FMS^2: UNIFIED FLOW MATCHING FOR SEGMENTATION AND SYNTHESIS OF THIN STRUCTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Segmenting thin structures such as infrastructure cracks and anatomical vessels is hampered by fragile connectivity, expensive annotations, and poor cross-domain generalization. We propose FMS^2, a flow-matching framework that addresses all three with two modules, ***SegFlow*** for image→mask segmentation and ***SynFlow*** for mask→image synthesis. SegFlow is an encoder–decoder that casts segmentation as deterministic transport, regressing a velocity field on interpolated image–mask states and integrating from the image itself. Two exact identities describe this sampler: the velocity objective is equivalent to time-weighted endpoint regression on interpolated reference states, and the K-step output equals the one-step prediction plus a sum of signed corrections. Together with margin and residual assumptions, these identities characterize possible pixel-label changes and give sufficient conditions for reconnecting two components or recovering the decoded binary mask exactly. SynFlow renders mask-conditioned images using multi-scale mask injection and boundary gating, while class-conditional mask generation and controlled propagation vary sparsity, width, and branching. The same endpoint view applies to SynFlow, where extrapolation drifts by time-weighted trajectory acceleration, giving a conditional stability criterion for a fixed structural image readout already aligned with the conditioning mask. Across five crack and vessel benchmarks, SegFlow improves both overlap and topological metrics over nine baselines: mean IoU 0.601 versus 0.546 ($+10.1\%$) and Betti matching error 51.524 versus 73.856 (-30.2%), each against the strongest baseline on that metric. Synthetic augmentation improves SegFlow when it receives 25% of the real training labels, recovering most of the full-label performance on two crack subsets and increasing source-only cross-domain IoU by 0.109 on average. We release code, model checkpoints, and 10k crack and 1k vessel image–mask pairs.

Keywords: Flow Matching · Thin-Structure Segmentation · Topology-Aware Segmentation · Mask-Conditioned Image Synthesis · Label-Efficient Learning

1 INTRODUCTION

Segmenting thin structures is critical for infrastructure inspection and medical imaging. Yet cracks and vessels can be as narrow as one pixel, some crack masks occupy less than 1% of the field of view, and a one-pixel gap can break the connection needed for downstream metrology and planning (Mou et al., 2021; Shit et al., 2021; Yang & Farsiu, 2023; Gupta et al., 2023; Zuo et al., 2024). Pixel-wise annotation is labor-intensive and varies along low-contrast, branching structures and ambiguous margins (Lesage et al., 2009; Zhou et al., 2023; Falcetta et al., 2025). Models trained on one dataset can also fail when texture, material, illumination, or acquisition protocol changes (Chen et al., 2024; Wu et al., 2025; Yang & Farsiu, 2023; Zhou et al., 2025a). Thus fragile connectivity, expensive supervision, and domain shift are coupled challenges.

Architectural advances have improved in-domain accuracy. Multi-scale CNNs with edge or contour heads, CNN–Transformer hybrids, and Mamba/SSM variants improve local detail, long-range context, and efficiency (Mou et al., 2021; Acebes et al., 2024; Chen et al., 2024; Zhou et al., 2025b; Liu et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

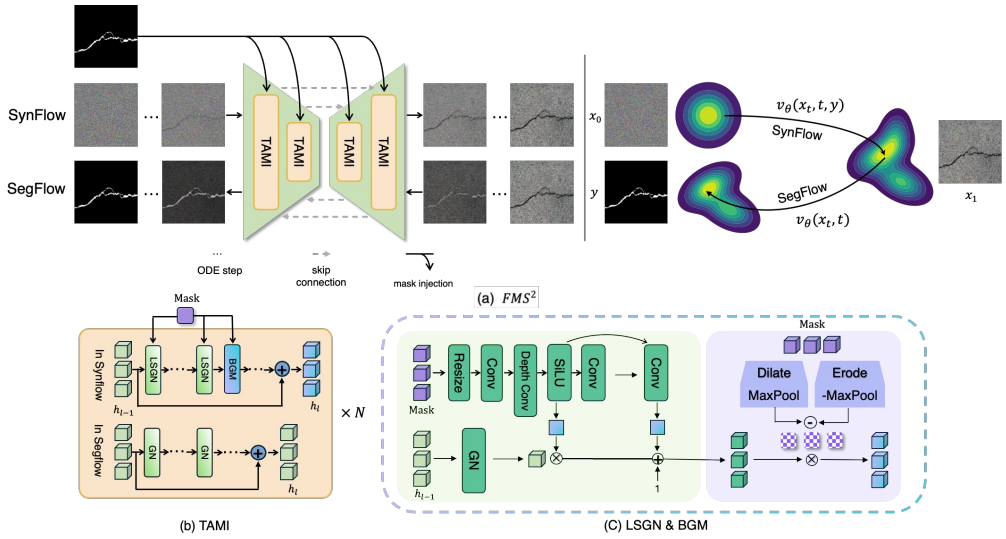


Figure 1: **Architecture and transport view of FMS^2. (a)** The U-Net flow modules transport image→mask (SegFlow) and mask-conditioned noise→image (SynFlow). **(b)** Topology-Aware Mask Injection (TAMI) combines lite-SPADE GroupNorm (LSGN) and boundary gating (BGM). **(c)** LSGN injects a mask pyramid into group-normalized features; BGM gates crack/vessel boundaries.

Limited labels and severe foreground imbalance constrain larger models, while an overlap loss can scarcely penalize the removal of a one-pixel bridge. Topology-aware losses such as clDice and component graphs target connectivity at the final prediction (Shit et al., 2021; Kirchhoff et al., 2024; Lux et al., 2025); they do not describe how the mask forms. Synthetic data can address label cost and appearance shift, but at one-pixel width normalization and latent conditioning can weaken the match between a generated image and its mask (Lei et al., 2024; 2025). Reusing only observed masks also limits structural diversity.

