TemSAM: Temporal-aware Segment Anything Model for Cerebrovascular Segmentation in Digital Subtraction Angiography Sequences

Liang Zhang[1], Xixi Jiang[2], Xiaohuan Ding[1], Zihang Huang[1], Tianyu Zhao[1], and Xin Yang[1] (✉)

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology
[2] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
xinyang2014@hust.edu.cn

Abstract. Digital Subtraction Angiography (DSA) is the gold standard in vascular disease imaging but it poses challenges due to its dynamic frame changes. Early frames often lack detail in small vessels, while late frames may obscure vessels visible in earlier phases, necessitating time-consuming expert interpretation. Existing methods primarily focus on single-frame analysis or basic temporal integration, treating all frames uniformly and failing to exploit complementary inter-frame information. Furthermore, existing pre-trained models like the Segment Anything Model (SAM), while effective for general medical video segmentation, fall short in handling the unique dynamics of DSA sequences driven by contrast agents. To overcome these limitations, we introduce TemSAM, a novel temporal-aware segment anything model for cerebrovascular segmentation in DSA sequences. TemSAM integrates two main components: (1) a multi-level Minimum Intensity Projection (MIP) global prompt that enhances temporal representation through a MIP-guided Global Attention (MGA) module, utilizing global information provided by MIP, and (2) a complementary information fusion module, which includes a frame selection module and a Masked Cross-Temporal Attention Module, enabling additional foreground information extraction from complementary frame. Our experimental results demonstrate that TemSAM significantly outperforms existing methods. Our code is available at https://github.com/zhang-liang-hust/TemSAM.

Keywords: Digital Subtraction Angiography · Segment Anything Model · Vessel Segmentation.

1 Introduction

Digital Subtraction Angiography (DSA) [2] is highly effective in displaying vascular abnormalities, making it the gold standard for the diagnosis and treatment planning of vascular diseases such as blood flow, stenosis, and thrombosis. DSA sequences (Fig. 1-(a)) capture only transient snapshots of the dynamic vascular structure, with early frames often lacking detail in thin vessels (yellow arrow) and late frames providing poor visualization of the early phase vessels (red arrow). In contrast, its corresponding minimum intensity projection (MIP) image (Fig. 1-(b)) exhibits good visibility of the entire vessel structure. Meanwhile, unlike video object segmentation tasks in natural scenes, the apparent emergence or disappearance of vascular structures in DSA sequences reflects temporal variations in contrast agent distribution rather than physical movement, necessitating expert interpretation [4]. Consequently, the analysis of DSA sequences heavily relies on the expertise of radiologists, a process that is both time-consuming and labor-intensive. This has driven the development of fully automated methods for segmenting vascular structures in DSA sequences.

Recently, Convolutional Neural Networks (CNN)-based methods [11, 19, 13, 14, 16, 17, 21] have been developed for automatic vessel segmentation. Some focus on single-frame DSA analysis. For instance, Zhang et al. [21] firstly proposed a U-shaped network for cerebrovascular segmentation in single-frame DSA images. Xu et al. [17] proposed an Edge Regularization Network for cerebral vessel segmentation, using erosion-dilation for pseudo-labels generation and a Hybrid Fusion Module for refined predictions. Other studies utilize temporal information in DSA sequences. Su et al. [13] introduced CAVE, which uses a ConvGRU module to encode temporal features from 2D+time DSA series for A/V segmentation. Xie et al. [16] developed DSNet, a spatio-temporal network that incorporates MIP images as a spatial branch to enhance accuracy. However, CNN-based approaches still face challenges such as limited domain-specific training samples and restricted representation capabilities due to inadequate model capacity, making it difficult to achieve optimal vessel segmentation in DSA sequences.

