Dual-level Adapter Boosting Prompt-free Curvilinear Structure Segmentation

Kai Zhu[1,2] Li Chen[1,2,✉] Jun Cheng[3]
[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan, China
[3]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
kylechuzk@gmail.com chenli@wust.edu.cn juncheng@ieee.org

Abstract

Curvilinear structure segmentation is essential in domains such as medical imaging, remote sensing, and materials science. Existing methods often require extensive domain-specific training and lack generalization to novel domains. To overcome these limitations, we propose the Segment Anything Curve Model (SACM) — a universal framework for curvilinear structure segmentation built upon the pretrained Segment Anything Model (SAM). SACM introduces a dual-level adapter architecture that enables both fine-grained local adaptation and robust cross-domain generalization: block-level internal adapters refine local structural representations, while external adapters facilitate cross-domain feature alignment. Specifically, the internal adapters are embedded within each Transformer block to locally adapt and refine features for thin and intricate curvilinear patterns, while the external adapters operate across blocks to capture global, multi-layer contextual information and facilitate domain adaptation. Furthermore, SACM introduces a feature fusion mechanism that aggregates multi-layer features from all external adapters via a feed-forward network module, and a dual-stage refinement process in the mask decoder to enhance topology and connectivity. This design enables prompt-free, data-efficient fine-tuning and achieves robust cross-domain generalization when trained with only 18 annotated images. Extensive experiments across twelve diverse curvilinear datasets validate that SACM achieves state-of-the-art performance. The code is available at https://github.com/kylechuuuuu/SACM.

1. Introduction

Curvilinear structures, such as vascular networks, neural pathways, and road systems, form the essential skeletons of objects across a vast spectrum of scientific and engineering domains [17]. Their accurate and automated segmentation is not merely a technical task but a critical enabler for fundamental applications, from quantifying neuronal degradation in neurodegenerative diseases [5] and monitoring infrastructural integrity [39] to analyzing root system architectures for crop improvement [13]. However, the manual delineation of these intricate patterns is notoriously laborious and unscalable, creating a significant bottleneck that impedes large-scale quantitative analysis and the deployment of high-throughput systems. This underscores a compelling and persistent need for automated segmentation methods that are not only precise but also robust and generalizable across diverse real-world scenarios.

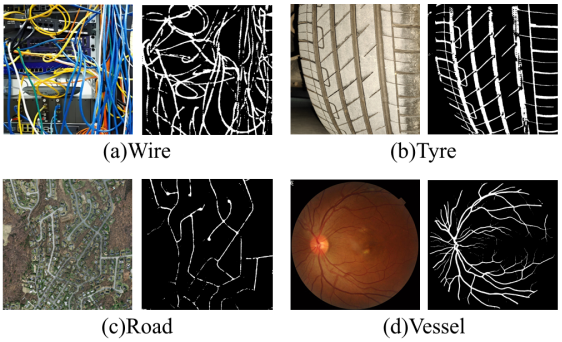


Figure 1. Segmentation results across four diverse domains using SACM. These results demonstrate the model's superior generalization in extracting complex curvilinear structures, including wire networks, tyre tread, remote sensing roads, and retinal vessels.

Curvilinear **S**tructure **S**egmentation (CSS) presents a formidable challenge rooted in a fundamental dichotomy: the need to reconcile local, fine-grained feature fidelity with global, long-range topological coherence. On one hand, these structures are defined by subtle local characteristics, such as faint boundaries, heterogeneous intensity profiles, and rapidly varying widths, which are often corrupted by noise or occluded by surrounding tissues [25, 28]. This demands a model with high sensitivity to low-level details.

✉ Corresponding author.

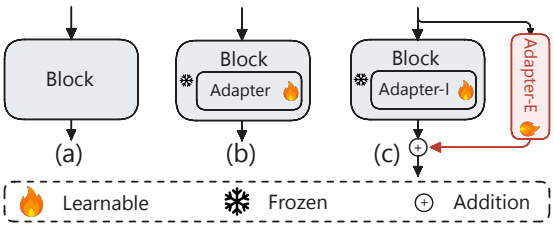


Figure 2. Structure of different SAM-based adapter mechanisms: (a) ViT Block in SAM without adaptation, (b) Medical Image Adapter [4] with internal-only adaptation, and (c) Our dual-level adapter architecture with both block internal and external adapter.

On the other hand, these structures are defined by their continuity, elongation, and branching topology, necessitating the preservation of global structural connectivity across extended regions. Many established CSS methods, from classic filter-based approaches to modern deep convolutional networks [15, 22, 43], often excel at one aspect at the expense of another. They tend to produce fragmented segments in low-contrast regions or lose fine details, thereby failing to maintain topological integrity. Furthermore, their heavy reliance on large, domain-specific annotated datasets severely limits their cross-domain generalization, as models trained on retinal vessels, for example, typically fail to segment satellite road imagery, hindering their practical utility.

The recent advent of large-scale foundation models, particularly the SAM [12], has marked a paradigm shift in computer vision, demonstrating unprecedented zero-shot generalization capabilities. However, these models were primarily designed for generic object segmentation and harbor intrinsic architectural biases that render them suboptimal for the unique geometric properties of curvilinear structures. A direct application is fundamentally challenged: (1) The prompting mechanism is ill-posed for CSS. The dense, tortuous, and interconnected nature of curvilinear networks makes discrete point- or box-based prompting impractical and inefficient for achieving complete segmentation. (2) Existing adaptation strategies are insufficient. Current fine-tuning or adapter-based methods [3, 4, 18, 24] typically insert lightweight modules into the MLP layers of the Transformer blocks (as shown in Fig. 2(b)). This single-level adaptation primarily refines block-level local features, while neglecting the global structural modeling mechanism itself. Consequently, it fails to explicitly enhance the model's capacity for capturing long-range spatial dependencies that are essential for maintaining the continuity of curved global structures. This architectural limitation restricts their ability to effectively model global topology and transfer structural knowledge across disparate domains.

