

VesSAM: Efficient Multi-Prompting for Segmenting Complex Vessel

Suzhong Fu^{1,2}, Rui Sun^{1,2}, Xuan Ding^{1,2}, Jingqi Dong^{1,2}, Yiming Yang^{1,2}, Yao Zhu³, Min Chang Jordan Ren⁴, Delin Deng⁵, Angelica Aviles-Rivero⁶, Shuguang Cui^{1,2}, Zhen Li^{1,2,*}

¹FNii-Shenzhen, CUHK-Shenzhen, Shenzhen, China

²School of Science and Engineering, CUHK-Shenzhen, Shenzhen, China

³Zhejiang University, Hangzhou, China ⁴Boston University, United States

⁵Vanderbilt University, United States ⁶Tsinghua University, Beijing, China

Abstract—Precise vessel segmentation is vital for clinical applications such as diagnosis and surgical planning but remains challenging due to thin, branching geometries and low texture contrast. Although foundation models such as the Segment Anything Model (SAM) show strong performance in general segmentation tasks, they remain suboptimal for vascular structures. In this work, we present VesSAM, a powerful and efficient framework tailored for 2D vessel segmentation. VesSAM integrates three core modules: a convolutional adapter that enhances local texture features, a multi-prompt encoder that fuses anatomical cues via hierarchical cross-attention, and a lightweight mask decoder that reduces jagged artifacts. We also introduce an automated pipeline to generate structured multi-prompt annotations, and curate a diverse benchmark dataset spanning 8 datasets across 5 imaging modalities. Extensive experiments show that VesSAM surpasses state-of-the-art PEFT-based SAM variants by over 10% Dice and 13% IoU, while maintaining competitive accuracy to fully fine-tuned methods with far fewer parameters. VesSAM also generalizes well to out-of-distribution (OoD) settings, outperforming all baselines in average OoD Dice and IoU.

Index Terms—vascular segmentation, segment anything model, multi-prompts fusion, information fusion

I. INTRODUCTION

Accurate segmentation of vascular structures is vital for clinical applications such as disease diagnosis, surgical planning, and treatment monitoring. Unlike solid organs, vessels display thin, elongated geometries with complex branching patterns, posing unique challenges for automatic segmentation. Although deep learning-based methods have achieved remarkable success in vessel segmentation [1], [11]–[13], they often rely on dense pixel-wise annotations, which are expensive and time-consuming to obtain in clinical workflows.

Recently, foundation models [14]–[17] such as the Segment Anything Model (SAM) [22] have enabled prompt-based, domain-adaptable segmentation. In the medical domain, SAM-based extensions [23]–[26] have demonstrated promising results for organ and tumor segmentation. These models use generic prompts—such as boxes or sparse points—to guide segmentation, reducing annotation burden. However, their performance on vessel segmentation remains limited.

As illustrated in Fig. 1, conventional prompting strategies perform adequately in simple vascular cases but degrade in

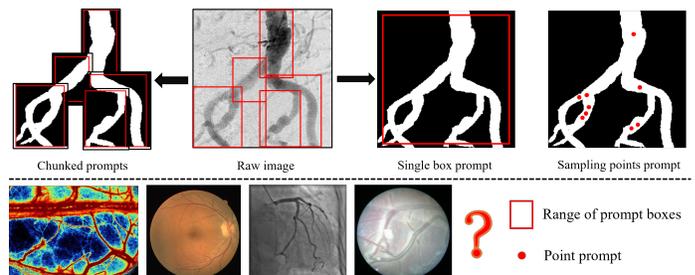


Fig. 1: Motivating examples showing the limitations of conventional prompt strategies in vessel segmentation.

complex ones. Full-image boxes offer little spatial constraint, while random point sampling within irregular masks introduces bias and uneven supervision. These methods struggle with fine, dense vascular networks characterized by intricate bifurcations and low contrast, where traditional prompts fail to capture detailed topology or maintain vessel continuity.

Beyond prompting, model architecture also constrains performance. Existing SAM-based frameworks rely heavily on ViT backbones. While transformers excel at modeling long-range dependencies, their patch-based tokenization can neglect local texture continuity—especially in non-convex, sparse structures like vessels. In organ segmentation, patch-wise attention aligns well with intra-mask continuity, but in vessel segmentation, masks often span multiple disconnected or narrow regions, undermining global modeling alone. Therefore, a successful vessel segmentation system must integrate both global context and local detail through texture-sensitive and topologically meaningful representations.