We introduce FMS^2 through two directed transports. In **SegFlow**, the observed image itself initializes the ODE; velocities on intermediate states between that image and its paired binary mask are learned in pixel space. This distinguishes its supervised path from image-conditioned random-mask denoising and latent joint segmentation–synthesis. Exact endpoint identities then express the integrated prediction as signed corrections to its one-step endpoint; under stated margin and residual assumptions, these can change pixel labels or reconnect components. Self-conditioning carries a running mask estimate, while additional integration steps can refine one-pixel-scale cracks in CrackTree260 and CrackLS315 and highly branched DRIVE vessels. SynFlow separately transports noise to an image conditioned on the target mask through multi-scale spatial injection and boundary gating. Class-conditional mask generation and a separate propagation mechanism vary sparsity, width, branching, and connectivity (Supplementary Sec. D.1); Supplementary Table S5 measures bidirectional KL divergence in resulting geometric distributions. Together they provide new mask-aligned training pairs for scarce supervision and source-only transfer. In the manual mask–image alignment test, SynFlow exceeds the evaluated latent-diffusion renderers LDM, ControlNet, and T2I-Adapter on all four reported metrics (Table 5).

Across five crack and vessel datasets, SegFlow improves overlap and topological metrics (Table 1). With only 25% of real training labels supplied directly to SegFlow, SynFlow augmentation approaches full supervision on CRACK500 and CrackTree260; XCAD retains a gap. Across three source-only crack transfers, the average relative mIoU gain is 46.6% over SegFlow (the mean of transfer-specific gains). Our contributions are:

- ***SegFlow***, with the best mean overlap and Betti error in the reported benchmark, exact endpoint-correction identities, and conditional connectivity analysis without auxiliary topology heads or post-processing.

- ***SynFlow***, a mask-conditioned renderer with multi-scale injection, boundary gating, and a controllable crack/vessel mask generator. Its paired images improve SegFlow with partial real data and under the reported shifts; manual annotations assess alignment.
- **A released collection** of **10k crack** and **1k vessel** synthetic image–mask pairs.

2 Related Work

Architecture and topology. Edge-aware CNNs retain local crack and vessel cues (Mou et al., 2021); Mask2Former, DConnNet, and SSM designs add global context or directional priors (Cheng et al., 2022; Yang & Farsiu, 2023; Liu et al., 2025). Yet high overlap can conceal a missing one-pixel bridge. clDice and Topograph therefore target centerline and component structure (Shit et al., 2021; Lux et al., 2025). They constrain the prediction; SegFlow also studies changes made along the image-to-mask inference trajectory.

Generative segmentation. MedSegDiff-V2 denoises image-conditioned masks (Wu et al., 2024); FlowSDF evolves an image-conditioned signed-distance representation from a prior (Bogensperger et al., 2025). SemFlow couples image and mask in a latent rectified flow (Wang et al., 2024), and SymmFlow uses a symmetric formulation for multiple tasks (Caetano et al., 2026). SegFlow instead supervises velocities on paired image–binary-mask interpolations, initializes its pixel-space ODE with the observed image, and characterizes signed integration corrections. Explicit margin and residual conditions, rather than flow matching alone, support its connectivity analysis.

Thin-mask-conditioned synthesis. LDM, ControlNet, and T2I-Adapter produce controllable latent-diffusion images (Rombach et al., 2022; Zhang et al., 2023; Mou et al., 2024). At one-pixel width, normalization and latent conditioning can weaken mask–image alignment; prior curvilinear and crack synthesis work explicitly addresses semantic fidelity (Lei et al., 2024; 2025; Li et al., 2025; Lin et al., 2026). SynFlow uses a separate pixel-space noise-to-image flow with multi-scale mask injection and boundary gating, supplied with generated and propagated masks. VesselFM employs flow-based synthetic vessel data, and SemFlow/SymmFlow connect flow models with synthesis (Wittmann et al., 2025; Wang et al., 2024; Caetano et al., 2026). Table 5 quantifies mask–image alignment against diffusion renderers.

3 Methodology

3.1 Preliminaries

Flow matching (FM) regresses reference-state velocities without training-time ODE integration. A state $\mathbf{x}_t$ evolves as

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{u}_\theta(\mathbf{x}_t, t \mid c), \tag{1}$$

Its learned endpoint need not equal the training target. SegFlow follows an image–mask path and SynFlow the noise–image path in Eq. (13); both integrate from their specified initial states (Fig. 1).

3.2 SegFlow: Segmentation as continuous transport

SegFlow trains a U-Net velocity field on paired image x_0 and binary mask x_1 in the same state space (Fig. 1). Along the linear path in Eq. (3), it targets the constant velocity $x_1 - x_0$:

$$\mathcal{L}_{\text{SegFlow}} \;=\; \mathbb{E}_{(x_0,x_1),\, t\sim[0,1]}\Big[\, \big\| \, u_\theta(x_t,t) - (x_1 - x_0)\big\|_2^2 \,\Big]. \tag{2}$$

We sample training times t and form

$$x_t \;=\; (1-t)\,x_0 \;+\; t\,x_1, \tag{3}$$

The endpoint estimate $D_\theta(z,t) = z + (1-t)u_\theta(z,t)$ satisfies, on reference states with $t < 1$,

$$D_\theta(x_t,t) - x_1 = (1-t)\big[u_\theta(x_t,t) - (x_1 - x_0)\big]. \tag{4}$$

Thus the same velocity MSE corresponds pointwise to endpoint MSE weighted by $(1-t)^{-2}$; no separate singular loss is optimized. Under a finite second moment, unrestricted population regression satisfies, for almost every reference state,

$$u^*(z,t) = \mathbb{E}[x_1 - x_0 \mid X_t = z, t], \qquad D^*(z,t) = \mathbb{E}[x_1 \mid X_t = z, t]. \tag{5}$$

Since $x_1 = X_t + (1-t)(x_1 - x_0)$, colliding training paths average target velocities. These reference-state identities do not guarantee paired-mask recovery along a learned rollout.