The recently proposed Segment Anything Model (SAM) [6], trained on the large-scale data set SA-1B, has powerful feature extraction capabilities and can accurately focus on semantics of interest based on user prompts (e.g., points and bounding boxes). Hence, several studies have applied SAM to improve segmentation accuracy in medical scenarios [1, 3, 5, 8, 9, 15, 22]. For 2D medical images, adapter-based methods have been proposed to adapt SAM on medical datasets. For instance, Med-SA [15] designs a lightweight bottleneck composed of a down-projection, ReLU activation, and up-projection for fine-tuning. Several approaches have been developed to enhance the Segment Anything Model (SAM) for medical video segmentation by incorporating temporal strategies. MediViSTA-SAM [5] integrates cross-frame attention into SAM to capture temporal information. MedSAM2 [22], a generalized auto-tracking model based on SAM2 [10], treats video segmentation as an object tracking problem. Despite these advancements, existing methods either focus solely on single-frame analysis without integrating temporal information or indiscriminately treat all frames equally, overlooking the fact that certain frames, being more informative, can offer enhanced guidance for the target vessels.

To address the above challenges, we propose a novel method for vessel segmentation in DSA sequences based on SAM, named TemSAM, which adopts a local-global fusion strategy. TemSAM fully utilizes the global information provided by the MIP image and the supplementary information from complementary frames to enhance contextual representation. Specifically, *we design a multi-level MIP global prompt that enables parallel encoding of video clips and MIP image, using a MIP-guided Global Attention module to integrate global information from the MIP image.* The MIP global feature also serves as dense prompts to guide the decoding process. Additionally, *we develop a novel complementary information fusion module consisting of a frame selection module and a Masked Cross-Temporal Attention (MCTA) module to achive complementary information fusion.* The frame selection module identifies complementary frames with significant foreground differences compared to current video clip. The MCTA module then aggregates features from complementary frames and temporal information using preliminary segmentation maps, thereby facilitating high-precision vessel segmentation. In comparisons with various existing methods, we achieve state-of-the-art performance for cerebrovascular segmentation in DSA sequences, particularly excelling in the segmentation of thin vessels.

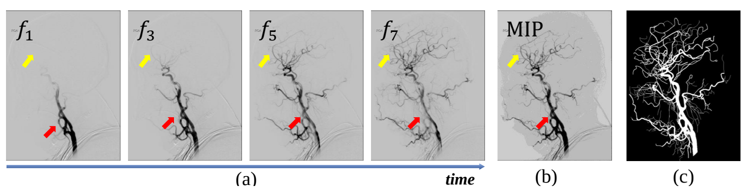

Fig. 1: Overview of the DSA sequence. (a) is a DSA sequence. (b) is the MIP image of the DSA sequence. (c) is the ground-truth of the DSA sequence.

2 Proposed Method

2.1 The Overall Architecture of TemSAM

Given a DSA sequence $x = \{x_1, x_2, \ldots, x_n\}$ and an MIP image, our objective is to predict a final segmentation map for the input sequence. As shown in Fig. 2, the DSA sequence comprises two components: current video clips $clip_t$ (in yellow bounding boxes) and remaining frames f_c (in green bounding boxes). f_c contains supplementary contextual information for $clip_t$. MIP image f_m provides comprehensive vascular structures information across all phases. We adopt a local-global fusion strategy: each $clip_t$ generate a prediction $Pred_t$ and the final prediction $Pred$ is the average of all clips' prediction, i.e., $Pred = Avg(\sum_{i=1}^{n} Pred_i)$, where n is the number of frames.

The overall architecture of TemSAM is built upon SAM. TemSAM employs a dual-branch image encoder to separately extract temporal information from local clips and global structural information from MIP images. To adapt SAM for cerebrovascular segmentation, we fine-tune it with an adapter module. In the local temporal branch and the global prompt branch, we utilize S-Adapter [15] and T-Adapter [5] separately. To effectively leverage MIP-derived global knowledge, we design a MIP-guided global attention (MGA) module between two branches, providing global context guidance for local feature extraction. Additionally, the deepest global feature F_g is fed into the mask decoder as dense prompts. Furthermore, we select the most complementary temporal information from the entire DSA sequence for the current clip and use it to refine the features of the current clip. We perform frame selection using cosine similarity, sampling the least similar one as complementary frame F_c from the remaining frames' features F_r with respect to F_t. To more effectively and precisely guide the integration of complementary information in foreground vascular regions, we introduce a two-stage hierarchical mask decoder. In the first stage, TemSAM uses SAM's original decoder to generate initial predictions M. In the second stage, we employ a Masked Cross-Temporal Attention (MCTA) module to guide feature interaction between F_t and F_c within the foreground region using mask M.