To address the limitations of adapting generic vision foundation models for CSS, we propose SACM, a universal framework built upon the pre-trained SAM. Fig. 1 demonstrates SACM's superior segmentation results across diverse domains. Our approach introduces a novel Dual-Level Adapter (DLAda) architecture (as shown in Fig. 2(c)), designed to enhance both fine-grained local adaptation and robust global context refinement. Complementing this, we develop a Prompt-Free Adapter-Fusion Decoder (PFAF-D) that integrates multi-layer features and employs a dual-stage refinement process to generate precise and topologically coherent segmentation masks. This synergistic design forms our contributions, which are summarized as follows:

- We propose SACM, a universal, prompt-free framework for curvilinear structure segmentation built on the pre-trained SAM, enabling direct mask prediction without user prompts, requiring only a few-shot training set for fine-tuning, and robust cross-domain generalization.
- We introduce a dual-level adapter architecture comprising: (1) block-internal adapters in each Transformer MLP path for fine-grained local feature tuning, and (2) block-external adapters between Transformer blocks for global, multi-layer feature injection. (3) An Adapter Fusion module aggregates multi-layer features from all external adapters and injects them into the mask decoder to enable prompt-free segmentation. The decoder then employs a two-stage refinement process to enhance segmentation accuracy and topological connectivity.
- Extensive experiments demonstrate that SACM consistently outperforms state-of-the-art baselines across twelve curvilinear datasets, achieving strong cross-domain generalization with only few-shot training data, highlighting its practical utility for data-scarce scenarios.

2. Related Work

2.1. Curvilinear Structure Segmentation

Curvilinear segmentation targets thin, elongated structures. U-Net [25] established the baseline; subsequent CNN improvements include residual blocks [1], multi-scale aggregation [14], attention modules [27], and deformable or dilated convolutions [47]. However, CNNs often produce fragmented outputs and rely on costly post-processing. Recent architectures emphasize global context and long-range dependencies: hybrid Transformer designs (e.g., TSNet [46]), state-space models with edge enhancement (MambaVesselNet [16]), semi-supervised Transformer methods for low-data regimes (EXP-Net [26]), and context-distillation decoders (FCoDT-Net [42]).

2.2. SAM-based Semantic Segmentation

Foundation models shifted segmentation toward prompt-driven, general-purpose frameworks. The SAM [12] enables zero-shot, promptable masks and has inspired three adaptation paths: (1) domain transfer via fine-tuning or adapters (SAM-Med2D [4], CWSAM [24]); (2) prompt and output refinement to improve boundary quality (HQ-

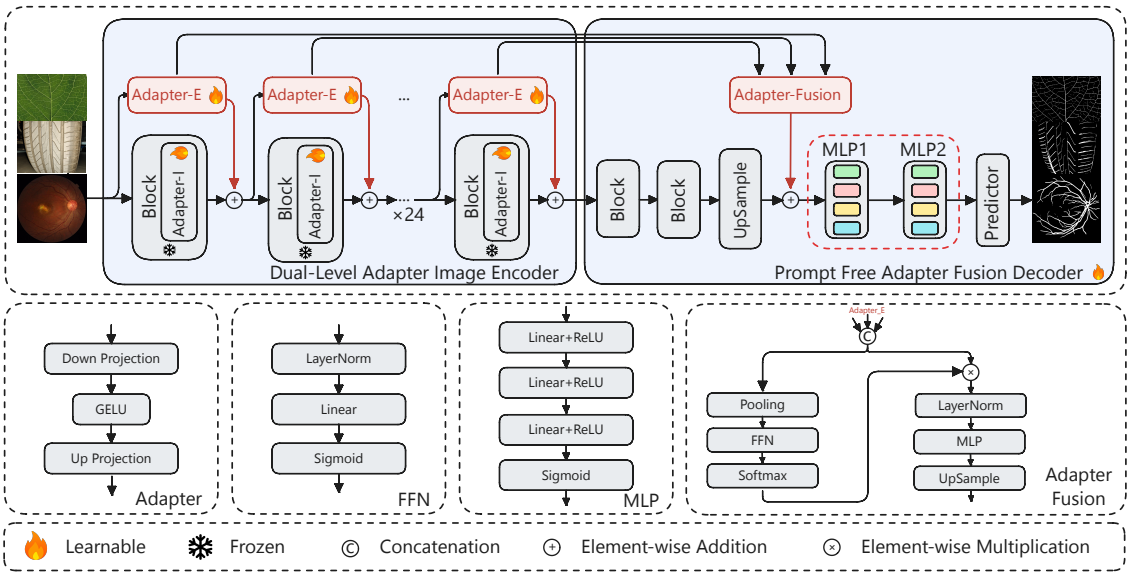


Figure 3. Overview of the SACM framework. SACM leverages the pretrained SAM as the foundation model, enhanced with dual-level adapter fine-tuning. Multi-layer features from external adapters are aggregated and integrated into the mask decoder via a dual-stage refinement mechanism to improve curvilinear structure segmentation.

SAM [11]); and (3) structure-aware prompting for geometrically constrained targets (e.g., SAM-OCTA [33]).

2.3. Adapter-Based Fine-tuning

Adapters provide parameter-efficient adaptation by freezing most pretrained weights and training small modules [3, 24, 37]. Originating in NLP [6, 8], adapters are effective across vision tasks [7, 34, 38] and help reduce memory and compute during transfer [32, 44]. However, by inserting adapters only locally within Transformer blocks, these methods lack a mechanism for global, multi-scale feature aggregation. This architectural limitation makes them inherently unsuitable for curvilinear segmentation, which demands long-range spatial dependency modeling to maintain topological continuity.

3. Methodology

This section provides a detailed description of SACM, our prompt-free framework for CSS, built upon a frozen SAM image encoder. SACM's architecture is centered around two key components: (1) DLAda for both local and global feature adaptation within the encoder, and (2) PFAF-D for multi-layer features fusion and dual-stage mask refinement. The overall pipeline is illustrated in Fig. 3, with each component elaborated in the subsequent subsections.

3.1. Dual-Level Adapter Architecture

The DLAda architecture is the core component enabling parameter-efficient fine-tuning and robust domain generalization for curvilinear structures. As demonstrated in Fig. 3, it is strategically integrated into every Transformer block across the depth of the frozen SAM image encoder. This design achieves feature adaptation across multiple granularities while effectively preserving the knowledge encoded within the pretrained foundation model backbone. The DLAda architecture comprises two distinct, yet complementary, components that collaboratively address the challenge of fine-grained local feature discrimination and global topological coherence: the block-internal adapter (Adapter-I) and the block-external adapter (Adapter-E).

3.1.1. Block-Internal Adapter

The block-internal adapter is designed for fine-grained local feature adaptation within the transformer block, focusing on refining feature representations crucial for distinguishing thin structures from the complex background. Following established parameter-efficient fine-tuning methodologies, Adapter-I is embedded within the residual pathway of the MLP sub-module of the l-th Transformer block.

