WAVEJOINT: FREQUENCY-GUIDED SCALE TRANSITIONS FOR OCT ANGIOGRAPHY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optical coherence tomography angiography (OCTA) supports two complementary biomarkers: the retinal vessel (RV) tree and the foveal avascular zone (FAZ). The capillaries of the RV tree span only 1–3 pixels and their decorrelation signals are weak, whereas the FAZ is a compact region whose accurate delineation depends on stable contextual information. Jointly segmenting both structures from a single en-face projection is desirable for clinical quantification, but the two targets place conflicting demands on the network. We present WaveJoint, a compact (12.4M-parameter) dual-task Transformer and an OCTA joint-segmentation architecture built around an explicit frequency decomposition. At each encoder scale transition, a wavelet-guided downsampling module (WaveDown) applies fixed orthogonal Haar filters, which are parameter-free and satisfy perfect reconstruction, to compute directional detail through a 2-D discrete wavelet transform and to reweight the downsampled path before aliasing suppresses high-frequency detail. Two task-specific heads then predict the RV and FAZ masks under a balanced Dice–BCE objective. A 2×2 factorial ablation over three seeds with interaction terms attributes the model's fine-vessel advantage to this downsampling path: this path shows a 2.9 percentage-point mean main effect on capillary-scale recall, significant in every seed, and is the only module whose effect is stable in direction. The same analysis explains why single-module ablations judged on aggregate Dice alone can suggest, incorrectly, that no module is important. We also contribute a structural evaluation protocol comprising skeleton recall stratified by vessel caliber, HD95 and ASSD boundary distances, a fragmentation ratio, and Holm-corrected paired Wilcoxon tests. Under this protocol, WaveJoint exceeds a protocol-matched U-Net by 3.25 percentage points of capillary recall with 48% lower fragmentation; this structural difference compresses to 2.0 percentage points of RV Dice. On the official OCTA-500 3 mm split, WaveJoint attains an RV Dice of 0.9220 ± 0.0002 and an FAZ Dice of 0.9820 ± 0.0013 over three seeds; its RV Dice matches the best published joint-segmentation systems, while using a single 2-D input and no pretraining. The structural advantage persists under shifts in field of view, device, and population (OCTA-500 6 mm; FAROS in-domain), with capillary-recall gains between 7.8 and 12.5 percentage points. These results indicate that OCTA segmentation should report structural metrics and that the contribution of an architectural mechanism should be localized rather than presumed.

1 INTRODUCTION

Optical coherence tomography angiography (OCTA) measures decorrelation between repeated B-scans to image the living retinal microvasculature without contrast injection (Spaide et al., 2018). Clinicians routinely quantify two biomarkers: the **retinal vessel (RV)** tree, whose density and caliber distribution index microvascular health, and the **foveal avascular zone (FAZ)**, whose enlargement and irregular shape are early signs of diabetic retinopathy and other maculopathies (Spaide et al., 2018; Li et al., 2024). Manual delineation is slow and variable, so automatic segmentation is a precondition for OCTA quantification at scale.

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.

The two structures are anatomically coupled—the FAZ is the avascular region bounded by the innermost capillary ring—so their masks are mutually constraining, motivating *joint* architectures that predict both from a shared representation (Hu et al., 2022; Jiang et al., 2025). But the two targets stress a network oppositely. Capillaries span 1–3 pixels with weak decorrelation signals and are destroyed precisely by the operations that build receptive field—strided downsampling and global attention pooling; the FAZ is a compact, smooth-contoured region that depends on stable mid-level context. Downsample aggressively and lose the capillaries, or conservatively and lose global structure.

The dilemma has a natural frequency-domain reading: fine vessels and edges concentrate in high-frequency directional subbands; broad background and lumens in the low-frequency approximation; the FAZ contour spreads across both. The discrete wavelet transform (DWT) (Li et al., 2021) separates these in one reversible, alias-free operation. Yet existing OCTA joint-segmentation architectures are spatial designs—dual decoders with mutual constraints (Hu et al., 2022), 3-D projection learning (Li et al., 2020; Jiang et al., 2025), foundation-model prompting (Chen et al., 2025)—none, to our knowledge, protects high-frequency content at the moment it is most fragile, the scale transition.