To address these issues, we propose VesSAM, a structure-aware and parameter-efficient segmentation framework tailored for vascular imaging. We construct a multi-modality vessel dataset with automatically generated prompt annotations to support structure-aware training. Code is publicly available.¹

Our key contributions are as follows:

- We propose **VesSAM**, a segmentation framework that enhances ViT-based encoders with localized texture mod-

*Corresponding author: lizhen@cuhk.edu.cn

¹<https://github.com/VersaceSu/VesSAM>

eling and anatomical prompt fusion, significantly improving accuracy on non-convex vascular structures.

- We design a **multi-prompt encoder** that unifies sparse (e.g., bifurcation points) and dense (e.g., skeletons, masks) anatomical cues using hierarchical cross-attention and graph-based topology reasoning.
- We develop an automated prompt generation pipeline and release a **benchmark dataset spanning eight vascular datasets across five imaging modalities**, supporting scalable, high-fidelity supervision.

II. EXPERIMENTS SETTING

A. Datasets

We construct a comprehensive vessel benchmark by aggregating **eight datasets across five imaging modalities**, each annotated with precise vessel delineations. Specifically, the dataset includes pelvic-iliac artery angiograms (Aorta) [4], coronary XCAD [5], retinal datasets (ARIA [6], DRIVE [7], HRF [8], IOSTAR [9]), placental vessel images (PSVFM) [10], and laser speckle contrast imaging (LSCI) [11].

Multi-Prompt Strategy. We employ a multi-prompt strategy that designs three anatomically grounded prompt types to guide segmentation: **Bifurcation points**: minimal key-points to resolve topological ambiguity; **Segment midpoints**: orientation-aware anchors that reduce redundancy; **Skeleton maps**: structural templates that preserve global vessel continuity. To systematically extract these prompts, we develop an automatic prompt generation algorithm, which processes vessel masks to generate informative prompt sets.

B. Experiments Setting

Baseline Methods. To ensure a fair comparison with existing methods, we benchmark VesSAM against five representative baselines. Among them, SAM-Med2D [25], MedSAM [23], and SAMed [24] represent parameter-efficient fine-tuning (PEFT) extensions of the Segment Anything Model (SAM) and are adapted to the vascular domain with lightweight adapters or LoRA modules. We also include two fully fine-tuned state-of-the-art methods: nnUNet [1] and TransUNet [2], which serve as strong supervised baselines in medical image segmentation. All methods are fine-tuned on the same training splits and evaluated under identical settings.

Ablation Studies. To analyze the contribution of each component in VesSAM, we conducted comprehensive ablation studies. We evaluated the impact of different prompt combinations by testing six configurations: using only bifurcation points, only segment midpoints, only skeleton maps, bifurcation+midpoints, bifurcation+skeletons, and all three combined. These experiments were performed under 512×512 input resolutions, as shown in Fig. 4.

III. METHOD

VesSAM addresses vessel segmentation challenges with a prompt-aware and texture-sensitive design. As shown in Fig. 2, it consists of three components: a convolution-enhanced

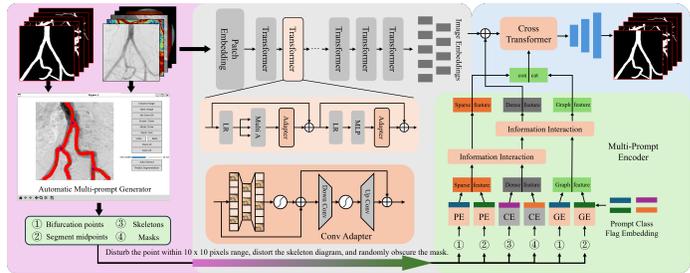


Fig. 2: Overview of the proposed VesSAM framework.

image encoder, a multi-prompt encoder integrating diverse anatomical cues, and a lightweight mask decoder for fine-grained segmentation.

Image Encoder with Convolutional Adapter. Although SAM employs a ViT-based [3] encoder to capture long-range dependencies, it lacks sensitivity to the local continuity and texture crucial for thin, non-convex vessels. Unlike organs with coherent and convex structures, vessels extend across spatially separated patches, requiring stronger modeling of local cues beyond patch boundaries.

To address this, we introduce a convolutional adapter that complements the ViT encoder. It employs depth-wise separable convolutions for efficient channel attention and a lightweight spatial attention branch using a downsampling–upsampling sequence to capture coarse patterns and refine details. By combining channel and spatial attention, the adapter enhances ViT’s global modeling with local structural awareness, improving performance on fine vessel structures with minimal parameter overhead.

