CSWinUNETR: Segmentation of Thin Anatomical Structures in Medical Images

Junho Moon[1 [0009−0004−3522−6357]], Haejun Chung[1✉ [0000−0001−8959−237X]], and Ikbeom Jang[2✉ [0000−0002−6901−983X]]

[1] Hanyang University, Seoul, Republic of Korea
[2] Hankuk University of Foreign Studies, Yongin, Republic of Korea
{jhmoon6807, haejun}@hanyang.ac.kr, ijang@hufs.ac.kr

Abstract. Accurate segmentation of thin, tortuous anatomical structures, such as retinal vessels, cerebral vasculature, and facial wrinkles, remains challenging due to low contrast, frequent discontinuities, and severe class imbalance. Although recent convolutional and Transformer-based models have improved performance, they often yield fragmented predictions and fail to recover fine branches. We propose CSWinUNETR, a task-oriented 2D/3D backbone for thin-structure segmentation. It employs cross-shaped stripe self-attention to model long-range principal-axis context and incorporates cyclic shifts to enhance information exchange across stripes. To better preserve fine-grained details, we further introduce a detail-enhanced multi-scale self-attention module that aggregates contextual features from multi-resolution representations. In addition, we propose sparse-control dynamic snake convolution, which reconstructs reliable dense curvilinear kernels from sparsely predicted control points to better follow tortuous geometry. Extensive experiments on four benchmarks across ophthalmology, neurovascular imaging, and dermatology demonstrate that CSWinUNETR consistently outperforms state-of-the-art methods without task-specific post-processing or topology-aware losses. Code is available at https://github.com/labhai/CSWinUNETR.

Keywords: Thin Structure Segmentation · CSWin Transformer · Multi-Scale Multi-Head Self-Attention · Dynamic Snake Convolution.

1 Introduction

Segmenting thin, tortuous structures, such as retinal vessels, cerebrovascular branches, and facial wrinkles, remains a long-standing challenge across heterogeneous imaging modalities [2,13,19,20,28]. These targets typically occupy only a minute fraction of the image domain, exhibit low contrast and frequent interruptions, and follow highly anisotropic, curvilinear paths, all of which hinder reliable segmentation [5,14,29]. Despite these challenges, precise delineation of such slender structures is crucial for both clinical interpretation and the development of reliable methodological benchmarks [7,28]. However, their extremely

✉ Corresponding authors

small caliber, complex branching geometry, and topological variability often confound standard architectures, leading to fragmented or missing predictions [24].

Despite substantial advances in medical image segmentation, robust and broadly applicable architectures for delineating thin, curvilinear structures remain underexplored across diverse datasets and clinical applications [2,7]. Accurate segmentation of these structures requires two complementary capabilities: (i) orientation-aware long-range information propagation to bridge low-contrast gaps and interruptions [4,29], and (ii) anisotropic, trajectory-aligned integration of local evidence to preserve continuity and recover fine terminal branches [1,22]. Cross-Shaped Window (CSWin) self-attention [4] is well-suited to addressing (i), as its stripe-wise attention efficiently captures long-range, orientation-aware context. Although CSWin-UNet [17] adopts this mechanism for segmentation, it is primarily designed for 2D slice-wise evaluation and does not explicitly condition attention features on fine-grained structural details. For (ii), Dynamic Snake Convolution (DSConv) [22] enables deformable feature aggregation along curvilinear structures. However, its stepwise offset tracking can propagate local prediction errors when structural evidence is weak or corrupted.

We introduce CSWinUNETR, a segmentation backbone designed for thin, tortuous anatomical structures in 2D and 3D medical images across diverse modalities. CSWinUNETR incorporates the above inductive biases in a unified architecture. To support (i), we employ CSWin self-attention [4] with cyclic shift [18] to enable efficient axis-wise global reasoning and derive a detail-enhanced Multi-Scale Multi-Head Self-Attention (MS-MHSA) to better preserve subtle structural cues. To address (ii), we integrate Sparse-control Dynamic Snake Convolution (SDSConv), which aggregates reliable curvilinear features through sparse control-trajectory construction.