This paper targets that design point. We introduce WaveJoint, a dual-task architecture placing an explicit, fixed-filter wavelet decomposition at every scale transition, and we pose—and answer—two questions a methods paper should answer: *which* component carries the gain, and is the gain *visible* to the metric the field habitually reports? Our contributions:

1. **WaveJoint (architecture).** We propose WaveJoint, a compact (12.4M-parameter) dual-task Transformer for joint RV and FAZ segmentation from a single 2-D OCTA(FULL) projection, building an explicit frequency decomposition into every scale transition: all wavelet operations use a shared set of fixed, parameter-free orthogonal Haar filters—orthogonal by construction and perfectly reconstructing—which protect high-frequency fine-vessel detail before aliasing suppresses it. Two task-independent prediction heads trained under a balanced Dice–BCE objective enable dual-task inference at half the cost of dedicated single-task models.
2. **Mechanism-level localization of the gain.** A 2×2 factorial ablation over three seeds (attention $\times$ downsampling, with interaction terms) attributes the fine-vessel advantage to the wavelet-guided downsampling path (WaveDown): a mean $+2.9$ pp main effect on capillary recall ($+1.3$/$+3.3$/$+4.2$ pp, significant in every seed) and the only factor whose effect is significant with a consistent direction. Across all configurations, mean Dice stays within a 0.18 pp band, and the conclusion also holds under the deployed fixed-Haar setting.
3. **A structural evaluation protocol.** We assemble caliber-stratified skeleton recall, HD95/ASSD boundary distances, a fragmentation ratio, and Holm-corrected paired non-parametric tests on topology-aware precedents (Shit et al., 2021) into a standard OCTA reporting bundle, revealing a $+3.25$ pp capillary-recall gap and 48% lower fragmentation against a protocol-matched U-Net—differences Dice poorly expresses.
4. **Cross-dataset generalization.** Three dual-labeled (RV+FAZ) datasets along two deployment axes: OCTA-500 6 mm (field-of-view shift; $+12.5$ pp zero-shot and $+7.8$ pp in-domain capillary recall) and FAROS (device/population shift; $+8.1$ pp in-domain recall, $2.5\times$ lower fragmentation). On FAROS zero-shot, dense semi-automatic labels floor both models' capillary recall; the surviving edges are FAZ Dice and fragmentation (Section 4.7).

2 Related work

OCTA biomarker segmentation. Since OCTA-500 (Li et al., 2024) enabled reproducible benchmarking, U-Net-style encoder–decoder CNNs (Ronneberger et al., 2015) have been the default backbone, later joined by Transformer designs such as OCT2Former's dynamic token aggregation (Tan et al., 2023). For *joint* segmentation, Joint-Seg (Hu et al., 2022) casts FAZ and RV as mutually constraining tasks in a single-encoder, dual-decoder framework; the 3-D line runs from IPN's pooled volume encoding (Li et al., 2020) to PL-Joint-Seg's dual-path volumetric fusion (Jiang et al., 2025); SAM-OCTA (Chen et al., 2025) fine-tunes a prompted foundation model (Kirillov et al., 2023). These surveyed systems are spatial designs: we are not aware of one that uses an explicit frequency

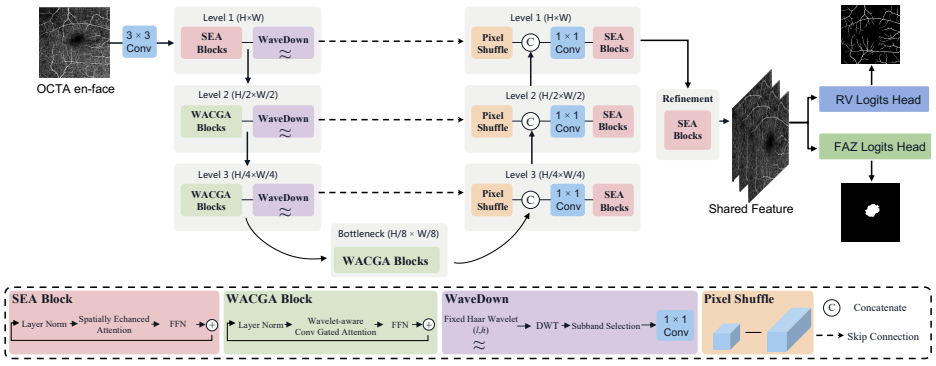

Figure 1: WaveJoint architecture. Level 1 uses SEA; levels 2–3 and the bottleneck use WACGA; every encoder stage ends with WaveDown, whose fixed-Haar analysis reweights the downsampled path. One shared fixed Haar bank (orthogonal, perfectly reconstructing, zero parameters); two parameter-independent heads map the shared feature F to RV and FAZ logits.

decomposition to protect high-frequency structure at the scale transitions where 1–3-pixel capillaries are lost, and the strongest depend on 3-D input or a prompted backbone.