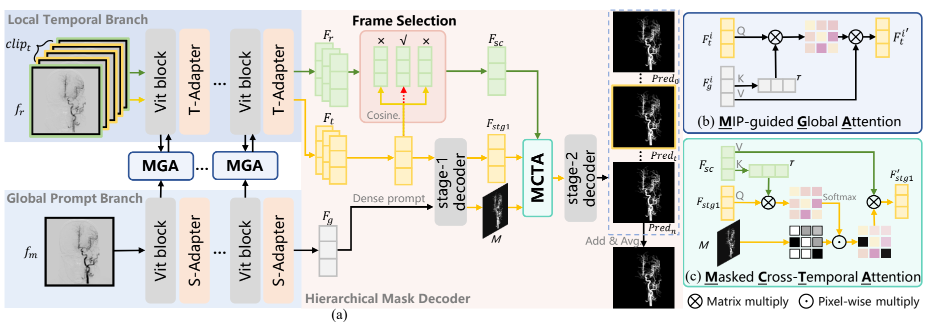

Fig. 2: (a) Overview of the proposed TemSAM. TemSAM integrates an adapter-based dual-branch image encoder coupled with a two-stage hierarchical decoder. (b) MIP-guided Global Attention module. (c) Masked Cross-Temporal Attention module.

2.2 Multi-level MIP Global Prompt

Dual-branch Encoder. the MIP image provides vascular prior information for the DSA sequence, and thus, we leverage MIP images as prompts to enhance feature extraction of temporal information. In our method, both the MIP image and DSA sequences are fed into the encoder for parallel encoding. Specifically, the global prompt branch employs an S-Adapter [15] in 2D dimension, which sequentially performs dimension reduction, ReLU activation, and dimension restoration. The local temporal branch utilizes a T-Adapter based on multi-head self-attention (MHSA) mechanisms [5] along the temporal dimension to capture temporal information. The procedure of the two adapters can be formulated as:

$$\mathcal{A}_s(x) = RELU(xW_d)W_u \tag{1}$$

$$\mathcal{A}_t(x) = MHSA(Reshape(x \in \mathbb{R}^{(B\times T)\times H\times W\times C} \rightarrow x \in \mathbb{R}^{(B\times H\times W)\times T\times C})) \tag{2}$$

where $W_d \in \mathbb{R}^{d\times\frac{d}{4}}$ and $W_u \in \mathbb{R}^{\frac{d}{4}\times d}$ are the projection matrices, d is the dimension of the feature. Through these operations, S-Adapter and T-Adapter enable the encoder to better adapt to the feature distribution of medical image domains.

MIP-guided Global Attention (MGA). The MGA module utilizes the global information of MIP to help aggregate the temporal information within the local window. This module establishes hierarchical interaction between global MIP features F_g^i and local temporal features F_t^i in multiple layers. MGA enables the MIP image to progressively guide and refine the extraction of local temporal information throughout the network. As shown in Fig. 2(b), we formulate F_t^i as queries while designating F_g^i as corresponding keys and values. The MGA can be formulated as:

$$F_t^{i'} = \text{softmax}(F_t^i W_q \cdot (F_g^i W_k)^T / \sqrt{d_m}) \cdot (F_g^i W_v) \tag{3}$$

where d_m is the dimension of MGA and i represents the current layer number in the encoder. $W_q \in \mathbb{R}^{d\times d_m}$, $W_k \in \mathbb{R}^{d\times d_m}$ and $W_v \in \mathbb{R}^{d\times d_m}$ are the learnable weight matrices used to project F_t^i and F_g^i to different subspaces. The output of MGA $F_t^{i'}$ is enhanced temporal feature which will be added to F_t^i and then used in subsequent encoding layer.