Our contributions include: (1) We propose CSWinUNETR, a task-oriented 2D/3D backbone tailored for segmenting thin, tortuous anatomical structures. (2) We introduce shifted CSWin self-attention with multi-scale contextual modeling to enhance long-range dependency reasoning while preserving fine-grained details. (3) We incorporate SDSConv to enable stable, trajectory-aligned local feature aggregation, strengthening geometry-aware feature representations. (4) Extensive experiments on four heterogeneous benchmarks demonstrate consistent improvements over representative baselines, supported by both quantitative and qualitative evaluations.

2 Method

An overview of the proposed method is shown in Fig. 1. Although CSWinUNETR supports both 2D and 3D inputs, we present the method using 2D notation for clarity. The extension to volumetric data is described in Sec. 2.2.

2.1 CSWinUNETR: A Backbone for Thin-Structure Segmentation

A key observation in thin-structure segmentation is that long-range dependencies in curvilinear anatomy tend to propagate predominantly along principal

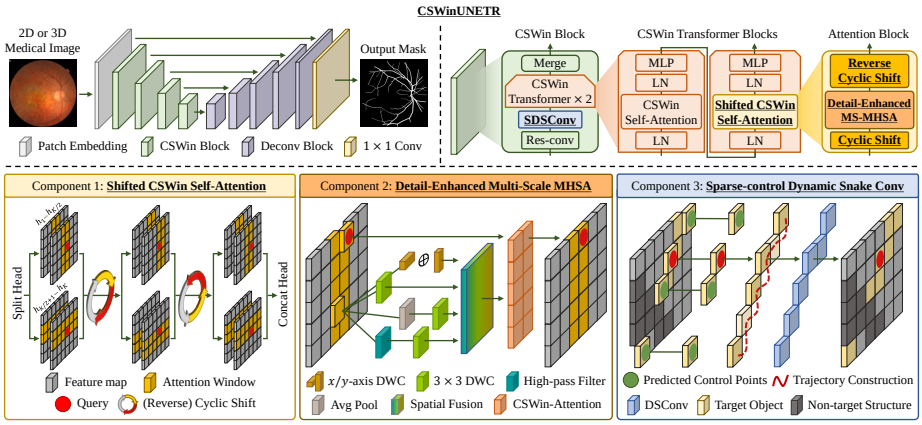

Fig. 1. Overview of the proposed CSWinUNETR for segmenting thin, tortuous anatomical structures in 2D and 3D medical images. The architecture comprises three core components. (1) Shifted CSWin Self-Attention captures orientation-aware long-range dependencies via cross-shaped stripe attention with cyclic shifts. (2) Detail-Enhanced MS-MHSA fuses contextual information across multi-scale features to better preserve fine-grained structural details. (3) SDSConv encourages continuity along tortuous trajectories by constructing reliable, geometry-aligned curvilinear responses from sparsely predicted control points. Newly introduced components are **highlighted**.

directions [2,5]. Accordingly, axis-aligned, elongated context windows provide an effective inductive bias for aggregating non-local cues.

Motivated by this, we adopt the CSWin self-attention [4] to efficiently model long-range context. Given an input medical image $\mathcal{I} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$, we first extract tokens using a convolutional embedding [26] with a 7×7 kernel and stride 2, producing $X^{(0)} \in \mathbb{R}^{C_0 \times \frac{H}{2} \times \frac{W}{2}}$. The encoder comprises four stages of CSWin blocks. For stage index $t \in \{1, 2, 3, 4\}$, the block takes $X^{(t-1)}$ as input, applies a residual convolutional refinement, followed by two consecutive CSWin Transformer blocks, and then downsamples features through a merge layer implemented by a 3×3 convolution with stride 2. This merge operation halves the spatial resolution and doubles the channel dimension, yielding $X^{(t)} \in \mathbb{R}^{C_t \times \frac{H}{2^{t+1}} \times \frac{W}{2^{t+1}}}$ where $C_t = 2^t C_0$. For the decoder, we employ a SwinUNETR-style deconvolutional pathway [10] to progressively upsample and fuse multi-scale encoder features, producing the predicted segmentation mask.