Wavelets, aliasing, and structural evaluation. The DWT gives a multiresolution, invertible decomposition into an approximation and three directional detail subbands (Li et al., 2021). Deep networks use wavelets as lossless downsamplers, frequency splitters, and token mixers: Wave-ViT (Yao et al., 2022) unifies wavelets with Transformer attention, and **LWTformer** (Ruan et al., 2026) couples wavelet-aware gated attention with a wavelet-guided downsampler using *learnable shared filters*, targeting ancient character image restoration. Unlike LWTformer, our architecture adopts fixed, parameter-free Haar filters for all wavelet computations in the small-data OCTA segmentation setting, and further supports joint RV–FAZ dual-task learning. A parallel line attacks aliasing directly: BlurPool (Zhang, 2019) low-pass filters before subsampling for shift invariance, removing high frequencies to protect the coarse grid—whereas WaveDown keeps a lossless downsampled path and uses the DWT detail as evidence to reweight it (Section 3.4). For tubular structures, clDice's skeleton-overlap score (Shit et al., 2021) is the key topology-aware precedent; our protocol assembles caliber-stratified recall, boundary distances, fragmentation, and paired Holm tests into a standard OCTA bundle (Section 4.3), within the U-shaped Transformer line of Restormer, TransNeXt, TransUNet, and Swin-Unet (Zamir et al., 2022; Shi, 2024; Chen et al., 2024; Cao et al., 2022).

3 Method

3.1 Overview

WaveJoint is a U-shaped network (Figure 1): a three-level encoder, bottleneck, three-level decoder, and refinement stage. The input is a single-channel OCTA en-face projection $X \in \mathbb{R}^{1\times 304\times 304}$, embedded to $C = 32$ channels by an overlapping 3×3 convolution. Each encoder stage is attention blocks followed by WaveDown (Section 3.4): level 1 uses spatially enhanced attention (SEA, Section 3.5) at full resolution; levels 2–3 and the bottleneck use **wavelet-aware convolutional gated attention** (WACGA, Section 3.3). The decoder pixel-shuffle upsamples, concatenates skips, reduces channels with 1×1 convolutions, and refines with SEA blocks (four more form the refinement stage). Two **parameter-independent** convolutional heads map the shared 64-channel feature to RV and FAZ logits. Every WACGA and WaveDown shares one fixed Haar pair $(\mathbf{l}, \mathbf{h})$: zero wavelet parameters, with the by-construction guarantees of Section 3.2.

3.2 Shared fixed-filter 2-D discrete wavelet transform

Let $X \in \mathbb{R}^{C\times H\times W}$ be a feature map and $\mathbf{l}, \mathbf{h} \in \mathbb{R}^K$ the fixed Haar filters $\mathbf{l} = \frac{1}{\sqrt{2}}(1, 1)$, $\mathbf{h} = \frac{1}{\sqrt{2}}(1, -1)$, frozen throughout training. The four separable 2-D analysis kernels are the outer products $\Phi_a = \mathbf{l}\mathbf{l}^\top$, $\Phi_h = \mathbf{h}\mathbf{l}^\top$, $\Phi_v = \mathbf{l}\mathbf{h}^\top$, $\Phi_d = \mathbf{h}\mathbf{h}^\top$; one 2-D DWT level is a stride-2 grouped convolution with these kernels, yielding the approximation and three directional detail subbands $Y_s \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$, $s \in \{a, h, v, d\}$ (exact convolution form in Appendix D).