Concretely, let $\mathbf{X}^{(l)} \in \mathbb{R}^{N\times D}$ be the input sequence to block l (where N is the sequence length and D the embedding dimension). The internal adapter is a token-wise bottleneck module:

$$\text{Adapter-}I(\mathbf{X}) \;=\; \mathcal{G}\big(\mathbf{X}\mathbf{W}^{I}_{\downarrow}\big)\,\mathbf{W}^{I}_{\uparrow}, \qquad (1)$$

where $\mathbf{W}^{I}_{\downarrow} \in \mathbb{R}^{D\times r}$ and $\mathbf{W}^{I}_{\uparrow} \in \mathbb{R}^{r\times D}$ are the down-/up-projection matrices, r is the bottleneck ratio, and $\mathcal{G}$ denotes GELU activation. This per-token design keeps the adapter's Jacobian block-diagonal across tokens, concentrating updates on channel-wise, local refinements.

The MLP output augmented by the internal adapter is incorporated via the residual connection:

$$\mathbf{H}_{\text{out}}^{I} \;=\; \text{MLP}\big(\text{LN}(\mathbf{Y})\big) \;+\; \mathbf{Y} + \;\text{Adapter-}I\big(\text{LN}(\mathbf{Y})\big), \quad (2)$$

where $\mathbf{Y}$ is the feature sequence after the MSA, and LN denotes LayerNorm. Embedding Adapter-I in the MLP path emphasizes local discriminative features (e.g., thin edges, fine vessel details) while preserving the pretrained transformer's global structure.

3.1.2. Block-External Adapter

Unlike conventional adapter approaches that operate solely within individual Transformer blocks, our block external adapter operates at the block level, creating direct pathways for cross-layer feature fusion and hierarchical representation learning. Positioned in the residual connection around the entire Transformer block, it facilitates global contextual propagation. The external adapter processes the normalized block output:

$$\mathbf{X}^{(l+1)} \;=\; F_l\big(\mathbf{X}^{(l)}\big) \;+\; \text{Adapter-}E\Big(\text{LN}\big(F_l(\mathbf{X}^{(l)})\big)\Big), \quad (3)$$

where F_l denotes the Transformer block at layer l, with the same bottleneck form but independent parameters:

$$\text{Adapter-}E(\mathbf{Y}) \;=\; \mathcal{G}\big(\mathbf{Y}\mathbf{W}_{\downarrow}^{E}\big)\,\mathbf{W}_{\uparrow}^{E}, \quad (4)$$

where $\mathbf{W}_{\downarrow}^{E} \in \mathbb{R}^{D\times r}$ and $\mathbf{W}_{\uparrow}^{E} \in \mathbb{R}^{r\times D}$.

Because Adapter-E receives features after attention-based token mixing in F_l, its perturbation is naturally propagated through subsequent attention layers. Unlike Adapter-I which refines tokens locally, Adapter-E injects global feature updates at the block boundary. Through successive layers, this effect compounds: each deeper layer increases cross-token interaction density, enabling Adapter-E's structural cues to influence distant tokens and capture long-range dependencies. This hierarchical propagation enables Adapter-E to inject global, multi-layer context for preserving long-range continuity in curvilinear structures.

This mechanism is validated by: (i) Grad-CAM visualizations showing Adapter-E captures broad vessel structures globally, while Adapter-I focuses on local details; and (ii) ablation studies demonstrating that combining both adapters yields synergistic gains, confirming their complementary roles in multi-layer feature adaptation.

3.2. Prompt-Free Adapter Fusion Decoder

To overcome the limitations of prompt-driven segmentation for thin, branching structures, SACM introduces a prompt-free mask decoder with two key components: (1) a prompt-free feature-fusion stream that aggregates multi-layer outputs from external adapters to provide global structural priors, and (2) a dual-stage refinement module that ensures both local boundary precision and global topological coherence. The following sections detail these components.

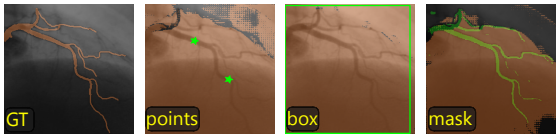


Figure 4. The orange regions indicate the SAM (ViT-L) output mask, while the green regions represent the input prompts. The visual evidence suggests that the segmentation performance across all three prompt modalities (points, box, and mask) is suboptimal for curvilinear structure segmentation.

3.2.1. Adapter Fusion for Prompt-Free Segmentation

Prompt-based segmentation (points, boxes, and masks) is inherently ill-suited for CSS due to three fundamental mismatches (as shown in Fig. 4): (i) *positional bias*—sparse prompts introduce localized cues that blur fine boundaries and disrupt thin structure continuity; (ii) *scale mismatch*—fixed prompts struggle to encode multi-scale information needed for both thin vessels and complex junctions; (iii) *topology agnostic*—prompts provide no global continuity guidance for maintaining branching topology.

SACM operates in a fully prompt-free manner by replacing interactive prompts with learned, feature-level structural guidance. The Adapter Fusion module aggregates multi-layer outputs from external adapters, which inherently capture cross-layer, attention-mixed context, and injects this fused descriptor into the mask decoder. External adapters encode long-range dependencies across encoder layers, naturally capturing thin, elongated geometry and branching structures, while internal adapters perform token-wise refinements. The fused multi-layer descriptor provides global feature priors, addressing CSS challenges and enabling automatic segmentation of curvilinear structures.

Formally, let $\{\mathcal{E}_1, \ldots, \mathcal{E}_L\}$ denote the external adapter outputs from L encoder layers, where each $\mathcal{E}_l$ corresponds to the output at layer l. We compute layer-wise descriptors via average pooling:

$$\mathbf{z}_l = \mathcal{A}(\mathcal{E}_l) \quad (l = 1, \ldots, L), \quad (5)$$

where $\mathcal{A}$ denotes average pooling.

Since different encoder layers contribute unequally to CSS, we then learn adaptive weights $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]$ that reflect each layer's contribution:

$$\boldsymbol{\alpha} = \text{Softmax}\left(\text{FFN}\left(\text{Concat}(\mathbf{z}_1, \ldots, \mathbf{z}_L)\right)\right), \quad (6)$$

where FFN denotes a feed-forward network.

The weighted fusion aggregates multi-layer structural information:

$$\mathcal{F}_{\text{fusion}} = \text{UP}\left(\text{MLP}\left(\sum_{l=1}^{L} \alpha_l \cdot \mathcal{E}_l\right)\right), \quad (7)$$

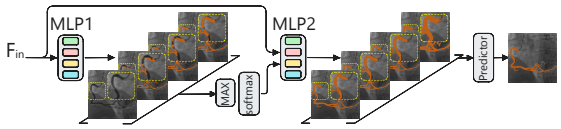


Figure 5. Dual-stage mask refinement: Stage-1 generates coarse descriptors and ranks heads by confidence scores; Stage-2 refines masks with reordered descriptors, balancing boundary precision and topological consistency.

where UP denotes upsample operations. The fused representation is then injected into the decoder via a residual connection:

$$\mathbf{F}^{\text{out}}_{\text{decoder}} = \mathbf{F}^{\text{in}}_{\text{decoder}} + \mathcal{F}_{\text{fusion}}. \tag{8}$$

3.2.2. Dual-stage Refinement

Single-pass mask decoding often produces locally plausible but globally inconsistent predictions for CSS, exhibiting topological inconsistencies such as discontinuous vessels or spurious branches. To address this, we introduce a two-stage refinement module that separates local boundary precision from global topological coherence. The structure of the dual-stage refinement module is shown in Fig. 5.