Within each CSWin block, both CSWin Transformer blocks use CSWin self-attention. Specifically, the attention heads are partitioned into two groups corresponding to vertical and horizontal stripes, and self-attention is computed independently within the associated axis-aligned stripe windows. We further apply a cyclic shift [18] in the second CSWin Transformer block to alleviate stripe-wise isolation and boundary artifacts. With this shift, tokens near stripe boundaries are reassigned to different stripe neighborhoods, facilitating cross-stripe interactions and improving information exchange between adjacent stripes.

While stripe attention is well-suited for orientation-aware long-range aggregation, fine-grained cues of thin structures can be attenuated by stage-wise downsampling. To better preserve high-frequency details while integrating multi-scale context, we propose a detail-enhanced MS-MHSA module, inspired by multi-branch attention designs [6]. Given an input feature map $X^{(t)}$, we construct several lightweight depthwise convolution (DWC) branches that operate in parallel. The axial strip branch applies a separable strip DWC along the horizontal and vertical axes and sums the two responses to reduce over-smoothing. The local-refinement branch captures regional patterns using a 3×3 DWC, and the semantic-context branch models coarser context by applying average pooling (AvgPool) followed by a 3×3 DWC. In addition, we incorporate a high-pass detail branch to explicitly enhance fine structures by computing $X^{(t)} - \mathrm{AvgPool}_{3\times3}(X^{(t)})$ and refining the resulting features with a 3×3 DWC.

Following multi-scale enhancement, we employ a spatial fusion gate to adaptively combine the branch outputs. Specifically, a 1×1 convolution predicts per-location branch logits, which are normalized by a softmax across branches to obtain spatially varying fusion weights. The fused feature map $\widetilde{X}^{(t)}$ is then computed as the weighted sum of the branch responses. Next, we construct the attention triplet (Q_a, K_a, V_a) using pointwise linear projections at each spatial location. The Q_a is projected from the original feature map $X^{(t)}$, whereas the K_a and V_a are projected from the fused feature map $\widetilde{X}^{(t)}$. Given (Q_a, K_a, V_a), stripe attention for each orientation branch $a \in \{\mathrm{v}, \mathrm{h}\}$ is computed as:

$$O_a = \mathrm{Softmax}\left(\frac{Q_a K_a^\top}{\sqrt{d}} + M_a\right) V_a + \mathrm{LePE}(V_a), \tag{1}$$

where d denotes the per-head feature dimension, M_a represents the (shifted) stripe attention mask, and $\mathrm{LePE}(\cdot)$ is a local positional encoding implemented by a 3×3 DWC.

2.2 Sparse-control Dynamic Snake Convolution

Segmenting thin, tortuous anatomy requires anisotropic evidence aggregation that can follow curved trajectories [1,5]. DSConv [22] partially addresses this need by deforming an axial convolutional kernel using location-adaptive offsets. However, DSConv relies on cumulative offset accumulation along the sampling path, which can amplify local prediction errors and propagate them through the entire trajectory. This issue is exacerbated when a small segment of the target structure is corrupted by non-target factors (e.g., low-contrast gaps, imaging artifacts, or occlusions), in which case subsequent sampling locations may progressively drift away from the true anatomy.