The U-Net embeds time as $\tau = 1000t$. Unlike end-state class logits or random-mask denoising, SegFlow supervises intermediate image–mask states and integrates from the observed image. It also differs from topology-aware end-state losses (Cheng et al., 2022; Shit et al., 2021); Tables 2–3 and Fig. 2 probe behavior without isolating every architectural difference.

At test time, deterministic fixed-step Euler integration starts from input x_0 at t=0 and takes K steps to t=1:

$$x_{t+\Delta t} \;=\; x_t \;+\; \Delta t\, u_\theta(x_t, t), \tag{6}$$

Here $\Delta t = 1/K$ and $t \in \{0, \Delta t, 2\Delta t, \ldots, 1-\Delta t\}$. Thresholding terminal state $\hat{M}$ gives a binary mask; the released implementation reads masks as RGB to match image channels and thresholds the negative channel mean at zero (Supplementary Sec. A.1). Thin connections are sensitive to local errors: later velocities can repair or introduce them. Overlap and Betti metrics evaluate their net effect without auxiliary skeleton losses (Shit et al., 2021) or graph post-processing.

Self-conditioning. By default, SegFlow concatenates evolving x_t with a running terminal-mask estimate $\tilde{x}_1$ and evaluates $u_\theta^{\mathrm{SC}}([x_t, \tilde{x}_1], t)$; Table 3 studies its removal. The straight-line endpoint estimate is

$$\hat{x}_1 \;=\; x_t + (1-t)\, u_\theta^{\mathrm{SC}}([x_t, \tilde{x}_1], t). \tag{7}$$

During training, with probability p_{sc}, a preliminary zero-memory pass supplies a stop-gradient terminal estimate to a second pass. The loss remains MSE velocity regression (Supplementary Sec. E.3).

Trajectory corrections and connectivity. For either variant, let Z_k be the Euler state at $t_k = k/K$, v_k the evaluated velocity (including self-conditioning history), and $D_k = Z_k + (1-t_k)v_k$ its terminal-mask estimate. This implemented-update view relates to Lagrangian and flow-map analyses (Milanfar, 2026; Boffi et al., 2024).

Fix an undirected foreground adjacency graph on the pixel set Ω and an affine pixelwise signed readout q implementing the mask threshold. Let $s_{k,p} = q(D_k)_p$ and $P_k = \{p : s_{k,p} > 0\}$.

Proposition 1 (Endpoint corrections and a connectivity certificate)**.** *For an integer $K \geq 1$, Euler sampling satisfies $Z_K = D_{K-1}$ and, for $0 \leq k \leq K-2$,*

$$D_{k+1} - D_k = (1 - t_{k+1})(v_{k+1} - v_k). \tag{8}$$

Consequently, the output equals its one-step prediction plus signed endpoint corrections:

$$Z_K = D_0 + \sum_{k=0}^{K-2} (1 - t_{k+1})(v_{k+1} - v_k). \tag{9}$$

For $0 \leq k \leq K-1$, define the remaining score variation

$$T_{k,p} = \sum_{j=k}^{K-2} |s_{j+1,p} - s_{j,p}|. \tag{10}$$

Then the final changed-label set obeys

$$P_k \triangle P_{K-1} \subseteq \{p : |s_{k,p}| \leq T_{k,p}\}. \tag{11}$$

Moreover, suppose the corrections make a nonempty connected gap $G \subseteq \Omega \setminus P_k$ foreground, preserve all labels outside G, and G is adjacent to exactly two distinct components of P_k. Then $\beta_0(P_{K-1}) = \beta_0(P_k) - 1$, where β_0 counts foreground components.

Proof sketch. Substituting $Z_{k+1} = Z_k + v_k/K$ into D_{k+1} and telescoping gives Eq. (9) with $D_{K-1} = Z_K$. The triangle inequality bounds future score changes by $T_{k,p}$; a connected gap touching exactly two components merges them (full proof in Supplementary Sec. A). Corrections reflect learned velocity changes, not step count; self-conditioning changes v_k but preserves the identity. Stronger recovery requires residual and rollout assumptions (Supplementary Sec. A.4). A label flip needs variation to reach its current margin, and a bridge improves annotation agreement only if

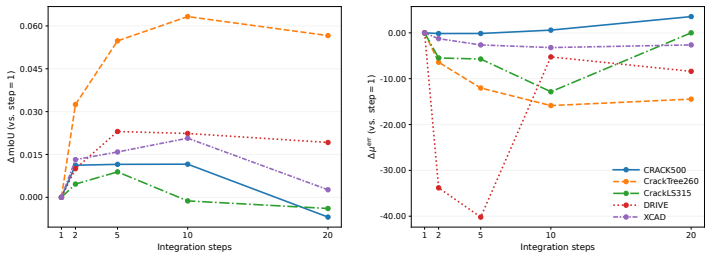


Figure 2: SegFlow mIoU and μ^{err} changes relative to one Euler step as the integration budget varies.

its pixels and attachments are correct. Component reduction need not lower Betti matching error. The future-dependent certificate gives neither an early-stopping rule nor an unconditional topology guarantee: an exactly fitted compatible straight path reaches the target in one step. We assess learned corrections with the integration-budget plot and topology metrics.