Multi-Prompt Encoder. VesSAM leverages rich anatomical cues through multi-type prompts that include sparse point sets (bifurcations and midpoints), dense maps (skeleton and mask), and topological relationships. These prompts are processed through a dedicated encoder designed to preserve their distinct spatial and semantic roles while allowing effective feature fusion.

For sparse prompts, each point set is first embedded with a learnable coordinate encoder and tagged with a prompt-type indicator (e.g., bifurcation or midpoint). The resulting representations are concatenated and passed through a lightweight convolutional block to yield sparse features. For dense prompts, both the skeleton map and mask are independently encoded through shallow convolutional encoders, producing dense features. To capture global vessel topology, we additionally construct a graph using the bifurcation and midpoint prompts as nodes, and apply graph convolution to encode their relational structure, resulting in graph features with topological knowledge.

These three sources of information—sparse features (**SF**), dense features (**DF**), and graph features (**GF**)—are then integrated via a two-stage cross-attention mechanism. In the first stage, sparse and dense features are fused.

$$\mathbf{SF}', \mathbf{DF}' = \text{CrossAttention}(\mathbf{SF}, \mathbf{DF}) \quad (1)$$

TABLE I: Segmentation Performance on 512*512 (H) resolution dataset. The best results are in bold with brown background, and the second-best are underlined.

Dataset	VesSAM		PEFT Methods		SAMED		MedSAM		Full Fine-tuning Methods		TransUNet	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
LSCI	77.10	87.01	61.49	75.73	55.76	70.97	74.15	84.88	82.66	90.43	81.26	89.16
Placenta	74.38	84.61	56.35	70.58	71.74	82.55	73.86	84.22	69.50	81.03	68.10	79.67
Retinal	70.10	82.32	24.29	38.91	41.57	58.50	<u>66.03</u>	<u>79.16</u>	60.86	75.58	59.05	74.12
Aorta	92.33	95.99	84.02	91.26	93.25	96.50	93.06	96.39	93.82	96.78	93.49	96.61
XCAD	85.54	92.14	47.05	63.62	70.74	82.78	<u>82.42</u>	<u>89.72</u>	71.42	83.23	68.82	81.41
ALL	78.66	87.36	51.93	64.45	54.01	62.52	76.39	86.08	78.14	86.70	77.19	86.21
Average(H)	79.89	88.41	54.64	68.02	66.61	78.26	<u>77.90</u>	<u>86.87</u>	75.65	85.41	74.14	84.19

TABLE II: Performance under OoD setting. The best results are in bold with brown background, and the second-best are underlined. Labeled in the table for testing, and the other four for training.

Dataset	VesSAM		PEFT Methods		SAMED		MedSAM		Full Fine-tuning Methods		TransUNet	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
LSCI	24.27	38.02	8.25	14.48	26.28	40.40	24.77	39.05	<u>27.73</u>	<u>40.99</u>	38.14	51.87
retinal	50.25	66.76	19.65	32.71	16.27	27.09	10.30	17.89	24.77	38.45	32.30	48.45
Placenta	23.94	36.99	10.93	17.59	<u>22.20</u>	<u>31.47</u>	0.87	1.67	7.53	12.13	12.32	18.85
Aorta	80.55	89.10	53.44	68.87	82.07	89.90	34.25	49.01	61.31	75.38	72.74	84.02
XCAD	61.55	76.16	41.39	58.16	<u>54.81</u>	<u>70.34</u>	30.86	46.89	38.22	53.81	48.06	64.33
Average	48.11	61.40	26.73	38.36	40.33	51.84	20.21	30.90	31.91	44.15	<u>40.71</u>	<u>53.50</u>

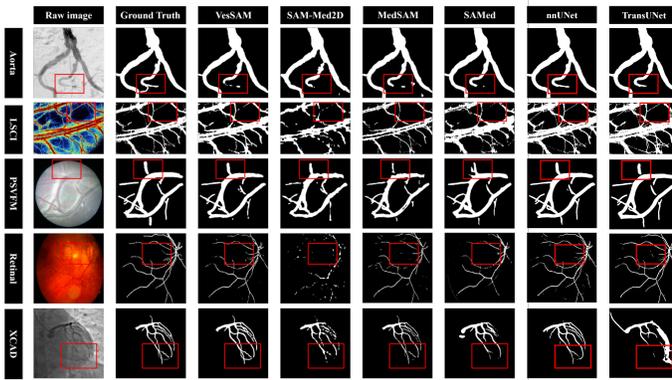


Fig. 3: Visualization of segmentation results from different methods.