To obtain a more stable curvilinear operator, we propose SDSConv, which predicts sparse control points and constructs a dense snake-like kernel via trajectory construction, avoiding cumulative drift. Given the feature map $F = X^{(t)}$ at encoder level t, let $\mathbf{u} = (y, x)$ denote a spatial location. SDSConv constructs a curvilinear sampling kernel of odd length $P = 2c + 1$ over a centered uniform

axial support $p_k = k$, $k \in \{-c, \ldots, c\}$, with $\xi_k = p_k/c \in [-1, 1]$. Instead of regressing P independent offsets at each $\mathbf{u}$, SDSConv predicts branch-wise lateral control offsets $\{\Delta_m^a(\mathbf{u})\}_{m=1}^M$ for $a \in \{\mathrm{v}, \mathrm{h}\}$, and branch-shared control locations $\{\mu_m(\mathbf{u})\}_{m=1}^M \subset [-1, 1]$ along the axial coordinate. The control locations are parameterized as bounded residual shifts from the base locations. Dense branch-wise lateral offsets are obtained by a normalized Gaussian radial basis function (RBF) blending:

$$o_a(k; \mathbf{u}) = \sum_{m=1}^{M} \alpha_{m,k}(\mathbf{u}) \Delta_m^a(\mathbf{u}), \quad \alpha_{m,k}(\mathbf{u}) = \frac{\exp\Big(-\frac{(\xi_k - \mu_m(\mathbf{u}))^2}{2\sigma^2}\Big)}{\sum_{j=1}^{M} \exp\Big(-\frac{(\xi_k - \mu_j(\mathbf{u}))^2}{2\sigma^2}\Big)}, \tag{2}$$

where $\sigma > 0$ denotes the RBF bandwidth. Here, o_{v} is the x-direction lateral offset for vertical traversal and o_{h} is the y-direction lateral offset for horizontal traversal. We further apply center anchoring with a learnable distance-ramped scope:

$$\hat{o}_a(k; \mathbf{u}) = s_a(|\xi_k|)\big(o_a(k; \mathbf{u}) - o_a(0; \mathbf{u})\big), \tag{3}$$

where $s_a(\cdot) > 0$ is a bounded monotonic ramp learned for each axial branch. This guarantees $\hat{o}_a(0; \mathbf{u}) = 0$ and removes the constant lateral offset component, suppressing global drift. The vertical and horizontal trajectories are then defined as:

$$\mathbf{g}_{\mathrm{v}}(k; \mathbf{u}) = \big(y + p_k,\ x + \hat{o}_{\mathrm{v}}(k; \mathbf{u})\big), \quad \mathbf{g}_{\mathrm{h}}(k; \mathbf{u}) = \big(y + \hat{o}_{\mathrm{h}}(k; \mathbf{u}),\ x + p_k\big). \tag{4}$$

The feature map F is sampled along these trajectories using bilinear interpolation, and the P sampled features are aggregated by axis-aligned DWC. The resulting vertical and horizontal responses are fused and injected into the CSWin blocks as a curvilinear residual prior. This sparse, non-cumulative construction generates dense curvilinear kernels in a fully parallel manner, mitigating error accumulation under unreliable local evidence.

For volumetric inputs, CSWin self-attention uses stripes oriented along the x-, y-, and z-axes, with cyclic shifts applied across all spatial axes. SDSConv is extended to three axial branches. Each branch traverses one principal axis and uses the two orthogonal components to construct a dense 3D trajectory. The predicted sparse control offsets are expanded via RBF weighting to generate dense trajectories, which are used to sample volumetric features through trilinear interpolation. The sampled features are aggregated using axis-aligned 3D DWC.

2.3 Optimization and Implementation Details

To assess generality, we adopt the same network architecture across all datasets, modifying only settings that depend on input dimensionality when switching between 2D and 3D. After convolutional tokenization, the base embedding dimension is fixed to $C_0 = 48$. The number of attention heads is set to $(2, 4, 8, 16)$ for 2D inputs and $(3, 6, 12, 24)$ for 3D inputs. For the SDSConv, the kernel length P is set to $(11, 9, 7, 5)$ from shallow to deep stages, and the number of predicted sparse control points is defined as $M = \lfloor P/2 \rfloor$. Model training is performed using a composite objective combining Dice and cross-entropy losses.

Table 1. Quantitative comparison on 2D thin-structure segmentation.