3.3 SynFlow: Rendering as Topology-Aware Transport

SynFlow is designed to render images whose thin structures follow a supplied mask; we measure alignment against it. Starting from noise seed $\mathbf{z} \sim p_0$, the mask-conditioned ODE produces

$$\hat{\mathbf{x}} \;=\; \mathbf{z} \;+\; \int_0^1 \mathbf{u}_\theta(\mathbf{x}_t, t \mid \mathbf{M})\, dt, \tag{12}$$

with numerical integration specified in the experimental protocol. For each training pair $(\mathbf{x}_1, \mathbf{M})$, the mask-conditioned FM interpolant has $(\mathbf{x}_0, \mathbf{x}_1) = (\mathbf{z}, \mathbf{x}_1)$,

$$\mathbf{x}_t \;=\; (1-t)\,\mathbf{z} \;+\; t\,\mathbf{x}_1, \qquad \partial_t \mathbf{x}_t = \mathbf{x}_1 - \mathbf{z}, \tag{13}$$

The implementation sets the flow matcher's interpolation noise scale to $\sigma = 0$, so the reference endpoints match the sampling source $\mathbf{z}$ and image target $\mathbf{x}_1$. Supplementary Sec. A.5 discusses how nonzero noise would alter this endpoint relation. The training objective regresses the conditional velocity along the reference path,

$$\mathcal{L}_{\text{SynFlow}} \;=\; \mathbb{E}_{(\mathbf{x}_1,\mathbf{M})}\, \mathbb{E}_{\mathbf{z}}\, \mathbb{E}_{t\sim\mathcal{U}(0,1)} \Big[\big\| \mathbf{u}_\theta(\mathbf{x}_t, t \mid \mathbf{M}) - \partial_t \mathbf{x}_t \big\|_2^2 \Big]. \tag{14}$$

Training samples an image–mask pair, noise seed, and time; the mask conditions every velocity evaluation. Three U-Net additions supply geometry and emphasize thin boundaries:

Multi-Scale Mask Injection. Mask pyramids enter matching feature resolutions; time enters through sinusoidal or MLP embeddings.

Lite-SPADE GroupNorm (LSGN). A small convolutional head maps each mask scale to spatial affine parameters for group-normalized features (Fig. 1):

$$\mathrm{LSGN}(h \mid \mathbf{M}^{(s)}) \;=\; \gamma(\mathbf{M}^{(s)}) \odot \mathrm{GN}(h) \;+\; \beta(\mathbf{M}^{(s)}). \tag{15}$$

This injects spatially varying mask information; in the implementation, $\gamma = 1 + \gamma_{\mathrm{raw}}$.

Boundary-gating Module (BGM). A clipped mask morphological gradient and a learned scalar gain form a boundary gate for decoder and skip features (Fig. 1):

$$h_{\text{out}}(\mathbf{p}) = G_B(\mathbf{p}) h_{\text{in}}(\mathbf{p}), \quad G_B(\mathbf{p}) = 1 + \alpha\, \mathrm{clip}\big[(\mathrm{dilate}\, M - \mathrm{erode}\, M)_{\mathbf{p}}, 0, 1\big], \tag{16}$$

The gate modulates locations where small errors interrupt thin structures; its gain is initialized at zero. Supplementary Sec. D.2 compares mask-injection locations and Figs. S5–S9 show dataset-specific renderings.

Conditional structural stability. Multi-scale injection, LSGN, and BGM supply mask-dependent features to learn alignment; the following statement concerns its persistence once attained. Fix $\mathbf{M}$ and let $\bar{I}_a = Z_a + (1-a)u_\theta(Z_a, a \mid \mathbf{M})$ be the endpoint extrapolation at $a < 1$. Let a fixed, mask-independent structural score readout R of the image agree in sign with $\mathbf{M}$ at $\bar{I}_a$, with minimum signed margin $\gamma > 0$. If $\|R(I) - R(J)\|_\infty \le L_R \|I - J\|_\infty$ for the compared images, the numerical output retains every readout label provided

$$L_R \left[\int_a^1 (1-t) \|A_t\|_\infty \, dt + \delta_{\text{out}} \right] < \gamma, \tag{17}$$

Here $A_t = (\partial_t u_\theta + J_z u_\theta \, u_\theta)(Z_t, t \mid \mathbf{M})$ is the ODE acceleration, and δ_{out} bounds total numerical and any output postprocessing error relative to the continuous endpoint (derivation and regularity assumptions in Supplementary Sec. A).

Along the continuous ODE, $\frac{d}{dt}\big[Z_t + (1-t)u_\theta(Z_t, t \mid \mathbf{M})\big] = (1-t)A_t$: remaining drift is acceleration-weighted even though the reference training velocity is constant. The readout is an analytical observable, not a training module. This condition assumes an initial margin and readout sensitivity; FM does not guarantee either. It concerns continuous and numerical endpoints rather than the SegFlow Euler identity; acceleration and error bounds must reflect the integrated field, including guidance if used. Table 5 uses manual annotation to assess mask–image alignment; FID and KID assess complementary distributional properties.

Controllable mask generator. We train a small class-conditional flow-matching mask generator on sparsity bins (e.g., 0–0.5%, 0.5–1%) and render sampled masks with SynFlow. In parallel, propagation grows regions, extends skeleton branches, and thins or thickens their radius, broadening sparsity, length, width, and connectivity (Supplementary Sec. D.1 and Fig. S4). Supplementary Table S5 reports lower bidirectional KL divergences for measured density, length, and branch-point distributions under augmentation. The separately trained modules couple through synthetic pairs.

4 Experiments

We evaluate SegFlow on five crack and vessel datasets, then assess SynFlow through mask alignment, partial-label learning, and source-only transfer.

Table 1: SegFlow versus nine baselines on five datasets. Best results are in **bold**; second-best results are underlined.