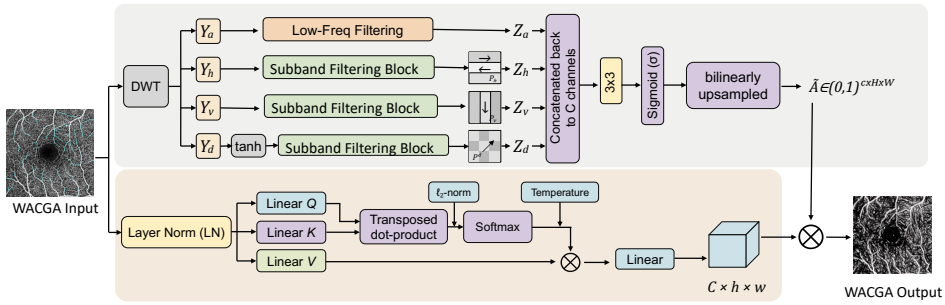

Figure 2: WACGA. The wavelet branch processes the four DWT subbands with structure-matched operators (Table 4) and fuses them into a sigmoid gate $\tilde{A}$; in parallel, MDTA computes channel attention O. The block output is $O \odot \tilde{A}$, wrapped in pre-norm residuals with a GDFN (Eq. (1)).

Three properties make fixed Haar right for the small-data OCTA regime, all by construction: **(i) orthogonality** ($\langle \mathbf{l}, \mathbf{h} \rangle = 0$, so subbands stay decorrelated—a guarantee learnable filters lack); **(ii) perfect reconstruction** (lossless and invertible); **(iii) zero parameters** (one shared fixed bank, no filter-source seed variance). The decomposition also matches target physics: capillary walls and edges project onto Y_h/Y_v; inter-capillary noise and bifurcations onto Y_d; lumens, background, and FAZ interior onto Y_a (full correspondence in Appendix D, Table 4).

3.3 Wavelet-aware convolutional gated attention (WACGA)

WACGA (Figure 2) replaces every attention block at encoder levels 2–3 and the bottleneck, fusing wavelet analysis and channel attention in one module. Given $X \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, two branches run in parallel.

Wavelet branch. The four subbands pass through content-matched operators—a 1×1 projection plus depthwise convolution for the approximation; grouped anisotropic convolutions with Prewitt-initialized directional kernels for h/v detail; $\tanh$ bounding with a Laplacian-initialized kernel for d detail (Appendix E; correspondence in Appendix D, Table 4)—each compressing $C \to C/4$ and concatenated back to C. The fusion gate is $A = \sigma(\mathrm{DWConv}_{3\times3}(Z))$, bilinearly upsampled to $\tilde{A} \in (0,1)^{C \times H \times W}$.

Attention branch and gating. In parallel, Restormer-style MDTA (Zamir et al., 2022) computes channel attention $O = \mathrm{softmax}\big((\hat{Q}\hat{K}^\top)\tau\big)V$ ($\hat{Q}, \hat{K}$ ℓ_2-normalized, τ learnable) at cost linear in $H \times W$, and $WACGA(X) = O \odot \tilde{A}$, wrapped in pre-norm residuals with a gated DConv feed-forward network (GDFN) (Zamir et al., 2022):

$$X' = X + WACGA(\mathrm{LN}(X)), \qquad X'' = X + \mathrm{GDFN}(\mathrm{LN}(X')), \tag{1}$$

the SEA blocks use the conventional form (Appendix E).

Design intent. MDTA answers *which channels* matter, but its spatially uniform aggregation dilutes fine structures, which contribute few positions to a channel's global statistics. The wavelet gate answers the complementary question—*where* the evidence sits and in *which band*—suppressing attention output where the detail subbands carry no vessel evidence. Its contribution and interaction with WaveDown are quantified in Section 4.6.

3.4 Wavelet-guided downsampling (WaveDown)

Every encoder stage ends with WaveDown (Figure 3), localized as the load-bearing wavelet component in Section 4.6. For stage output $X \in \mathbb{R}^{C \times H \times W}$, a 3×3 convolution to $C/2$ channels plus pixel unshuffle gives the main path $U \in \mathbb{R}^{2C \times \frac{H}{2} \times \frac{W}{2}}$; in parallel the shared-filter DWT yields four concatenated subbands $S \in \mathbb{R}^{4C \times \frac{H}{2} \times \frac{W}{2}}$ through a subband-attention branch (SAB: depthwise conv, group norm, GELU, 1×1 conv $4C \to 2C$, sigmoid):

$$A_{\mathrm{down}} = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{GELU}(\mathrm{GN}(\mathrm{DWConv}_{3\times3}(S))))), \tag{2}$$