Method	FIVES (2D)				FFHQ-Wrinkle (2D)				#P
	Dice↑	clDice↑	β↓	HD95↓	Dice↑	clDice↑	β↓	HD95↓	
nnUNetv2 [12]	89.26*	87.96*	37.62*	33.75*	63.27*	70.69*	**11.60**	71.89*	
+clDice [24]	89.88*	90.14*	36.63*	29.98*	64.14*	70.87*	12.14*	66.26*	
+cbDice [23]	89.40*	88.83*	36.78*	35.69*	61.96*	67.47*	15.53*	73.79*	46M
+ske-recall [15]	89.42*	89.66*	35.62*	30.75*	64.40	71.90*	12.09*	65.78*	
+GLCP [29]	88.56*	89.01*	48.95*	39.09*	62.91*	70.25*	12.03*	65.88*	
UNETR [9]	87.65*	88.50*	56.02*	36.18*	55.62*	62.81*	17.79*	99.96*	88M
SwinUNETR [8]	89.35*	90.25*	41.46*	26.96	62.64*	70.28*	12.25*	65.75*	25M
SwinUNETRv2 [10]	89.90*	90.20*	39.37*	27.18*	63.57*	71.48	11.83	65.89*	29M
IncepFormer [6]	89.57*	89.89*	56.27*	28.59*	62.46*	69.69*	11.70	63.43*	15M
DSCNet [22]	88.22*	88.46*	63.02*	30.43*	61.40*	69.47*	11.95*	66.12*	2M
CSWin-UNET [17]	89.80*	90.83	34.34	27.00	62.57*	70.00*	11.91	63.18	24M
CS^2-Net [21]	87.69*	88.75*	60.66*	36.21*	59.91*	67.54*	12.10*	72.18*	8M
SGAT-Net [16]	88.95*	89.42*	42.68*	33.43*	55.27*	55.81*	18.03*	94.82*	8M
TMP+U-Net [20]	86.64*	87.17*	64.56*	37.67*	61.44*	68.67*	12.27*	85.11*	17M
CSWinUNETR (Ours)	**90.71**	**90.88**	**33.07**	**25.62**	**64.75**	**72.18**	11.67	**61.76**	31M

Table 2. Quantitative comparison on 3D thin-structure segmentation.

Method	TopCoW-MRA (3D)				TopCoW-CTA (3D)				#P
	Dice↑	clDice↑	β↓	HD95↓	Dice↑	clDice↑	β↓	HD95↓	
nnUNetv2 [12]	80.84*	89.45*	0.74*	7.78*	75.23*	83.89*	0.56*	8.14*	
+clDice [24]	82.65*	91.64*	0.68*	6.64*	76.14*	84.75*	0.48*	7.89*	
+cbDice [23]	82.49*	91.32*	0.66	6.85*	76.43*	86.35*	0.45*	7.35	31M
+ske-recall [15]	81.94*	91.28*	0.71*	7.46*	75.86*	82.64*	0.54*	8.02*	
+GLCP [29]	81.78*	90.68*	0.77*	6.89*	74.69*	86.68	0.59*	7.64*	
UNETR [9]	78.80*	84.56*	0.94*	7.55*	65.84*	73.13*	0.92*	9.48*	93M
SwinUNETR [8]	80.98*	84.54*	0.71*	7.69*	72.48*	82.40*	0.62*	6.89	62M
SwinUNETRv2 [10]	79.96*	85.38*	0.74*	7.54*	72.87*	84.29*	0.44	8.18*	73M
DSCNet [22]	82.49*	90.55*	0.67	6.16	73.32*	83.53*	0.69*	8.45*	8M
CS^2-Net [21]	80.30*	84.54*	0.98*	6.51*	73.38*	84.32*	0.53*	7.77*	23M
ER-Net [27]	73.86*	78.13*	0.87*	8.76*	73.50*	83.46*	0.91*	8.73*	5M
CSWinUNETR (Ours)	**84.74**	**93.55**	**0.58**	**4.72**	**77.06**	**90.71**	**0.36**	**6.11**	81M