In the second stage, the updated dense features interact with the graph features to yield topology-enhanced representations:

$$GF', DF'' = \text{CrossAttention}(DF', GF) \quad (2)$$

The outputs SF' , DF'' , and GF' are then forwarded to the mask decoder for final prediction.

Lightweight Mask Decoder. The decoder module integrates features from the image encoder and the multi-prompt encoder to produce the final segmentation mask. First, global image tokens extracted by the ViT encoder are concatenated with the prompt features SF' , DF'' , and GF' . Second, the combined sequences processed by a transformer-based attention module for cross-modal fusion.

IV. EXPERIMENTS RESULTS

Performance under In-Distribution (ID) Settings. We evaluated at 512×512 resolution datasets, VesSAM exhibits

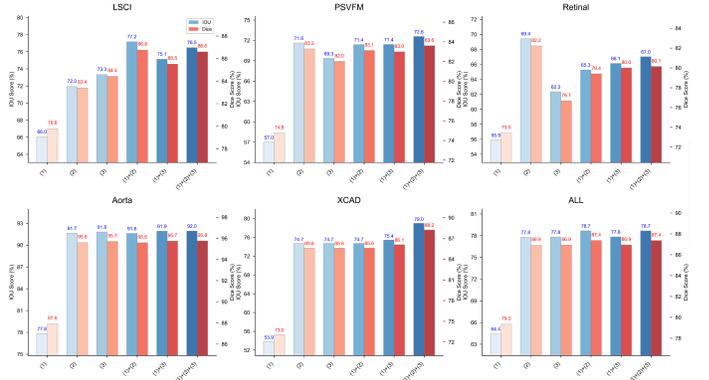


Fig. 4: Ablation on different prompt combinations.

even stronger performance. As shown in Table I, the model not only outperforms SAMed by more than 10% in Dice and 13% in IoU, but also surpasses nnUNet by 3% in Dice score. The advantage becomes especially pronounced on datasets such as LSCI and Retina, which require fine-grained boundary preservation and vascular continuity. The multi-prompt mechanism is particularly effective in suppressing noise, preserving thin structures, and improving segmentation robustness. VesSAM also yields strong results on PSVFM and XCAD datasets, highlighting its ability to handle blurry boundaries and small-scale vessel features.

Performance under Out-of-Distribution (OoD) Settings.

To evaluate cross-domain generalization, we conduct experiments under the out-of-distribution (OoD) setting by holding out one modality as the test domain while training on the remaining ones. As shown in Table II, VesSAM achieves an average Dice score of 61.4% and an IoU of 48.11%, marking an improvement of 9.56% and 9.78% over SAMed,

respectively. Compared to other methods, VesSAM exhibits greater resilience on challenging datasets such as Retinal and XCAD, where complex vascular structures and terminal branches require detailed supervision. In contrast, SAM-Med2D, SAMed, nnUNet, and MedSAM suffer substantial performance drops on these datasets, particularly in scenarios involving small vessels and boundary ambiguity. These results underscore the importance of structure-aware prompting in achieving robust generalization across heterogeneous imaging domains.

Visualization Analysis. Qualitative comparisons further highlight VesSAM’s strengths. As illustrated in Fig. 3, across five modalities, VesSAM consistently produces smoother and more anatomically accurate vessel masks. For instance, in the Aorta dataset, where noise artifacts resemble vessel endpoints, MedSAM is easily prone to false positives, while VesSAM leverages keypoint prompts to preserve structure and reject irrelevant features. TransUNet and nnUNet, although competitive, often miss thin branches or introduce artifacts in low-contrast regions. VesSAM also outperforms others on the LSCI dataset through effective suppression of background noise while preserving fine vascular continuity. On PSVFM and XCAD, where boundary quality is compromised, VesSAM benefits from stronger prompt encoding and convolutional refinement, yielding sharper and more complete segmentation. In contrast to the jagged artifacts observed in SAM-Med2D outputs, VesSAM’s mask decoder produces more consistent and artifact-free results.