		CRACK500				CrackTree260				CrackLS315			
Method	**Venue**	**Overlap ↑**		**Topological**		**Overlap ↑**		**Topological**		**Overlap ↑**		**Topological**	
		mIoU	F1	clDice↑	μ^{err} ↓	mIoU	F1	clDice↑	μ^{err} ↓	mIoU	F1	clDice↑	μ^{err} ↓
Mask2Former	CVPR 2022	0.562	0.710	0.821	27.868	0.332	0.497	0.552	119.536	0.308	0.467	0.522	93.310
DConnNet	CVPR 2023	0.571	0.720	0.821	23.651	0.341	0.506	0.546	252.225	0.295	0.451	0.481	120.670
SemFlow	NeurIPS 2024	0.172	0.286	0.287	154.372	0.045	0.085	0.085	338.955	0.041	0.079	0.078	142.060
FlowSDF	IJCV 2025	0.531	0.687	0.763	36.558	0.577	0.728	0.729	157.830	0.221	0.358	0.358	137.150
Topograph	ICLR 2025	0.471	0.619	0.740	43.860	0.184	0.308	0.272	161.975	0.114	0.203	0.186	180.210
SCSegamba	CVPR 2025	0.549	0.702	0.794	24.791	0.437	0.604	0.609	202.800	0.386	0.548	0.557	94.130
CrackSSM	CVPR 2026	0.612	0.753	0.841	**23.512**	0.461	0.629	0.677	190.805	0.379	0.542	0.585	**91.570**
MixerCSeg	CVPR 2026	0.551	0.701	0.802	25.151	0.424	0.592	0.611	206.230	0.372	0.531	0.548	99.400
SymmFlow	AAAI 2026	0.553	0.705	0.788	33.930	0.470	0.629	0.630	117.960	0.358	0.520	0.521	108.770
SegFlow	**Ours**	**0.624**	**0.763**	**0.847**	24.884	**0.645**	**0.771**	**0.772**	**76.070**	**0.442**	**0.603**	**0.603**	96.260

		DRIVE				XCAD				Mean (5 datasets)			
Method	**Venue**	**Overlap ↑**		**Topological**		**Overlap ↑**		**Topological**		**Overlap ↑**		**Topological**	
		mIoU	F1	clDice↑	μ^{err} ↓	mIoU	F1	clDice↑	μ^{err} ↓	mIoU	F1	clDice↑	μ^{err} ↓
Mask2Former	CVPR 2022	0.444	0.615	0.661	280.480	0.616	0.761	0.812	15.923	0.452	0.610	0.674	107.423
DConnNet	CVPR 2023	0.567	0.722	0.780	**48.104**	0.636	0.776	**0.835**	**9.827**	0.482	0.635	0.693	90.895
SemFlow	NeurIPS 2024	0.295	0.454	0.449	149.208	0.157	0.269	0.225	148.942	0.142	0.235	0.225	186.707
FlowSDF	IJCV 2025	0.635	0.776	0.776	130.313	0.592	0.742	0.780	18.125	0.511	0.658	0.681	95.995
Topograph	ICLR 2025	0.553	0.711	0.676	239.062	0.536	0.696	0.676	120.500	0.372	0.507	0.510	149.121
SCSegamba	CVPR 2025	0.605	0.753	0.765	65.563	0.564	0.720	0.759	23.442	0.508	0.665	0.697	82.145
CrackSSM	CVPR 2026	0.626	0.769	0.819	61.063	**0.654**	**0.790**	**0.835**	12.808	0.546	0.696	0.751	75.951
MixerCSeg	CVPR 2026	0.603	0.751	0.788	83.688	0.574	0.727	0.786	17.269	0.505	0.661	0.707	86.348
SymmFlow	AAAI 2026	0.462	0.629	0.580	98.063	0.579	0.732	0.770	10.558	0.484	0.643	0.658	73.856
SegFlow	**Ours**	**0.653**	**0.789**	**0.830**	49.792	0.642	0.780	0.831	10.615	**0.601**	**0.741**	**0.777**	**51.524**

4.1 EXPERIMENTAL SETTINGS

Datasets. We use five datasets: three crack sets (CRACK500, CrackTree260, CrackLS315 (Zhang et al., 2016; Zou et al., 2018; 2012)) and two vessel sets (DRIVE, XCAD (Staal et al., 2004; Ma et al., 2021)). Crack images range from handheld to controlled laser acquisition, with expert masks and 0.25%–2.84% mean foreground coverage (Supplementary Table S1). DRIVE has color fundus photographs with ophthalmologist vessel maps; XCAD has coronary X-ray angiograms with curated, highly branched arterial trees and variable contrast flow. Statistics, preprocessing, and splits are in Supplementary Sec. B and Table S1.

Comparison methods. Nine baselines span transformers (Mask2Former (Cheng et al., 2022)), directional connectivity (DConnNet (Yang & Farsiu, 2023)), topology (Topograph (Lux et al., 2025)), state-space designs (SCSegamba, CrackSSM, MixerCSeg (Liu et al., 2025; Gu et al., 2026; Zhao et al., 2026)), and generative/flow models (SemFlow, FlowSDF, SymmFlow (Wang et al., 2024; Bogensperger et al., 2025; Caetano et al., 2026)). They include crack-specific methods and complementary supervision on the reported curated splits.

Evaluation metrics. We report mIoU, Dice (F1), clDice (Shit et al., 2021), and spatially matched Betti error μ^{err} (Stucki et al., 2023) for β_0 and β_1; definitions are in Supplementary Sec. B.1.

Implementation details. Inputs are cropped/resized to 256×256; each crack set uses 500 curated patches split $80/10/10$, while DRIVE/XCAD follow Supplementary Sec. B (XCAD: $100/26$ images). The nine baselines use official code on these splits. SegFlow/SynFlow train with Adam on two NVIDIA A100 GPUs for 42k/150k iterations; schedules and inference settings are in Supplementary Sec. B.2.