2.3 Complementary Information Fusion Module

Frame selection. Local video clip often lacks some information regarding vascular visibility, while distant frames can provide necessary complementary information. Therefore, we aim to identify the most complementary frames for $clip_t$. Specifically, given current clip's mean feature $F_t^{'}$ (generated by averaging F_t) and remaining frames' features F_r , we select the feature F_c that is most dissimilar to F_t from m candidate frames. This module outputs the index of selected frame with the lowest similarity, which can be interpreted as

$$ind = argmin_{i\in[1,m]}\{Sim(F_r, F_t^{'})\} \tag{4}$$

where idx is the index of the selected complementary frame, $i \in [1, m]$ indicates the ith candidate frames, and $argmin$ is used to calculate the index of the frame with the lowest similarity, and Sim is the similarity function as follows:

$$\text{Sim}(X, Y) = (X \cdot Y) / \sqrt{dim} \tag{5}$$

According to equ.5, cosine distance is used to calculate the similarity between the F_r and $F_t^{'}$, and dim represents the dimension of the input feature.

Masked Cross-Temporal Attention (MCTA). The MCTA module establishes a connection between the stage-1 decoder's feature F_{stg1} and the complementary frame's feature F_c. Unlike cross-attention which attends to the global context, MCTA is guided by the predictions from the stage-1 and operates within the predicted mask, thereby further supplementing missing local features in the foreground areas. The MCTA module can be formulated as:

$$F_{stg1}^{'} = M \odot \mathrm{softmax}(F_{stg1}W_q \cdot (F_{sc}W_k)^T / \sqrt{d_m}) \cdot F_{stg1}W_v \tag{6}$$

which employs the probabilistic map M of stage-1 resized to the same spatial resolution as the attention map. W_q, W_k and W_v are the learnable weight matrices and d_m is the dimension of MCTA. With pixel-wise multiplication between M and the attention map, background will be ignored by multiplying a near-zero probability. The output $F_{stg1}^{'}$ is added to F_{stg1} and fed into the stage-2 for subsequent decoding process.

Additionally, the training loss will be applied to each stage in our mask decoder. This ensures thorough supervision and facilitates high-resolution prediction through hierarchical feature fusion.

3 Experiments

Datasets. The proposed TemSAM is evaluated on two cerebrovascular segmentation datasets DIAS [7] and DSCA [20]. DIAS annotates 60 sequences with 321 frames. DSCA includes 1792 frames from 224 sequences. All DSA sequences are resampled to a length of 8. We train on DIAS and split the training, validation, and test sets with a ratio of 3:1:2 following [7]. The DSCA dataset is used to test the generalization ability.
Evaluation Metrics. We evaluate the segmentation performance using DSC, clDice [12], IoU, Acc, and AUC. Given that accurately segmenting thin vessels presents a significant challenge, we conduct evaluation for vessels thinner than 7 pixels, following [18]. A 5-pixel search range is assigned to each thin vessel, and pixels within this range are counted for pixel-to-pixel matching. The performance is evaluated using three metrics: DSC_{thin}, Acc_{thin} and IOU_{thin}.
Implementation Details. We adopt Med-SA [15] as the baseline. The model is initialized with pre-trained ViT-B weights from SAM and implemented using Pytorch on three NVIDIA RTX A6000 GPUs. All frames are resized to 800×800 pixels with a clip length of 3. We apply data augmentation, including horizontal and vertical flipping, brightness/contrast adjustment, and random rotations. AdamW optimizer is utilized ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a StepLR scheduler that decreases the learning rate by 0.9 every 10 epochs. For a fair comparison, all models are re-implemented and trained for 200 epochs under the same settings.
Comparison with SOTA Methods. To demonstrate the superiority of our proposed method, we compare our method with some state-of-the-art (SOTA) methods, including *task-specific* methods [7, 13, 11], *SAM-based 2D* methods [8, 1, 15], and *SAM-based temporal* methods [3, 5, 9, 22].