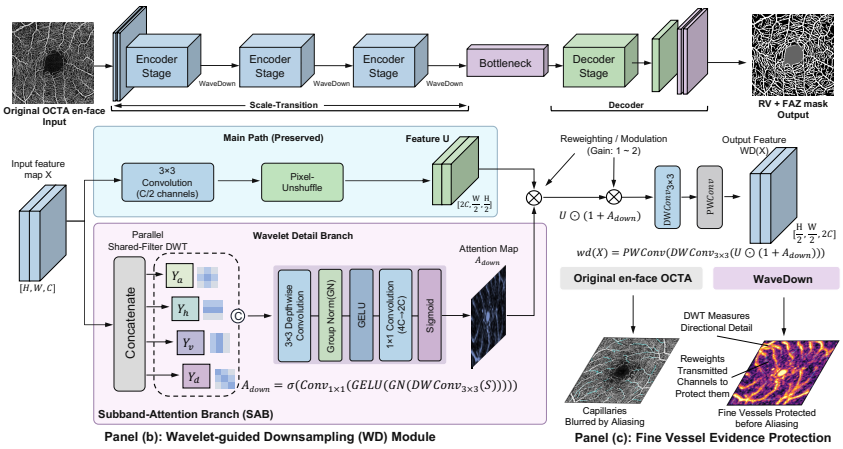

Figure 3: WaveDown. (a) Insertion at every encoder scale transition. (b) A parallel shared-filter DWT measures pre-downsampling directional detail and the SAB reweights transmitted channels via $U \odot (1 + A_{\text{down}})$ (Eqs. (2)–(3)); the main path is preserved. (c) Standard downsampling aliases capillaries; WaveDown protects fine-vessel channels before aliasing occurs.

and the output fuses the two paths:

$$WaveDown(X) = \text{PWConv}(\text{DWConv}_{3\times3}(U \odot (1 + A_{\text{down}}))) \,. \tag{3}$$

Role in segmentation. Downsampling is where capillaries die: strided convolutions and pixel unshuffle alike alias or average high-frequency content onto the coarse grid. WaveDown keeps the main path intact but uses the DWT of the *pre-downsampling* map to measure directional detail per position and channel, then weights exactly the downsampled channels carrying capillary evidence (gain $\in (1, 2)$), leaving detail-free channels at ≈ 1—protecting, before aliasing occurs, the information the decoder needs at every scale transition. Frequency evidence thus acts on transmitted features rather than attention output alone (contrast with anti-aliased downsampling in Section 5).

3.5 Spatially Enhanced Attention (SEA) in the Decoder

Decoder and refinement blocks use SEA—an MDTA variant with value-conditioned channel gating in the spirit of ConvolutionalGLU (Shi, 2024) (Eq. (6), Appendix E): the gate reweights each aggregated value channel by the evidence in V itself, sharpening boundary-transition channels when the decoder reintegrates wavelet-protected skips. Encoder level 1 uses SEA for the same reason.

3.6 Dual-Task Heads and Joint Objective

RV and FAZ masks are strictly disjoint on OCTA-500 3 mm (zero shared pixels across all 200 cases). The backbone is shared—including wavelet filters shaped by both tasks—while two parameter-independent 3×3 convolutions $H_{\text{RV}}, H_{\text{FAZ}}$ map the shared feature $F \in \mathbb{R}^{64\times304\times304}$ to task logits: independent heads prevent the FAZ (only $\sim$3–4% of pixels) from being dominated by RV gradients, while sharing regularizes both tasks (Vandenhende et al., 2022).

Each task is supervised by a weighted Dice–BCE objective (Milletari et al., 2016): Dice optimizes bulk overlap, BCE supplies dense per-pixel gradients favoring sharp boundaries (definitions in Appendix E), combined with equal task weights:

$$\mathcal{L}_{\text{total}} = \textstyle\sum_{t\in\{\text{RV,FAZ}\}} (0.6\,\mathcal{L}^t_{\text{Dice}} + 0.4\,\mathcal{L}^t_{\text{BCE}}) \,. \tag{4}$$

Appendix B verifies this joint configuration matches two dedicated single-task models on RV and FAZ Dice—both biomarkers at half the cost, without interference, while capillary recall favors the joint model.