4.2 COMPARISON OF SEGFLOW WITH SOTA METHODS

Overall results. Among nine baselines on five datasets (Table 1), SegFlow has the highest mean IoU (0.601 versus CrackSSM's 0.546; $+10.1\%$) and lowest mean Betti error (51.524 versus SymmFlow's 73.856; -30.2%). The benchmark excludes SynFlow images; common-backbone and integration studies probe the formulation.

SegFlow leads mIoU on all three crack sets and has Betti error 76.070 on CrackTree260 versus SymmFlow's 117.960. Other baselines achieve lower errors on CRACK500, CrackLS315, DRIVE, or XCAD; SegFlow leads the five-set mean. Inference cost appears in Supplementary Table S2.

Flow-based comparators. FlowSDF samples signed-distance masks, SemFlow transports image–mask states in a latent space, and SymmFlow uses symmetric tasks (Bogensperger et al., 2025; Wang et al., 2024; Caetano et al., 2026). Their mean IoUs are 0.511, 0.142, and 0.484, respectively, versus SegFlow's 0.601 without SynFlow pairs. SymmFlow is the closest flow comparator in mean Betti error (73.856 versus 51.524); these model-level results do not isolate individual design choices.

Supervision and topology. Under a common U-Net backbone (Table 2), mean Betti error is 51.524 versus 71.348 for clCE and 73.738 for Skeleton Recall, although both losses improve on CrackLS315. This comparison does not isolate supervision from inference.

Table 2: Betti matching error μ^{err} under common-backbone supervision. Lower is better.

Method	C500	CT260	CLS315	DRIVE	XCAD	Mean
clCE (MICCAI)	25.163	166.415	**92.700**	55.479	16.982	71.348
Skeleton Recall (ECCV 2024)	26.779	144.850	92.930	85.898	18.231	73.738
SegFlow	**24.884**	**76.070**	96.260	**49.792**	**10.615**	**51.524**

Table 3: Self-conditioning (SC) ablation across five datasets.

Metric / variant	C500	CT260	CLS315	DRIVE	XCAD
↑ mIoU w/o SC	0.612	0.639	0.438	0.646	0.638
↑ mIoU w/ SC	**0.624**	**0.645**	**0.442**	**0.653**	**0.642**
↓ μ^{err} w/o SC	27.640	77.175	101.040	61.958	11.558
↓ μ^{err} w/ SC	**24.883**	**76.070**	**96.260**	**49.791**	**10.615**

Integration and self-conditioning. Figure 2 shows nonmonotonic gains: at 20 steps, CRACK500 and CrackLS315 IoU fall below one-step results. Mean IoU/Betti error improve from $0.595/55.874$ without self-conditioning to $0.601/51.524$ with it (Table 3); inference cost appears in Supplementary Table S2.

Ablation of LSGN and BGM. Table 4 ablates LSGN and BGM separately; together they give the best overlap, topology, and FID. FID measures distributional fidelity; manual mask–image alignment appears below.

Table 4: LSGN/BGM ablation on two crack sets; arrows give the favorable metric direction.

Components		CRACK500					CrackTree260				
LSGN	BGM	mIoU↑	F1↑	clDice↑	μ^{err} ↓	FID↓	mIoU↑	F1↑	clDice↑	μ^{err} ↓	FID↓
×	✓	0.593	0.737	0.818	30.2	26.44	0.567	0.710	0.710	97.2	26.68
✓	×	0.581	0.727	0.803	34.2	30.48	0.562	0.706	0.707	94.3	28.80
✓	✓	**0.613**	**0.754**	**0.835**	**28.0**	**18.36**	**0.603**	**0.739**	**0.740**	**88.0**	**20.14**

4.3 SynFlow versus latent-diffusion synthesis

Figure 3 compares images and segmentation predictions for ControlNet, T2I-Adapter, and SynFlow. Table 5 adds LDM and compares manual annotations of generated images with conditioning masks. SynFlow exceeds ControlNet in mIoU/clDice ($0.561/0.852$ versus $0.467/0.738$) and improves on T2I-Adapter's Betti error (4.583 versus 5.917).

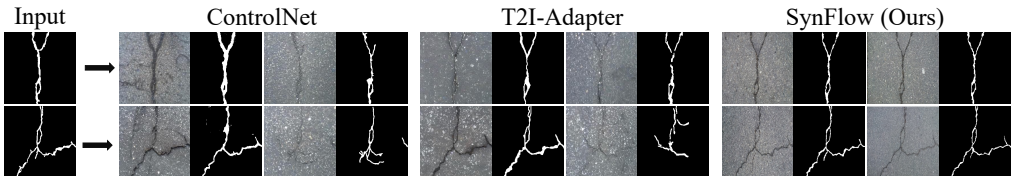


Figure 3: Mask-conditioned crack synthesis: each input mask (left) appears with a generated image and its segmentation prediction for ControlNet, T2I-Adapter, and SynFlow.

Table 5: Manual mask–image alignment. Generated images were annotated and compared with conditioning masks; higher is better except for Betti error.

Method	Venue	mIoU↑	F1↑	clDice↑	μ^{err} ↓
LDM	CVPR 2022	.329	.493	.575	6.833
ControlNet	ICCV 2023	.467	.632	.738	8.667
T2I-Adapter	AAAI 2024	.413	.574	.676	5.917
SynFlow	Ours	**.561**	**.717**	**.852**	**4.583**

Qualitative comparison. Figure 4 compares crack/vessel predictions; arrows mark missed branches and spurious fragments. Supplementary Figs. S1 and S2 show all datasets and failures.