ACKNOWLEDGMENT

The work was supported in part by NSFC with Grant No. 12326610&62573371, by Guangdong S&T Programme with Grant No. 2024B0101030002, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao ShenzhenHK S&T Cooperation Zone, by the Shenzhen General Program No. JCYJ2022053014360001, by the Shenzhen Outstanding Talents Training Fund 202002, by the Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of BigData Computing CHUKShenzhen, by the NSFC 61931024&62293482, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027), by China Association for Science and Technology Youth Care Program, by the Shenzhen-Hong Kong Joint Funding No.SGDGX20211123112401002, and by Tencent & Huawei Open Fund.

REFERENCES

- [1] F. Isensee, P. F. Jaeger, S. A. A. Kohl et al., “nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [2] J. Chen, Y. Lu, Q. Yu et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” in *arXiv preprint arXiv:2102.04306*, 2021.

- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *arXiv preprint arXiv:2010.11929*, 2020.
- [4] V. Zohranyan, M. Haghghi, S. Wang et al., “Dr-SAM: An End-to-End Framework for Vascular Segmentation, Diameter Estimation and Anomaly Detection on Angiography Images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5113–5121.
- [5] Y. Ma, M. Liu, S. Huang et al., “Self-supervised vessel segmentation via adversarial learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7536–7545.
- [6] D. J. J. Farnell, A. Papageorgiou, A. S. French et al., “Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators,” *J. Franklin Inst.*, vol. 345, no. 7, pp. 748–765, 2008.
- [7] J. Staal, M. D. Abramoff, M. Niemeijer et al., “Ridge-based vessel segmentation in color images of the retina,” *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, 2004.
- [8] T. Köhler, F. Maier, A. Hornegger et al., “Automatic no-reference quality assessment for retinal fundus images using vessel segmentation,” in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst. (CBMS)*, 2013, pp. 1–6.
- [9] J. Zhang, S. Dashtbozorg, G. P. Looman et al., “Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores,” *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2631–2644, 2016.
- [10] S. Bano, F. Vasconcelos, L. M. Shepherd et al., “Deep placental vessel segmentation for fetoscopic mosaicking,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, vol. 12263, pp. 763–772.
- [11] S. Fu, J. Xu, S. Chang et al., “Robust vascular segmentation for raw complex images of laser speckle contrast based on weakly supervised learning,” *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 39–50, 2023.
- [12] S. Yao, H. Wu, S. Fu et al., “High-performance laser speckle contrast image vascular segmentation without delicate pseudo-label reliance,” *J. Innov. Opt. Health Sci.*, vol. 18, no. 1, 2025.
- [13] K. Yang, S. Chang, J. Yuan et al., “Robust vascular segmentation in laser speckle contrast images based on semi-weakly supervised learning,” *Phys. Med. Biol.*, vol. 68, no. 14, Art. no. 145008, 2023.
- [14] H. Touvron, T. Lavril, G. Izacard et al., “LLaMA: Open and efficient foundation language models,” in *arXiv preprint arXiv:2302.13971*, 2023.
- [15] J. Achiam, S. Adler, S. Agarwal et al., “GPT-4 technical report,” in *arXiv preprint arXiv:2303.08774*, 2023.
- [16] GLM Team, A. Zeng, B. Xu et al., “ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools,” in *arXiv preprint arXiv:2406.12793*, 2024.
- [17] S. Liu, Z. Zeng, T. Ren et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.
- [18] Z. Yang, L. Li, K. Lin et al., “The dawn of LMMs: Preliminary explorations with GPT-4V(ision),” in *arXiv preprint arXiv:2309.17421*, 2023.
- [19] A. Ramesh, M. Pavlov, G. Goh et al., “Zero-shot text-to-image generation,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
- [20] J. Li, D. Li, C. Xiong et al., “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.
- [21] R. Rombach, A. Blattmann, D. Lorenz et al., “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.
- [22] A. Kirillov, E. Mintun, N. Ravi et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.
- [23] J. Ma, Y. He, F. Li et al., “Segment anything in medical images,” *Nat. Commun.*, vol. 15, no. 1, Art. no. 654, 2024.
- [24] K. Zhang and D. Liu, “Customized segment anything model for medical image segmentation,” in *arXiv preprint arXiv:2304.13785*, 2023.
- [25] J. Cheng, J. Ye, Z. Deng et al., “SAM-Med2D,” in *arXiv preprint arXiv:2308.16184*, 2023.
- [26] H. Wang, S. Guo, J. Ye et al., “SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images,” in *European Conference on Computer Vision (ECCV)*, 2024.