3 Experiments

3.1 Datasets and Evaluation metrics

We evaluate four benchmarks for thin, tortuous-structure segmentation across ophthalmology, neurovascular imaging, and dermatology. FIVES [13] comprises 800 fundus images at 2048×2048 resolution with binary vessel annotations, split into 500/100/200 images for training/validation/testing. TopCoW [28] includes paired 3D CTA and MRA Circle-of-Willis volumes with multiclass annotations for 13 vessel classes. We use the public 125-patient cohort and adopt a 100/12/13 patient-level split for training/validation/testing. FFHQ-Wrinkle [20] provides binary wrinkle masks for 1,000 facial images at 1024×1024 resolution, with an 800/100/100 training/validation/testing split. Segmentation performance is measured using overlap metrics (Dice [3] and clDice [24]), topology error (Betti

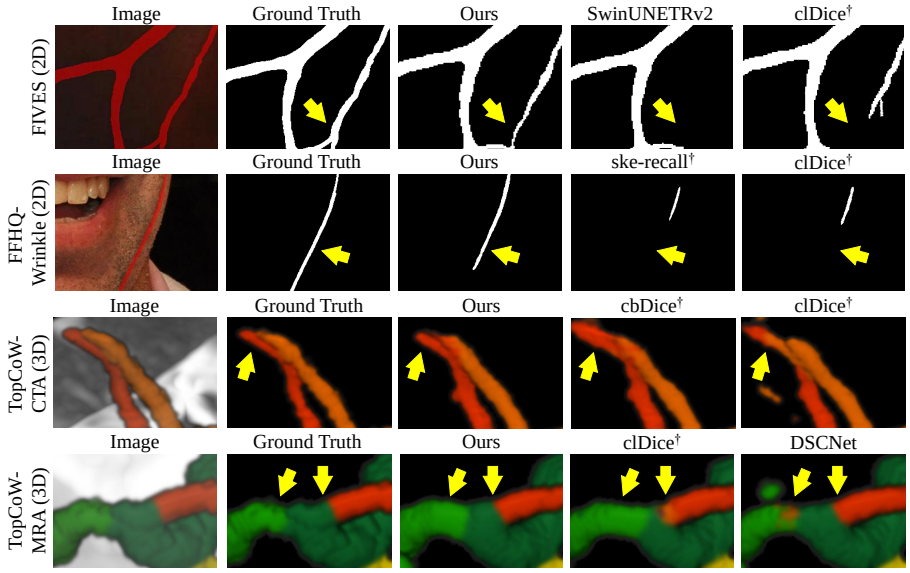

Fig. 2. Qualitative comparison of 2D and 3D thin-structure segmentation. From left to right: the input image overlaid with the ground truth, the ground truth, our CSWin-UNETR prediction, and the second- and third-ranked methods by Dice score (method names are shown above). † denotes methods trained with an nnUNetv2 framework. Yellow arrows highlight regions with the largest discrepancies.

Error [11], $\beta = \beta_0 + \beta_1$), and boundary distance error (95th-percentile Hausdorff Distance; HD95 [25]).

3.2 Comparison Results and Ablation Study

To evaluate CSWinUNETR on thin-structure segmentation, we first consider general-purpose medical segmentation backbones, including nnUNetv2 [12], UNETR [9], SwinUNETR [8], and SwinUNETRv2 [10]. We then include methods that incorporate components closely related to our design, namely InceptFormer [6], DSCNet [22], and CSWin-UNET [17]. We also compare with thin-structure-oriented methods, including CS^2-Net [21], SGAT-Net [16], Texture-Map Pretraining (TMP)+U-Net [20], GLCP [29], and ER-Net [27]. In addition, we retrain nnUNetv2 using clDice [24], cbDice [23], and skeleton-recall (ske-recall) loss [15]. For statistical analysis, we use a paired t-test with the Benjamini–Hochberg correction. In all tables, statistical significance is indicated by * ($p < 0.05$), and #P denotes the number of trainable parameters. The best and second-best results are highlighted in **bold** and underlined, respectively.