Let $\mathbf{U}$ denote the fused feature map from the decoder upsampling path. The module employs two identical-architecture MLP networks: MLP_1 for coarse mask generation and MLP_2 for refined mask prediction. In the first stage, coarse masks $\mathbf{M}^{(1)}$ are generated to assess global topological coherence:

$$\mathbf{M}^{(1)} = \text{MLP}_1(\mathbf{U}), \quad \mathbf{w} = \text{Softmax}\left(\mathcal{M}(\mathbf{M}^{(1)})\right), \tag{9}$$

where $\mathcal{M}$ denotes the max pooling and $\mathbf{w}$ represents confidence weights. Heads are ranked by their maximum spatial activation strength: $\mathbf{s} = \text{argsort}(\mathbf{w},\ \text{descending})$.

In the second stage, refined masks $\mathbf{M}^{(2)}$ are generated using the same feature map but conditioned on the ranking:

$$\mathbf{M}^{(2)} = \text{MLP}_2(\mathbf{U}, \mathbf{s}), \tag{10}$$

An MLP-based IoU predictor selects the optimal mask from the ordered candidates. This two-stage design ensures topologically sound heads dominate final predictions while maintaining boundary precision, producing segmentation that balances sharp local boundaries with preserved vessel connectivity.

3.3. Loss Function and Evaluation Metrics

The loss function is formulated as a weighted sum of Binary Cross-Entropy (BCE) loss and Dice loss to achieve optimal segmentation performance:

$$\mathcal{L}_{SACM} = \mathcal{L}_{BCE} + \lambda \cdot \mathcal{L}_{Dice}, \tag{11}$$

Split	Distribution	Dataset	Modality	Target	Samples
Train	-	DRIVE [31]	Fundus	Vessel	3
	-	CHASEDB1 [23]	Fundus	Vessel	3
	-	DCA1 [2]	X-ray	Vessel	3
	-	CrackTree [50]	RGB	Crack	3
	-	CREMI[1]	Microscopy	Boundary	3
	-	CORN [21]	Microscopy	Nerve	3
Test (Base)	Seen	DRIVE [31]	Fundus	Vessel	20
	Seen	CHASEDB1 [23]	Fundus	Vessel	8
	Seen	DCA1 [2]	X-ray	Vessel	100
	Seen	CORN [21]	Microscopy	Nerve	100
	Unseen	FIVES [10]	Fundus	Vessel	200
	Unseen	DSCA [45]	X-ray	Vessel	45
	Unseen	XCAD [19]	X-ray	Vessel	126
	Unseen	CRACK	RGB	Crack	31
Test (Novel)	Unseen	ROAD [20]	Aerial	Road line	49
	Unseen	LEAF	RGB	Venation	31
	Unseen	TYRE	RGB	Tread line	36
	Unseen	WIRE	RGB	Wire	31

Table 1. Details of the datasets for experiments.

where λ controls the trade-off between the two loss items.

We use both pixel-level and topology-aware metrics to provide a comprehensive assessment of segmentation quality. Specifically, Dice [41] and Intersection-over-Union (IoU) [35] measure pixel-wise agreement between predictions and ground truth, clDice [29] evaluates topological correctness by comparing skeletonized centerlines, and the 95th-percentile Hausdorff distance (HD95) [36] assesses boundary localization while reducing sensitivity to extreme outliers. All metrics are computed on binarized masks (threshold = 0.5); clDice is computed on skeletonized masks obtained via standard morphological thinning. For HD95, lower values indicate better boundary agreement.

4. Experiments

4.1. Datasets

To fine-tune SACM, we randomly sampled a training set of only 3-shot per dataset (18 images) from six curvilinear benchmarks: DRIVE [31], CHASEDB1 [23], DCA1 [2], CrackTree [50], CREMI[1], and CORN [21]. These datasets cover diverse curvilinear segmentation tasks, including retinal vessel extraction, neuron boundary delineation, crack detection, and nerve fiber segmentation. Modalities, targets, and case counts for all training and evaluation datasets are summarized in Table 1.

For evaluation, we partitioned the test data along two dimensions: class familiarity (Base or Novel) and data distribution (Seen or Unseen). Base refers to curvilinear structure classes that appeared in the training set, while Novel refers to entirely new classes never encountered during training. Seen indicates test images from the same datasets used in training, while Unseen indicates test images from different datasets with distinct data distributions.

The Base-Seen subset comprises test images from the training datasets (DRIVE [31], CHASEDB1 [23], DCA1 [2], and CORN [21]), evaluating in-distribution per-

[1] https://cremi.org/

Method	Prompt	DRIVE [31]				CHASEDB1 [23]				DCA1 [2]				CORN [21]			
		Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓
U-Net(2015) [25]	✗	74.66	59.81	25.19	17.48	71.53	55.89	13.27	34.06	67.41	51.77	30.40	29.78	22.46	12.88	11.81	60.92
CS^{2}Net(2021) [22]	✗	75.83	60.12	26.45	16.92	72.14	56.78	14.23	33.45	68.68	53.12	30.96	27.86	23.89	13.45	12.34	59.67
BCUNet(2023) [43]	✗	78.08	64.32	**29.40**	9.07	78.24	64.35	**17.20**	23.52	73.35	58.29	30.98	21.02	29.76	17.57	15.35	41.35
MaskVSC(2025) [48]	✗	67.47	51.06	14.37	13.56	66.39	49.76	9.22	28.07	59.78	43.02	20.21	50.11	35.16	21.46	19.13	23.64
nnWNet(2025) [49]	✗	68.12	52.34	15.67	12.89	67.45	50.23	10.45	27.34	71.92	56.47	29.84	23.49	36.23	22.67	19.78	22.45
SAM-Med2d(2023) [4]	✓	59.23	42.20	7.21	16.68	53.01	36.18	4.47	83.79	45.76	29.71	16.89	71.25	25.34	14.67	5.89	55.23
SAMOCTA(2024) [33]	✓	60.45	43.12	8.34	15.67	54.23	37.89	5.12	82.34	61.84	44.75	24.75	42.88	26.78	15.34	6.45	54.12
CWSAM(2025) [24]	✓	61.01	43.56	8.55	15.03	54.85	38.22	5.31	81.52	63.69	46.91	25.88	38.76	27.07	15.53	6.62	53.54
KnowSAM(2025) [9]	✓	58.13	41.14	7.05	17.12	52.42	35.88	4.03	84.52	43.87	28.10	15.97	75.92	24.56	13.89	5.67	56.34
SegDINO(2025) [40]	✗	61.78	44.56	9.12	14.89	55.67	39.23	6.34	81.45	48.95	32.41	18.16	66.35	27.34	16.12	7.01	53.67
Ours (SACM)	✗	**78.89**	**65.24**	29.02	**8.34**	**79.27**	**65.72**	16.66	**14.58**	**75.67**	**61.10**	**32.65**	**16.67**	**55.38**	**38.44**	**34.70**	**16.07**