As quantitatively demonstrated in Tab.1 and 2, TemSAM significantly outperforms existing SOTA task-specific methods across all evaluation metrics. These improvements can be primarily attributed to the multi-level MIP prompt,

Table 1: Comparative results on DIAS dataset.

Type	Method	DSC	Acc	IOU	AUC	clDice	DSC_{thin}	Acc_{thin}	IOU_{thin}
task-specific	VSS-Net [7]	0.7577	0.9631	0.6131	0.9762	0.6794	0.7186	0.9681	0.5641
	3D-UNet [11]	0.7618	0.9637	0.6184	0.9818	0.6879	0.7156	0.9678	0.5676
	ST-UNet [13]	0.7702	0.9630	0.6340	0.9835	0.7122	0.7147	0.9602	0.5630
SAM-based 2D	MedSAM [8]	0.6656	0.9512	0.5004	0.9339	0.5181	0.5592	0.9595	0.3907
	SAM-Med2D [1]	0.6948	0.9556	0.5344	0.9597	0.5684	0.6220	0.9599	0.4537
	Med-SA [15]	0.7107	0.9548	0.5526	0.9688	0.6277	0.6534	0.9597	0.4869
SAM-based temporal	VP-SAM [3]	0.5756	0.9406	0.4072	0.9363	0.4284	0.5188	0.9435	0.3536
	MedSAM2 [22]	0.7070	0.9549	0.5485	0.9697	0.6126	0.6321	0.9592	0.4629
	SAM+ST-Adapter [9]	0.7548	0.9605	0.6074	0.9826	0.7047	0.7011	0.9635	0.5413
	MediViSTA-SAM [5]	0.7597	0.9631	0.6142	0.9827	0.6898	0.7035	0.9659	0.5445
	ours	**0.7816**	**0.9655**	**0.6428**	**0.9836**	**0.7330**	**0.7360**	**0.9686**	**0.5834**

Table 2: Generalization comparison on DSCA Dataset.

Type	Method	DSC	Acc	IOU	AUC	clDice	DSC_{thin}	Acc_{thin}	IOU_{thin}
task specific	VSS-Net [7]	0.6464	0.9742	0.4862	0.9101	0.6596	0.5233	0.9767	0.3824
	3D-UNet [11]	0.7077	0.9739	0.5514	0.9790	0.5990	0.6394	0.9795	0.5706
	ST-UNet [13]	0.7286	0.9762	0.5766	0.9878	0.6265	0.6623	0.9789	0.5756
SAM-based 2D	MedSAM [8]	0.6055	0.9682	0.4375	0.9486	0.4062	0.4391	0.9740	0.3191
	SAM-Med2D [1]	0.6337	0.9679	0.4678	0.9577	0.4920	0.5580	0.9737	0.4665
	Med-SA [15]	0.6660	0.9702	0.5032	0.9660	0.5460	0.5948	0.9747	0.5212
SAM-based temporal	VP-SAM [3]	0.5151	0.9622	0.3500	0.9363	0.3249	0.4242	0.9654	0.3211
	MedSAM2 [22]	0.7286	0.9721	0.5753	0.9837	0.6341	0.6661	0.9766	0.6009
	SAM+ST-Adapter [9]	0.7241	0.9748	0.5703	0.9858	0.6342	0.6505	0.9769	0.6088
	MediViSTA-SAM [5]	0.7392	0.9771	0.5896	0.9882	0.6385	0.6640	0.9788	0.5865
	ours	**0.7476**	**0.9779**	**0.6044**	**0.9933**	**0.6596**	**0.6817**	**0.9800**	**0.6281**

the advanced complementary information fusion module, and the effective integration of SAM's inherent design advantages. Furthermore, the 2D variant of SAM exhibits significantly inferior performance, underscoring the critical role of temporal information in DSA sequences. Notably, our approach surpasses existing SAM-based temporal methods, demonstrating superior capability in both temporal information utilization and contextual feature extraction.