Tables 1 and 2 indicate that CSWinUNETR consistently outperforms the competing methods across all four benchmarks. Notably, the observed improvements are statistically significant ($p < 0.05$) for most metrics relative to the other

Table 3. Ablation study of the proposed core components. "Cyc" denotes the cyclic shift, "MS" is the detail-enhanced MS-MHSA, and "SDSC" represents the SDSConv.

Method			TopCoW-MRA (3D)					FFHQ-Wrinkle (2D)				
Cyc	MS	SDSC	Dice↑	clDice↑	β↓	HD95↓	#P	Dice↑	clDice↑	β↓	HD95↓	#P
			82.81	91.93	0.77	6.76	79.9M	64.09	71.55	11.86	63.89	30.5M
✓			83.39	92.28	0.74	5.91	79.9M	64.24	71.78	11.84	63.61	30.5M
✓	✓		83.96	92.77	0.64	5.16	80.0M	64.35	71.94	11.79	63.29	30.6M
✓	✓	✓	**84.74**	**93.55**	**0.58**	**4.72**	80.8M	**64.75**	**72.18**	**11.67**	**61.76**	31.1M

approaches. The qualitative comparisons in Fig. 2 further corroborate these findings, showing that CSWinUNETR more faithfully reconstructs fine branches and terminal structures, while reducing false positives in non-target regions. We further evaluate the deployment cost on TopCoW-MRA using a 96^3 input patch under identical inference settings. Specifically, DSCNet, SwinUNETRv2, and CSWinUNETR require 4.23/1.51/1.35 GB of memory, 425/346/346 GFLOPs, and 321/250/292 ms of per-patch latency, respectively. These results indicate that CSWinUNETR achieves competitive computational efficiency despite its larger parameter count. Table 3 shows that incorporating cyclic shift, MS-MHSA, and SDSConv yields consistent and stepwise improvements across all evaluation metrics. In particular, the proposed components improve performance with only marginal parameter overhead, suggesting that the observed gains primarily arise from the architectural design rather than increased model capacity.

4 Discussion and Conclusion

The proposed CSWinUNETR consistently outperforms strong baselines across four heterogeneous benchmarks that emphasize fine-grained and tortuous anatomical structures. Notably, these benchmarks span visually and modality-diverse targets, from vascular segmentation to facial wrinkle delineation. Despite this diversity, we evaluate a single end-to-end model using a largely fixed architecture and training protocol across all datasets. The improvements observed under this unified setting suggest that the performance gains are more likely attributable to the model's architectural inductive biases than to dataset-specific hyperparameter tuning. Taken together, these results support CSWinUNETR as a practical backbone for thin-structure segmentation across diverse imaging modalities.

The qualitative results further indicate that CSWinUNETR better preserves structural continuity while reducing spurious predictions outside the target anatomy. We attribute these improvements to the combined effects of orientation-aware long-range contextual modeling and trajectory-aligned feature aggregation. In particular, SDSConv constructs curvilinear sampling trajectories from sparse control points and aggregates features along anatomically plausible paths. By integrating consistent evidence across these trajectories, this mechanism may restore anatomical continuity in thin, tortuous structures that appear fragmented due to limited spatial resolution, motion, or acquisition artifacts. Consequently, the more consistently connected segmentations produced by CSWinUNETR may

improve the reliability of downstream centerline-based analyses, including the estimation of anatomical length, branching patterns, and tortuosity.

To conclude, CSWinUNETR is a unified segmentation architecture for thin, tortuous anatomical structures across heterogeneous medical imaging modalities, encompassing both 2D images and 3D volumes. Extensive experiments demonstrate that CSWinUNETR consistently outperforms existing methods, including approaches tailored to thin-structure segmentation. These results suggest that CSWinUNETR offers a principled framework for geometry-aware attention and robust curvilinear feature aggregation, enabling reliable dense prediction in challenging curvilinear segmentation scenarios.

