Y. D. Wang, S. Tamura, "Accurate vessel segmentation with constrained B-snake," *IEEE Trans. Image Process.* **24**(8), 2440–2455 (2015).
9. S. Wang, B. Peplinski, L. Lu, W. Zhang, J. Liu, Z. Wei, R. M. Summers, "Sequential Monte Carlo tracking for marginal artery segmentation on CT angiography by multiple cue fusion," *Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Nagoya, Japan, pp. 518–525 (2013).

10. A. F. Frangi, W. J. Niessen, K. L. Vincken, M. A. Viergever, "Multi-scale vessel enhancement filtering," *Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cambridge, MA, USA, pp. 130–137 (1998).
11. Y. Zhao, Y. Zheng, Y. Liu, Y. Zhao, L. Luo, S. Yang, T. Na, Y. Wang, J. Liu, "Automatic 2D/3D vessel enhancement in multiple modality images using a weighted symmetry filter," *IEEE Trans. Med. Imag.* **37**(2), 438–450 (2018).
12. Y. Y. Wang, O. V. Glinskii, F. Bunyak, K. Palaniappan, "Ensemble of deep learning cascades for segmentation of blood vessels in confocal microscopy images," *IEEE Conf. Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–7, IEEE (2021).
13. W. Tahir, S. Kura, J. Zhu, X. Cheng, R. Damseh, F. Tadesse, A. Seibel, B. S. Lee, F. Lesage, S. Sakadžic, "Anatomical modeling of brain vasculature in two-photon microscopy by generalizable deep learning," *BME Front.* **2020**, 8620932 (2020).
14. R. Damseh, P. Pouliot, L. Gagnon, S. Sakadzic, D. Boas, F. Cheriet, F. Lesage, "Automatic graph-based modeling of brain microvessels captured with two-photon microscopy," *IEEE J. Biomed. Health Inform.* **23**(6), 2551–2562 (2019).
15. M. Haft-Javaherian, L. Fang, V. Muse, C. B. Schaffer, N. Nishimura, M. R. Sabuncu, "Deep convolutional neural networks for segmenting 3D *in vivo* multi-photon images of vasculature in Alzheimer disease mouse models," *PLoS One* **14**(3), e0213539 (2019).
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszko-reit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," *Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
17. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, Y. Zhou, "TransUnet: Transformers make strong encoders for medical image segmentation," preprint (2021), arXiv: 2102.04306.
18. Y. Zhang, H. Liu, Q. Hu, "Transfuse: Fusing Transformers and CNNs for medical image segmentation," preprint (2021), arXiv:2102.08005.
19. C. Chen, K. Zhou, Z. Wang, R. Xiao, Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA," *IEEE Trans. Med. Imag.* **42**(2), 346–353 (2023).
20. Q. Wu, Y. Chen, N. Huang, X. Yue, "Weakly-supervised cerebrovascular segmentation network with shape prior and model indicator," *Int. Conf. Multimedia Retrieval (ICMR)*, Newark, NJ, USA, pp. 668–676 (2022).
21. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows," *IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 9992–10002, IEEE (2021).
22. L. Zhang, Y. Wen, "A transformer-based framework for automatic COVID19 diagnosis in chest CTs," *IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 513–518, IEEE (2021).
23. Z. Liu, Y. Tan, Q. He, Y. Xiao, "SwinNet: Swin Transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.* **32**(7), 4486–4497 (2022).
24. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, "Swin-Unet: Unet-like pure Transformer for medical image segmentation," preprint (2021), arXiv:2105.05537.
25. H. Gong, D. Xu, J. Yuan, X. Li, C. Guo, J. Peng, Y. Li, L. A. Schwarz, A. Li, B. Hu, B. Xiong, Q. Sun, Y. Zhang, J. Liu, Q. Zhong, T. Xu, S. Zeng, Q. Luo, "High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level," *Nat. Commun.* **7**, 12142 (2016).
26. Q. Zhong, A. Li, R. Jin, D. Zhang, X. Li, X. Jia, Z. Ding, P. Luo, C. Zhou, C. Jiang, Z. Feng, Z. Zhang, H. Gong, J. Yuan, Q. Luo, "High-definition imaging using line-illumination modulation microscopy," *Nat. Methods* **18**(3), 309–315 (2021).
27. F. Milletari, N. Navab, S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Int. Conf. 3D Vision*, Stanford, CA, pp. 565–571 (2016).
28. Y. Jiang, J. Liang, T. Cheng, X. Lin, Y. Zhang, J. Dong, "MTPA_Unet: Multi-scale Transformer-position attention retinal vessel segmentation network joint Transformer and CNN," *Sensors* **22**(12), 4592 (2022).
29. T. A. Soomro, A. J. Afifi, J. Gao, O. Hellwich, M. Paul, L. Zheng, "Strided U-Net model: Retinal vessels segmentation using dice loss," *Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, ACT, Australia, pp. 1–8 (2018).
30. S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. Pluim, U. Bauer, B. H. Menze, "clDice — a novel topology-preserving loss function for tubular structure segmentation," *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 16560–16569, IEEE (2021).
31. D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 850–863 (1993).
32. O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, V. I. Madai, "On the usage of average Hausdorff distance for

segmentation performance assessment: Hidden bias when used for ranking," preprint (2020), arXiv:2009.00215.

33. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," preprint (2016), arXiv:1606.06650.
34. J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, IEEE (2015).
35. H. Chen, Q. Dou, L. Yu, J. Qin, P. A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage* **170**, 446–455 (2017).