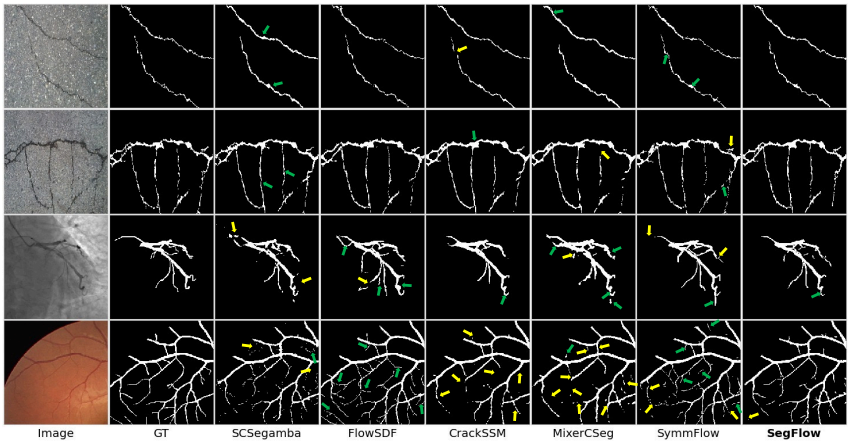


Figure 4: Visual comparison of SegFlow against SOTA methods. Yellow arrows denote missed structures (FNs), green arrows indicate spurious predictions (FPs).

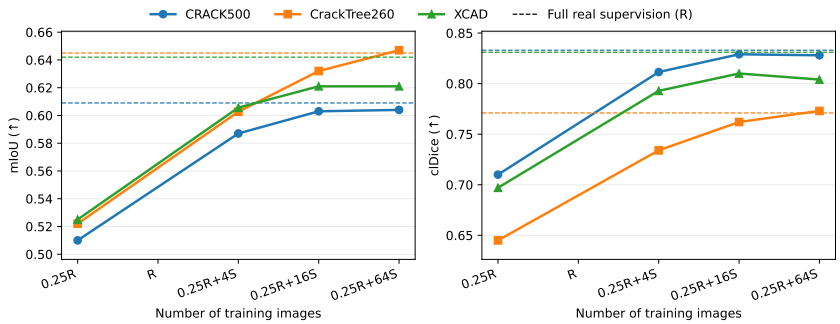


Figure 5: **SynFlow label efficiency.** SegFlow mIoU/clDice for $0.25R$ real plus $4S/16S/64S$ synthetic pairs ($S = 0.25R$); dashed lines: full R real. $R = 400$ for CRACK500/CrackTree260, 100 for XCAD.

Label efficiency. Figure 5 approaches full supervision on CRACK500 and CrackTree260 with $0.25R$ real SegFlow labels plus SynFlow pairs; XCAD retains a gap, and CRACK500/XCAD clDice decline slightly from 16S to 64S. Generated pairs and mask/image FID appear in Supplementary Figs. S3–S4 and S10 and Table S4. Full annotation cost also depends on labels for SynFlow and its mask generator (Supplementary Sec. D.1).

Cross-domain robustness. Across three source-only crack transfers (Supplementary Table S6), classic DA adds 0.006 mean IoU (2.4% relative); SynFlow adds 0.109 (46.6%) over SegFlow. Mask variants span sparsity, width, and branching; Supplementary Table S5 reports smaller bidirectional KL gaps in density, length, and branch-point distributions. These data do not isolate geometric diversity from appearance or training volume.

4.4 Scope of exact-mask recovery

Aggregate scores cannot certify individual masks or bound rollout error. For a fixed pair under base SegFlow, let $h = 1/K$, $V = x_1 - x_0$, $X_k = x_0 + t_k V$, and $r_k = u_\theta(X_k, t_k) - V$. If $\|u_\theta(Z_k, t_k) - u_\theta(X_k, t_k)\|_\infty \le L_k \|Z_k - X_k\|_\infty$, the Euler error recursion from $Z_0 = X_0$ gives

$$\|Z_K - x_1\|_\infty \le B_K := \sum_{j=0}^{K-1} h\|r_j\|_\infty \prod_{i=j+1}^{K-1} (1 + hL_i). \tag{18}$$

For an affine readout q with Lipschitz constant c_q and target scores $q(x_1)_p \in \{-1, +1\}$, $c_q B_K < 1$ recovers the exact decoded binary mask and its topology. Mean training MSE cannot ensure these samplewise residual and rollout bounds; evolving self-conditioning also needs a memory-discrepancy bound. Supplementary Sec. A.4 proves the criterion and gives a one-pixel counterexample.

5 Conclusion

We presented FMS^2, a flow-matching framework that links thin-structure segmentation with mask-conditioned image synthesis. SegFlow transports the observed image toward a mask, while SynFlow combines multi-scale mask injection, boundary gating, and controllable mask generation to produce image–mask pairs with varied structures. Our endpoint and multistep identities establish conditions under which trajectory corrections change pixel labels, reconnect components, and recover the decoded mask. Across five crack and vessel datasets, SegFlow improves mean IoU by 10.1% and reduces mean Betti matching error by 30.2% relative to the strongest of nine competing baselines on each respective metric. With 75% fewer real SegFlow training labels, SynFlow augmentation recovers most full-label performance on two crack subsets. Under the reported source-only transfer protocols, it increases average cross-domain IoU by 0.109 (46.6% relative) over SegFlow. We release code, checkpoints, and 10k crack and 1k vessel image–mask pairs to support further research on thin-structure segmentation and synthesis.

REFERENCES

Cesar Acebes, Abdel Hakim Moustafa, Oscar Camara, and Adrian Galdran. The centerline-cross entropy loss for vessel-like structure segmentation: Better topology consistency without sacrificing accuracy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 710–720. Springer, 2024.

Nicholas Matthew Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Trans. Mach. Learn. Res.*, 2025, 2024. URL https://api.semanticscholar.org/CorpusID:270380005.

Lea Bogensperger, Dominik Narnhofer, Alexander Falk, Konrad Schindler, and Thomas Pock. Flowsdf: Flow matching for medical image segmentation using distance transforms. *International Journal of Computer Vision*, 133(7):4864–4876, 2025.