Acknowledgments. This work was supported by NRF (RS-2024-00338048, RS-2024-00455720, RS-2026-25487796), MOTIE Technology Innovation Program (Biotechnology) (RS-2025-13002970), National Institute of Health (2026-ER0901-00, 2026-ER0904-00), Hankuk University of Foreign Studies Research Fund of 2025, KOCCA R&D Program (RS-2024-00332210), IITP AI Graduate School Program (RS-2020-II201373, Hanyang University), IITP AI Semiconductor Support Program (RS-2023-00253914), and AI Seoul Tech Research Support Program of Seoul Future Foundation.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Benmansour, F., Cohen, L.D.: Tubular structure segmentation based on minimal path method and anisotropic enhancement. International Journal of Computer Vision **92**(2), 192–210 (2011)
2. Bibiloni, P., González-Hidalgo, M., Massanet, S.: A survey on curvilinear object segmentation in multiple applications. Pattern Recognition **60**, 949–970 (2016)
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
4. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12124–12134 (2022)
5. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: International conference on medical image computing and computer-assisted intervention. pp. 130–137. Springer (1998)
6. Fu, L., Tian, H., Zhai, X.B., Gao, P., Peng, X.: Incepformer: efficient inception transformer with pyramid pooling for semantic segmentation. arXiv preprint arXiv:2212.03035 (2022)
7. Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., Wang, J.: Medical image segmentation: A comprehensive review of deep learning-based methods. Tomography **11**(5), 52 (2025)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)

9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
10. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
11. Hu, X., Li, F., Samaras, D., Chen, C.: Topology-preserving deep image segmentation. Advances in neural information processing systems **32** (2019)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific data **9**(1), 475 (2022)
14. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: International conference on medical imaging with deep learning. pp. 285–296. PMLR (2019)
15. Kirchhoff, Y., Rokuss, M.R., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Vollmuth, P., Kleesiek, J., Isensee, F., 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)
16. Lin, J., Huang, X., Zhou, H., Wang, Y., Zhang, Q.: Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images. Medical Image Analysis **89**, 102929 (2023)
17. Liu, X., Gao, P., Yu, T., Wang, F., Yuan, R.Y.: Cswin-unet: Transformer unet with cross-shaped windows for medical image segmentation. Information Fusion **113**, 102634 (2025)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Moon, I.J., Moon, J., Jang, I.: Weakly supervised pretraining and multi-annotator supervised finetuning for facial wrinkle detection. arXiv preprint arXiv:2408.09952 (2024)
20. Moon, J., Chung, H., Jang, I.: Facial wrinkle segmentation for cosmetic dermatology: Pretraining with texture map-based weak supervision. In: International Conference on Pattern Recognition. pp. 319–334. Springer (2024)
21. Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al.: Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. Medical image analysis **67**, 101874 (2021)
22. Qi, Y., He, Y., Qi, X., Zhang, Y., Yang, G.: Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6070–6079 (2023)
23. Shi, P., Hu, J., Yang, Y., Gao, Z., Liu, W., Ma, T.: Centerline boundary dice loss for vascular segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2024)

24. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16560–16569 (2021)
25. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**(1), 29 (2015)
26. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22–31 (2021)
27. Xu, W., Yang, H., Shi, Y., Tan, T., Liu, W., Pan, X., Deng, Y., Gao, F., Su, R.: Ernet: edge regularization network for cerebral vessel segmentation in digital subtraction angiography images. IEEE Journal of Biomedical and Health Informatics **28**(3), 1472–1483 (2023)
28. Yang, K., Musio, F., Ma, Y., Juchler, N., Paetzold, J.C., Al-Maskari, R., Höher, L., Li, H.B., Hamamci, I.E., Sekuboyina, A., et al.: Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. ArXiv pp. arXiv–2312 (2025)
29. Zhou, F., Gao, Z., Zhao, H., Xie, J., Meng, Y., Zhao, Y., Lip, G.Y., Zheng, Y.: Glcp: Global-to-local connectivity preservation for tubular structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 237–247. Springer (2025)