Qualitative results, as shown in Fig. 3, visually demonstrate TemSAM's performance advantages over other methods. TemSAM effectively preserves vascular connectivity and enhances the detection of thin vessels in low-contrast regions. Meanwhile, TemSAM achieves an optimal balance between false positives and false negatives, demonstrating remarkable robustness in challenging clinical scenarios (1^{st} row).

Effectiveness of each component in TemSAM. We conducted ablation studies by incrementally integrating each component (i.e. Dual-Branch Encoder (DBE), MGA Module, and MCTA Module) into the backbone. Additionally, we compare our Local-Global Fusion (LGF) strategy backbone with the Global Fu-

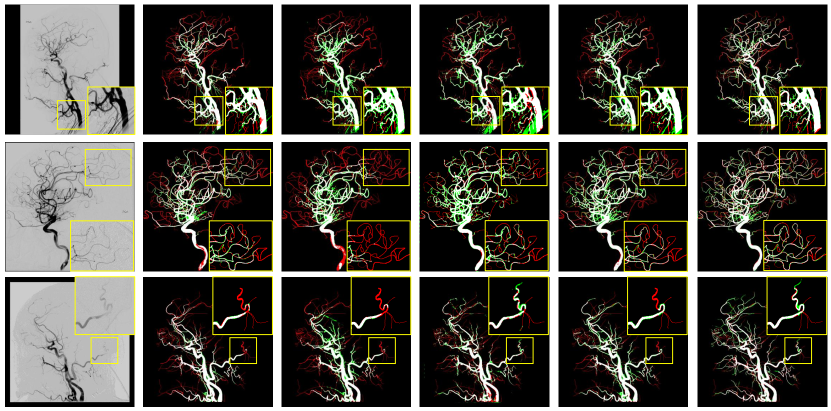

Fig. 3: Qualitative comparison results of cerebrovascular segmentation in DSA sequences. The white, green, and red denote the true positive, false positive, and false negative, respectively. Enlarged local viewing for better clarity.

sion (GF) strategy, which inputs the entir DSA sequence and directly outputs one final segmentation map. The results demonstrate that the LGF strategy preforms superior to GF strategy. This superiority can be attributed to the local video clip's ability to integrate local temporal information. In contrast, a global views may miss certain vessels only visible for a short time. Meanwhile, incorporating any single component of TemSAM significantly enhances the segmentation performance. Notably, combining all components achieve the optimal segmentation accuracy, especially for thin vessels (DSC_{thin}+1.92%, Acc_{thin}+0.16%, and IOU_{thin}+2.32%).

Table 3: Ablation study on different component combinations of TemSAM.

DBE	MGA	MCTA	DSC	Acc	IOU	AUC	clDice	DSC_{thin}	Acc_{thin}	IOU_{thin}
GIO Backbone			0.7593	0.9571	0.6003	0.9819	0.7079	0.7019	0.9612	0.5418
✗	✗	✗	0.7702	0.9644	0.6276	**0.9839**	0.7068	0.7168	0.9670	0.5602
✓	✗	✗	0.7748	0.9633	0.6335	0.9834	0.7266	0.7285	0.9665	0.5740
✓	✓	✗	0.7755	0.9634	0.6342	0.9834	0.7277	0.7300	0.9665	0.5759
✓	✓	✓	**0.7816**	**0.9655**	**0.6428**	0.9836	**0.7330**	**0.7360**	**0.9686**	**0.5834**

4 Conclusion

We propose TemSAM, a novel framework for cerebrovascular segmentation in DSA sequences that fully leverages the MIP's global structural information and complementary frames' temporal information. By introducing a multi-level MIP global prompt and a complementary information fusion module, TemSAM adaptively refines segmentation through structural priors encoded from the MIP image while aggregating contextually complementary information to enhance feature extraction. Experimental results demonstrate that TemSAM significantly outperforms existing methods.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China under Grant 62472184, and in part by the Fundamental Research Funds for the Central Universities.