Table 2. Comparison of Dice(%), IoU(%), clDice(%) and HD95 (px) on 4 base datasets. Best results are in **bold**, second best are underlined. The upper part of the table represents models without pre-trained weights, while the lower part represents models with pre-trained weights.

Method	Prompt	DSCA [45]				CRACK				XCAD [19]				FIVES [10]			
		Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓
U-Net(2015) [25]	✗	34.49	21.17	12.37	80.66	24.56	15.52	13.57	165.02	48.37	32.22	8.62	112.54	42.94	29.33	6.37	150.53
CS^{2}Net(2021) [22]	✗	35.63	23.55	13.68	80.32	19.57	12.84	12.48	120.36	55.50	40.40	9.97	88.23	37.54	25.10	8.42	168.16
BCUNet(2023) [43]	✗	30.54	18.63	12.78	125.58	25.62	15.56	13.88	93.88	21.15	12.63	6.01	98.09	56.01	41.74	9.73	128.16
MaskVSC(2025) [48]	✗	31.85	19.11	8.65	225.17	33.38	20.55	10.89	236.64	37.80	23.56	5.72	124.57	64.71	48.87	5.56	120.24
nnWNet(2025) [49]	✗	18.99	10.70	13.01	176.30	5.66	3.37	2.48	282.24	63.27	47.78	11.48	78.03	32.86	21.34	5.40	364.26
SAM-Med2d(2023) [4]	✓	36.52	24.35	10.85	123.45	5.68	4.54	14.37	160.18	25.59	16.67	5.31	102.34	10.71	8.57	2.98	120.65
SAMOCTA(2024) [33]	✓	65.65	49.46	13.96	128.01	52.30	37.40	16.83	76.23	68.58	54.86	6.68	92.34	74.51	60.94	10.44	87.26
CWSAM(2025) [24]	✓	44.31	28.86	7.75	197.47	49.21	34.23	15.94	109.14	68.32	52.24	10.33	112.92	64.95	48.65	8.78	105.13
KnowSAM(2025) [9]	✓	38.10	23.91	1.60	115.79	14.38	8.11	3.55	140.77	25.02	14.71	4.03	81.53	45.09	29.64	1.51	129.88
SegDINO(2025) [40]	✗	35.46	22.40	1.81	294.60	14.42	8.24	2.01	168.17	57.54	40.74	3.28	204.22	46.93	31.20	2.89	78.66
Ours (SACM)	✗	**68.43**	**53.51**	**15.51**	**78.22**	**63.17**	**47.80**	**19.19**	**62.03**	**74.29**	**57.74**	**13.46**	**43.33**	**75.48**	**62.26**	**12.24**	**69.47**

Table 3. Cross-dataset Performance on Four Unseen Datasets.

Method	Prompt	ROAD [20]				LEAF				TYRE				WIRE			
		Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓	Dice↑	IoU↑	clDice↑	HD95↓
U-Net(2015) [25]	✗	1.42	0.72	0.27	180.17	0.70	0.36	0.18	166.81	31.27	20.27	**16.44**	109.80	35.13	21.82	12.11	106.90
CS^{2}Net(2021) [22]	✗	2.21	1.77	0.36	150.22	0.36	0.29	0.12	192.28	24.78	16.14	13.43	129.48	26.48	17.26	11.32	112.31
BCUNet(2023) [43]	✗	1.05	0.53	0.22	143.38	4.87	2.90	0.87	180.40	32.46	21.19	14.20	88.26	29.91	18.20	12.73	93.49
MaskVSC(2025) [48]	✗	5.62	2.97	0.66	203.99	10.16	5.69	2.04	115.11	31.59	19.43	10.25	99.63	35.03	22.46	12.14	83.88
nnWNet(2025) [49]	✗	0.97	0.49	0.49	323.95	13.98	7.93	6.26	163.94	31.68	19.60	11.59	121.32	15.74	9.31	6.31	191.72
SAM-Med2d(2023) [4]	✓	0.52	0.42	0.12	340.12	2.35	1.88	0.35	150.23	12.06	9.65	1.14	120.45	5.99	4.79	2.42	145.28
SAMOCTA(2024) [33]	✓	21.62	14.11	5.24	150.22	20.43	13.37	6.43	94.12	20.08	13.16	8.34	110.23	44.20	30.37	14.67	79.53
CWSAM(2025) [24]	✓	24.79	14.53	2.99	182.50	27.85	17.01	7.94	115.80	30.77	19.19	10.43	99.04	42.61	29.55	14.65	80.42
KnowSAM(2025) [9]	✓	0.42	0.21	0.04	160.29	8.41	4.46	1.18	106.54	13.68	7.70	1.47	165.81	17.13	9.81	3.31	95.75
SegDINO(2025) [40]	✗	0.81	0.41	0.13	267.29	4.54	2.39	0.67	104.30	18.53	10.71	2.27	108.82	26.46	15.88	2.27	100.99
Ours (SACM)	✗	**40.43**	**26.31**	**6.80**	**131.79**	**36.80**	**23.61**	**9.32**	**48.88**	**37.43**	**24.12**	10.56	**75.52**	**54.60**	**38.25**	**17.43**	**55.66**

Table 4. Cross-dataset Performance on Four Unseen Datasets with Novel Classes.

formance on familiar classes. The Base-Unseen subset includes FIVES [10], DSCA [45], XCAD [19], and the CRACK collection [50], testing generalization to new data distributions for familiar structure classes. The Novel category contains the public ROAD benchmark [20] and our in-house datasets LEAF, TYRE, and WIRE, representing completely new curvilinear structure classes from unseen distributions. The in-house images were captured with mobile phones and annotated with pixel-level masks; most images have a resolution of 1024×768.