Francisco Caetano, Christiaan Viviers, Peter H. N. de With, and Fons van der Sommen. Symmetrical flow matching: Unified image generation, segmentation, and classification with score-based generative models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):2498–2506, 2026.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical image analysis*, 97:103280, 2024.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299. IEEE, 2022.

Daniele Falcetta, Vincenzo Marciano, Kaiyuan Yang, Jon Cleary, Loic Legris, Massimiliano Domenico Rizzaro, Ioannis Pitsiorlas, Hava Chaptoukaev, Benjamin Lemasson, Bjoern Menze, and Maria A Zuluaga. Vesselverse: A dataset and collaborative framework for vessel annotation. In *Medical Image Computing and Computer Assisted Intervention*, pp. 655–665, 2025.

Yubin Gu, Boyang Hou, Yuan Meng, Wenting Luo, Jiayi Ji, and Xiaoshuai Sun. CrackSSM: Reviving SSMs for crack segmentation via dynamic scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10721–10730, 2026.

Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. *Advances in Neural Information Processing Systems*, 36: 8186–8207, 2023.

Yannick Kirchhoff, Maximilian R Rokuss, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Philipp Vollmuth, Jens Kleesiek, Fabian Isensee, et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In *European Conference on Computer Vision*, pp. 218–234. Springer, 2024.

Qin Lei, Jiang Zhong, and Qizhu Dai. Enriching information and preserving semantic consistency in expanding curvilinear object segmentation datasets. In *European conference on computer vision*, pp. 233–250. Springer, 2024.

Qin Lei, Jiang Zhong, Mianxiong Dong, and Kaoru Ota. Faithful crack image synthesis from evolutionary pixel-level annotations via latent semantic diffusion model. *Expert Systems with Applications*, 275:126986, 2025.

David Lesage, Elsa D Angelini, Isabelle Bloch, and Gareth Funka-Lea. A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical image analysis*, 13(6): 819–845, 2009.

Danfeng Li, Hui Zhang, Sheng Wang, Jiacheng Li, and Zuxuan Wu. Seg2any: Open-set segmentation-mask-to-image generation with precise shape and semantic control. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuqi Lin, Hao Zhang, Wenqi Shao, Shiqu Liu, Zhihong Gu, Wenxiao Wang, Xiaofei He, and Kaipeng Zhang. Matchmask: Mask-centric generative data augmentation for label-scarce semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42126–42136, 2026.

Hui Liu, Chen Jia, Fan Shi, Xu Cheng, and Shengyong Chen. Scsegamba: Lightweight structure-aware vision mamba for crack segmentation in structures. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29406–29416. IEEE, 2025.

Laurin Lux, Alexander H Berger, Alexander Weers, Nico Stucki, Daniel Rueckert, Ulrich Bauer, and Johannes Paetzold. Topograph: An efficient graph-based framework for strictly topology preserving image segmentation. In *International Conference on Learning Representations*, volume 2025, pp. 80207–80231, 2025.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7516–7525. IEEE, 2021.

Peyman Milanfar. A lagrangian view of flow matching. arXiv preprint arXiv:2609.00198v2, 2026.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Lei Mou, Yitian Zhao, Huazhu Fu, Yonghuai Liu, Jun Cheng, Yalin Zheng, Pan Su, Jianlong Yang, Li Chen, Alejandro F. Frangi, Masahiro Akiba, and Jiang Liu. Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. *Medical Image Analysis*, 67:101874, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101874. URL `https://www.sciencedirect.com/science/article/pii/S1361841520302383`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16560–16569. IEEE, 2021.

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.

Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, pp. 32698–32727. PMLR, 2023.

Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. Semflow: Binding semantic segmentation and image synthesis via rectified flow. *Advances in Neural Information Processing Systems*, 37:138981–139001, 2024.

Bastian Wittmann, Yannick Wattenberg, Tamaz Amiranashvili, Suprosanna Shit, and Bjoern Menze. vesselFM: A foundation model for universal 3D blood vessel segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20874–20884, 2025.

Chun-Hung Wu, Shih-Hong Chen, Chih-Yao Hu, Hsin-Yu Wu, Kai-Hsin Chen, Yu-You Chen, Chih-Hai Su, Chih-Kuo Lee, and Yu-Lun Liu. Denver: Deformable neural vessel representations for unsupervised video vessel segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15682–15692. IEEE, 2025.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6030–6038, 2024.

Ziyun Yang and Sina Farsiu. Directional connectivity-based segmentation of medical images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11525–11535. IEEE, 2023.

Lei Zhang, Fan Yang, Yimin Daniel Zhang, and Ying Julie Zhu. Road crack detection using deep convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3708–3712. IEEE, 2016.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Zilong Zhao, Zhengming Ding, Pei Niu, Wenhao Sun, and Feng Guo. MixerCSeg: An efficient mixer architecture for crack segmentation via decoupled Mamba attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17969–17978, 2026.

Jiale Zhou, Wenhan Wang, Shikun Li, Xiaolei Qu, Xin Guo, Yizhong Liu, Wenzhong Tang, Xun Lin, and Yefeng Zheng. Topotta: Topology-enhanced test-time adaptation for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24123–24134, 2025a.

Shanglian Zhou, Carlos Canchila, and Wei Song. Deep learning-based crack segmentation for civil infrastructure: Data types, architectures, and benchmarked performance. *Automation in Construction*, 146:104678, 2023.

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)*, pp. 20852–20862. IEEE, 2025b.

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.

Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE transactions on image processing*, 28 (3):1498–1512, 2018.

Xin Zuo, Yu Sheng, Jifeng Shen, and Yongwei Shan. Topology-aware mamba for crack segmentation in structures. *Automation in Construction*, 168:105845, 2024.