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

References

1. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
2. Chng, S.M., Alvarez, H., Rodesch, G., Lasjaunias, P.: ARTERIOVENOUS MALFORMATIONS OF THE BRAIN AND SPINAL CORD, p. 595–608 (Jan 2002). https://doi.org/10.1016/b978-0-323-03354-1.50048-1, `https://doi.org/10.1016/b978-0-323-03354-1.50048-1`
3. Fang, Z., Liu, Y., Wu, H., Qin, J.: Vp-sam: Taming segment anything model for video polyp segmentation via disentanglement and spatio-temporal side network. In: European Conference on Computer Vision. pp. 367–383. Springer (2024)
4. Heiss, W.D., Forsting, M., Diener, H.C.: Imaging in cerebrovascular disease. Current Opinion in Neurology **14**(1), 67–75 (2001)
5. Kim, S., Kim, K., Hu, J., Chen, C., Lyu, Z., Hui, R., Kim, S., Liu, Z., Zhong, A., Li, X., et al.: Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation. arXiv preprint arXiv:2309.13539 (2023)
6. 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)
7. Liu, W., Tian, T., Wang, L., Xu, W., Li, L., Li, H., Zhao, W., Tian, S., Pan, X., Deng, Y., et al.: Dias: a dataset and benchmark for intracranial artery segmentation in dsa sequences. Medical Image Analysis p. 103247 (2024)
8. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
9. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022)
10. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
11. 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)

12. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16560–16569 (2021)
13. Su, R., van der Sluijs, P.M., Chen, Y., Cornelissen, S., van den Broek, R., van Zwam, W.H., van der Lugt, A., Niessen, W.J., Ruijters, D., van Walsum, T.: Cave: Cerebral artery–vein segmentation in digital subtraction angiography. Computerized Medical Imaging and Graphics **115**, 102392 (2024)
14. Van Asperen, V., Van Den Berg, J., Lycklama, F., Marting, V., Cornelissen, S., Van Zwam, W.H., Hofmeijer, J., Van Der Lugt, A., Van Walsum, T., Van Der Sluijs, M., et al.: Automatic artery/vein classification in 2d-dsa images of stroke patients. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 12034, pp. 366–377. SPIE (2022)
15. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
16. Xie, Q., Zhang, D., Mou, L., Wang, S., Zhao, Y., Guo, M., Zhang, J.: Dsnet: A spatio-temporal consistency network for cerebrovascular segmentation in digital subtraction angiography sequences. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 199–208. Springer (2024)
17. Xu, W., Yang, H., Shi, Y., Tan, T., Liu, W., Pan, X., Deng, Y., Gao, F., Su, R.: Ernet: Edge regularization network for cerebral vessel segmentation in digital subtraction angiography images. IEEE Journal of Biomedical and Health Informatics (2023)
18. Yan, Z., Yang, X., Cheng, K.T.: Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. IEEE Transactions on Biomedical Engineering **65**(9), 1912–1923 (2018)
19. Zhang, J., Xie, Y., Wang, Y., Xia, Y.: Inter-slice context residual learning for 3d medical image segmentation. IEEE Transactions on Medical Imaging **40**(2), 661–672 (2020)
20. Zhang, J., Xie, Q., Mou, L., Zhang, D., Chen, D., Shan, C., Zhao, Y., Su, R., Guo, M.: Dsca: A digital subtraction angiography sequence dataset and spatio-temporal model for cerebral artery segmentation. IEEE Transactions on Medical Imaging (2025)
21. Zhang, M., Zhang, C., Wu, X., Cao, X., Young, G.S., Chen, H., Xu, X.: A neural network approach to segment brain blood vessels in digital subtraction angiography. Computer methods and programs in biomedicine **185**, 105159 (2020)
22. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)