4.2. Implementation Details

For fair comparison, all SAM-based baselines and SACM use the ViT-L image encoder from SAM at the same scale and are initialized with identical pretrained weights. All compared methods were trained under the same 3-shot per-dataset protocol (18 images total from DRIVE, CHASEDB1, DCA1, CrackTree, CREMI, and CORN). Experiments were conducted on a server with an NVIDIA RTX 4090 GPU (24 GB) and PyTorch version 2.9.1. Fine-tuning is performed for 50 epochs with a batch size of 1 and a learning rate of 3×10^{-4}, using the AdamW optimizer (β_1 = 0.9 and β_2 = 0.999) and a cosine learning rate scheduler. The image encoder block is frozen during fine-tuning; only the DLAda and the PFAF-D are updated.

4.3. Comparison to State-of-the-art

4.3.1. Quantitative Results

We evaluate SACM against representative curvilinear segmentation methods, including classic CNN-based models (U-Net [25], CS^{2}Net [22], BCUNet [43], MaskVSC [48], nnWNet [49]) and recent SAM-based adaptations (SAM-Med2d [4], SAM-OCTA [33], CWSAM [24]). We further include KnowSAM [9], a semi-supervised SAM adaptation, and SegDINO [40], which builds on DINOv3 [30] pretraining, to cover diverse adaptation strategies and pre-training regimes. According to the released code of the compared methods, all compared SAM-based methods use

point prompts during fine-tuning or inference, whereas our method is completely prompt-free and does not rely on any additional prompt information. Tables 2, 3 and 4 report the Dice, IoU, clDice, and HD95 scores on twelve curvilinear datasets. The upper part of the table represents small models without pre-trained weights, while the lower part represents fine-tuned models with pre-trained weights.

For the seen (Base) test sets (Table 2), SACM achieves state-of-the-art performance across all reported metrics. Concretely, SACM attains Dice scores of 78.89, 79.27, 75.67 and 55.38 on DRIVE, CHASEDB1, DCA1 and CORN respectively, with corresponding IoU values of 65.24, 65.72, 61.10 and 38.44, improved clDice measures and substantially reduced HD95 errors (e.g., 8.34 on DRIVE). These results indicate that, under the same 3-shot fine-tuning protocol, SACM consistently improves pixel-level overlap and boundary accuracy compared to both conventional CNN baselines and recent SAM-based adaptations.

On the unseen domains (Tables 3 and 4), SACM demonstrates strong cross-domain generalization without any prompt engineering. On the open unseen benchmarks DSCA, CRACK, XCAD and FIVES, SACM reports the following Dice and IoU results: DSCA 68.43 and 53.51, CRACK 63.17 and 47.80, XCAD 74.29 and 57.74, and FIVES 75.48 and 62.26. On the in-house novel domains ROAD, LEAF, TYRE and WIRE, SACM achieves Dice and IoU of 40.43 and 26.31 on ROAD, 36.80 and 23.61 on LEAF, 37.43 and 24.12 on TYRE, and 54.60 and 38.25 on WIRE. These improvements, particularly on structurally diverse datasets such as XCAD, FIVES and WIRE, suggest that our prompt-free approach effectively transfers curvilinear priors across domains.

4.3.2. Qualitative Results

Fig. 6 presents a qualitative comparison of SACM against BCUNet [43], nnWNet [49], CWSAM [24], KnowSAM [9], and SegDINO [40], across eight curvilinear segmentation datasets (DSCA [45], XCAD [19], ROAD [20], FIVES [10], CRACK, LEAF, TYRE, WIRE). SACM produces cleaner masks with improved structural continuity and sharper boundaries.

On FIVES and XCAD, the colorized segmentation overlays show more complete and anatomically plausible vascular and tubular structures, with fewer spurious fragments and fewer missing branches, demonstrating SACM's superior ability to maintain long-range structural continuity. On LEAF and ROAD, SACM preserves long-range connectivity and fine curvilinear detail under cluttered backgrounds and partial occlusions, whereas competing methods often fragment or over-cover the target structures. On CRACK and TYRE, dominated by thin, irregular, high-contrast patterns, SACM accurately captures subtle fractures and tread features with minimal noise and fewer breaks. Similar trends are observed across the remaining datasets.

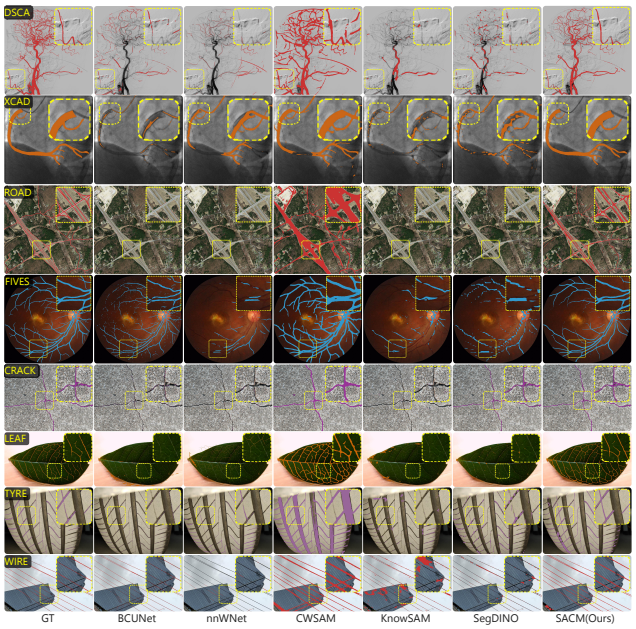


Figure 6. Visual comparison of segmentation results showing SACM's strong generalization to varying curvilinear datasets. Different colors are used to highlight differences between datasets for better visualization.

4.4. Ablation Studies

4.4.1. Impact of Different Components

We conduct an ablation study to assess the individual and combined effects of the block-internal adapter (Adapter-I), block external adapter (Adapter-E), adapter fusion (Adapter-F), and Dual-stage. Table 5 reports results on the WIRE dataset. The baseline SAM performs poorly (Dice 5.61%, clDice 2.32%), confirming that domain adaptation is essential. Adding Adapter-I or Adapter-E individually yields large gains (Dice 46.11%, clDice 14.83% and Dice 45.02%, clDice 15.24%, respectively), indicating both internal and external adapters contribute substantially. Using both adapters raises Dice to 50.38% (clDice 15.52%), and adding the Adapter-Fusion further increases Dice to 53.93% (clDice 16.02%). The full model with Dual-Stage refinement achieves the best performance (Dice 54.60%, clDice 17.43%), showing that internal adapters, external adapters, and multi-layer fusion provide complementary and cumulative improvements for curvilinear structure segmentation.

4.4.2. Impact of Training Shot Size

We studied SACM's data efficiency by varying the number of training shots per dataset. Experiments were conducted under 1-shot to 7-shot settings. As shown in Fig. 7, SACM achieves strong performance even with only 3-shot

Adapter-I	Adapter-E	Adapter-F	Dual-stage	Dice↑	clDice↑
-	-	-	-	5.61	2.32
✓	-	-	-	46.11	14.83
-	✓	-	-	45.02	15.24
✓	✓	-	-	50.38	15.52
✓	✓	✓	-	53.93	16.02
✓	✓	✓	✓	**54.60**	**17.43**

Table 5. Ablation study of different components on WIRE dataset. ✓ indicates the component is enabled.

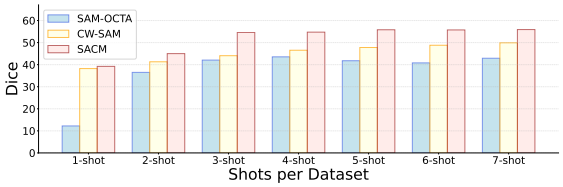


Figure 7. Dice of SACM versus the number of training shots per dataset on the WIRE dataset.

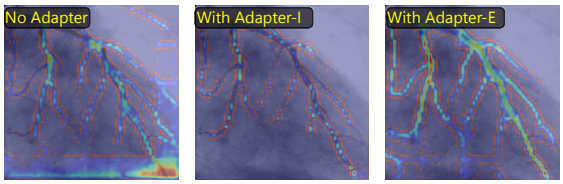


Figure 8. Grad-CAM visualizations of the original SAM, SAM with only Adapter-I, and SAM with only Adapter-E.

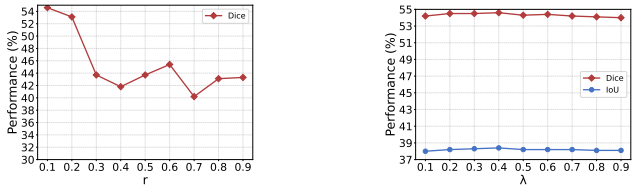
per dataset, outperforming other SAM-based models fine-tuned with the same amount of data. Performance improves with more shots, but the gains diminish beyond 5-shots, indicating SACM leverages limited annotations for robust curvilinear segmentation.

4.4.3. Grad-CAM of Different Adapters

To visualize the distinct contributions of the internal and external adapters, we generate Grad-CAM heatmaps for three configurations: original SAM without adapters, SAM with only Adapter-I, and SAM with only Adapter-E. As shown in Fig. 8, the original SAM exhibits diffuse attention patterns. With only Adapter-I, the model attends more to local vessel regions but lacks global context. In contrast, with only Adapter-E, the attention highlights broader vessel structures but introduces more irrelevant regions. This demonstrates that Adapter-I enhances local feature representation while Adapter-E captures global context, and their combination in SACM leads to comprehensive curvilinear segmentation.

4.4.4. Impact of Hyperparameters

We evaluate the effect of two key hyperparameters on the WIRE dataset via grid search: the adapter bottleneck ratio r and the loss weight λ. As shown in Fig. 9(a), performance peaks at $r = 0.1$. A smaller bottleneck constrains the adapter capacity, leading to more moderate updates to the pretrained encoder, while a larger r increases the adapter's degrees of freedom and can lead to overfitting on limited



(a) Bottleneck ratio r (b) Loss weight λ

Figure 9. Validation score with different bottleneck ratio r and loss weight λ.

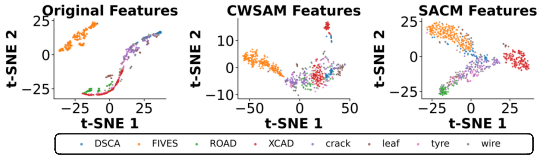


Figure 10. t-SNE visualization of SAM, CWSAM, and SACM; colors represent different data domains.

data. Based on this observation, we set $r = 0.1$ as the default setup. Similarly, for the loss weight λ, Fig. 9(b) shows that performance peaks near $\lambda = 0.4$, indicating that a balanced combination of BCE and Dice loss yields optimal results for curvilinear structure segmentation.

4.4.5. Visualization of t-SNE

To demonstrate the generalization of the SACM, we visualize feature embeddings from eight domains using t-SNE. As an illustration in Fig. 10, we compare the baseline of CWSAM with SACM. The CWSAM exhibits heavily overlapping, diffuse clusters with poor intra-domain cohesion and blurred domain boundaries. In contrast, SACM produces tighter, well-separated intra-domain groups, clearly distinguishing each domain in the embedding space. This improvement stems from dual-level adapter fine-tuning, which enhances domain-specific invariance.

5. Conclusion

In this paper, we presented SACM, a prompt-free framework for curvilinear structure segmentation built on SAM. By improving both local detail and global structural continuity, SACM achieves strong few-shot and cross-domain performance. Experiments on twelve datasets show that SACM consistently outperforms existing baselines using only 18 training images. Future work will explore continual adaptation, more efficient fine-tuning, and improved robustness to severe domain shifts.

6. Acknowledgment

The work was supported in part by the National Natural Science Foundation of China (62271359) and the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018. 2

[2] Fernando Cervantes-Sanchez, Ivan Cruz-Aceves, Arturo Hernandez-Aguirre, Martha Alicia Hernandez-Gonzalez, and Sergio Eduardo Solorio-Meza. Automatic segmentation of coronary arteries in x-ray angiograms using multiscale analysis and artificial neural networks. *Applied Sciences*, 9 (24), 2019. 5, 6

[3] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3359–3367, 2023. 2, 3

[4] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2, 6

[5] Venkateswararao Cherukuri, Vijay Kumar B.G., Raja Bala, and Vishal Monga. Deep retinal image segmentation with regularization under geometric priors. *IEEE Transactions on Image Processing*, 29:2552–2567, 2020. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3

[7] Ryunosuke Hamada, Tsubasa Minematsu, Cheng Tang, and Atsushi Shimada. Analysis of adapter in attention of change detection vision transformer. In *Proceedings of the Asian Conference on Computer Vision*, pages 34–49, 2024. 3

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[9] Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(5):2295–2306, 2025. 6, 7

[10] Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475, 2022. 5, 6, 7

[11] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, pages 29914–29934, 2023. 3

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 2

[13] Mónica G Larese, Rafael Namías, Roque M Craviotto, Miriam R Arango, Carina Gallo, and Pablo M Granitto. Automatic classification of legumes using leaf vein image features. *Pattern Recognition*, 47(1):158–168, 2014. 1

[14] Xiang Li, Chong Fu, Qun Wang, Wenchao Zhang, Chiu-Wing Sham, and Junxin Chen. Dmsa-unet: Dual multi-scale attention makes unet more strong for medical image segmentation. *Knowledge-Based Systems*, 299:112050, 2024. 2

[15] Ji Lin, Xingru Huang, Huiyu Zhou, Yaqi Wang, and Qianni Zhang. Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images. *Medical Image Analysis*, 89:102929–102943, 2023. 2

[16] Tianyong Liu, Zhiqing Zhang, Guojia Fan, Bin Li, Shoujun Zhou, Chengwu Xu, Gang Zhao, and Fuxia Yang. Mambavesselnet: A novel approach to blood vessel segmentation based on state-space models. *IEEE Journal of Biomedical and Health Informatics*, 29(3):2034–2047, 2025. 2

[17] Liana M Lorigo, Olivier D Faugeras, W Eric L Grimson, Renaud Keriven, Ron Kikinis, Arya Nabavi, and C-F Westin. Curves: Curve evolution for vessel segmentation. *Medical image analysis*, 5(3):195–206, 2001. 1

[18] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654–663, 2024. 2

[19] Yuxin Ma, Yang Hua, Hanming Deng, Tao Song, Hao Wang, Zhengui Xue, Heng Cao, Ruhui Ma, and Haibing Guan. Self-supervised vessel segmentation via adversarial learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7536–7545, 2021. 5, 6, 7

[20] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 5, 6, 7

[21] Lei Mou, Yitian Zhao, Li Chen, Jun Cheng, Zaiwang Gu, Huaying Hao, Hong Qi, Yalin Zheng, Alejandro Frangi, and Jiang Liu. Cs-net: channel and spatial attention network for curvilinear structure segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–730. Springer, 2019. 5, 6

[22] Lei Mou, Yitian Zhao, Huazhu Fu, Yonghuai Liu, Jun Cheng, Yalin Zheng, Pan Su, Jianlong Yang, Li Chen, Alejandro F Frangi, et al. Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. *Medical image analysis*, 67:101874, 2021. 2, 6

[23] Christopher G Owen, Alicja R Rudnicka, Robert Mullen, Sarah A Barman, Dorothy Monekosso, Peter H Whincup, Jeffrey Ng, and Carl Paterson. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. *Investigative ophthalmology & visual science*, 50(5):2004–2010, 2009. 5, 6

[24] Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 2, 3, 6, 7

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 1, 2, 6

[26] Ning Shen, Tingfa Xu, Shiqi Huang, Feng Mu, and Jianan Li. Expert-guided knowledge distillation for semi-supervised vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5542–5553, 2023. 2

[27] Yuhe Shen, Jiang Li, Weifang Zhu, Kai Yu, Meng Wang, Yuanyuan Peng, Yi Zhou, Liling Guan, and Xinjian Chen. Graph attention u-net for retinal layer surface detection and choroid neovascularization segmentation in oct images. *IEEE Transactions on Medical Imaging*, 42(11):3140–3154, 2023. 2

[28] Tianyi Shi, Xiaohuan Ding, Wei Zhou, Feng Pan, Zengqiang Yan, Xiang Bai, and Xin Yang. Affinity feature strengthening for accurate, complete and robust vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(8):4006–4017, 2023. 1

[29] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16560–16569, 2021. 5

[30] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 6

[31] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004. 5, 6

[32] Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24036–24045, 2024. 3

[33] Chengliang Wang, Xinrun Chen, Haojian Ning, and Shiying Li. Sam-octa: A fine-tuning strategy for applying foundation model octa image segmentation tasks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1771–1775, 2024. 3, 6

[34] Qijie Wang, Guandu Liu, and Bin Wang. Caps-adapter: Caption-based multimodal adapter in zero-shot classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5442–5450, 2024. 3

[35] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. 5

[36] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 34–43, Cham, 2022. Springer Nature Switzerland. 5

[37] Zhaozhi Xie, Bochen Guan, Weihao Jiang, Muyang Yi, Yue Ding, Hongtao Lu, and Lei Zhang. Pa-sam: Prompt adapter sam for high-quality image segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 3

[38] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023. 3

[39] Fan Yang, Lei Zhang, Sijia Yu, Danil Prokhorov, Xue Mei, and Haibin Ling. Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1525–1535, 2020. 1

[40] Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833*, 2025. 6, 7

[41] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 605–613. Springer International Publishing, 2019. 5

[42] Qin YuTao, Yang SiZhe, Hu Bang, and Ren Wei. Fcodt-net: A novel framework for high-precision medical image segmentation using contextual distillation transformer. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 2

[43] Hongbin Zhang, Xiang Zhong, Guangli Li, Wei Liu, Jiawei Liu, Donghong Ji, Xiong Li, and Jianguo Wu. Bcu-net: Bridging convnext and u-net for medical image segmentation. *Computers in Biology and Medicine*, 159:106960–106976, 2023. 2, 6, 7

[44] Haoran Zhang, Shuanghao Bai, Wanqi Zhou, Jingwen Fu, and Badong Chen. Promptta: Prompt-driven text adapter for source-free domain generalization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 3

[45] Jiong Zhang, Qihang Xie, Lei Mou, Dan Zhang, Da Chen, Caifeng Shan, Yitian Zhao, Ruisheng Su, and Mengguo Guo. Dsca: A digital subtraction angiography sequence dataset and spatio-temporal model for cerebral artery segmentation. *IEEE Transactions on Medical Imaging*, 44(6):2515–2527, 2025. 5, 6, 7

[46] Yishuo Zhang and Albert C. S. Chung. Retinal vessel segmentation by a transformer-u-net hybrid model with dual-

path decoder. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5347–5359, 2024. 2

[47] Yindong Zhang, Jie Chen, Li Wang, Miaohong Chen, Guoming Zhang, and Jianqiang Li. Dd-unet: Densely dilated u-net for curvilinear structure segmentation in fundus image. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2374–2379, 2023. 2

[48] Yi Zhou, Thiara Sana Ahmed, Meng Wang, Eric A. Newman, Leopold Schmetterer, Huazhu Fu, Jun Cheng, and Bingyao Tan. Masked vascular structure segmentation and completion in retinal images. *IEEE Transactions on Medical Imaging*, 44(6):2492–2503, 2025. 6

[49] Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnWNet: Rethinking the Use of Transformers in Biomedical Image Segmentation and Calling for a Unified Evaluation Benchmark . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20852–20862, 2025. 6, 7

[50] Qin Zou, Yu Cao, Qingquan Li, Qingzhou Mao, and Song Wang. Cracktree: Automatic crack detection from pavement images. *Pattern Recognition Letters*, 33(3):227–238, 2012. 5, 6