4 EXPERIMENTS

4.1 DATASETS AND PROTOCOLS

OCTA-500 (main benchmark). The 3 mm subset (Li et al., 2024): 200 OCTA(FULL) projections with dense RV/FAZ labels under the *official* 140/10/50 split; every architecture is trained under one identical protocol (preprocessing in Appendix A).

OCTA-500 6 mm (field-of-view shift). 300 scans (400×400), official 180/20/100 split, same pipeline. Coarser pixel pitch ($\approx$15 vs. $\approx$9.9 μm/px) means a vessel of given physical diameter spans fewer pixels—probing frequency protection exactly where capillaries are most vulnerable; the fully convolutional network needs no architectural change.

FAROS (device and population shift). 40 scans (304×304) from a ZEISS PlexElite 9000 swept-source OCTA (Ma et al., 2026) with semi-automatic capillary-inclusive vessel labels and FAZ labels, official 24/8/8 split (Appendix A). Its labels are markedly denser than OCTA-500's manual ones (vessel area $\approx$25–37% vs. $\approx$6–8%)—a protocol gap kept in view throughout Section 4.7.

4.2 IMPLEMENTATION DETAILS

The main configuration has 12.4M parameters, base width $C = 32$, fixed Haar filters. Training uses AdamW (Loshchilov & Hutter, 2019), lr 10^{-4}, weight decay 10^{-2}, warmup plus cosine decay, batch size 2, mixed precision, 150 epochs ($\sim$29 s/epoch on one RTX 3090); block layout and full hyperparameters in Appendix B. Baselines (U-Net, Attention U-Net, Swin-Unet) are retrained from their published architectures under the same protocol (padding in Appendix A).

Model-selection protocol. WaveJoint is evaluated at the *last* checkpoint (epoch 150): the 10-case validation set makes epoch selection a noise source, and last-checkpoint evaluation cut three-seed FAZ std from 0.0107 to 0.0013 and mean-Dice std from 0.0061 to 0.0008 (an example in Appendix B). Baselines show no such instability (best/last within 0.2 pp) and use best checkpoints; the residual asymmetry is disclosed in Section 5.

Cross-dataset protocol. Zero-shot transfer evaluates the frozen 3 mm model directly on the target test set with training-time normalization and no target-domain tuning; in-domain runs use identical hyperparameters (target test sets and seeds in Section 4.7, Limitation (i)).

4.3 EVALUATION PROTOCOL: AGGREGATE AND STRUCTURAL

Predictions are binarized at 0.5. **Aggregate metrics** (per-case means): Dice, IoU, sensitivity, specificity. **Structural metrics**, computed per case on RV: (i) *caliber-stratified skeleton recall*—the ground-truth skeleton is bucketed by local caliber ($\leq$ 2 px capillary scale, 2–4, 4–8 px; $>$ 8 px excluded) and recall computed per bucket; (ii) *boundary distances* (HD95, ASSD) for RV and FAZ; (iii) a *fragmentation ratio* (predicted vs. ground-truth skeleton components); (iv) *Holm-corrected paired Wilcoxon tests* over the 50 cases, pairing models on the same case (values averaged over seeds) and correcting within metric families (Appendix C). This is deliberate: the literature routinely reports 0.1–0.2 pp Dice gaps with no statistical treatment—squarely inside seed noise (our main-model std $\approx$0.1 pp).

4.4 COMPARISON WITH PUBLISHED METHODS AND PROTOCOL-MATCHED BASELINES

Protocol anchoring. To verify protocol comparability with the literature, we retrained the *original* U-Net (Ronneberger et al., 2015) under ours: RV Dice 0.9020 ± 0.0013 versus the published 0.9068 on the same official split (as compiled by Chen et al., 2025); a control "modernized" U-Net with a contemporary recipe reaches 0.9139. Published numbers are therefore an anchor range, which is how we use them.

Table 1 reports the comparison. On aggregate Dice, WaveJoint (three seeds, last checkpoint) exceeds every protocol-matched retrained baseline—$+2.0$ pp RV / $+0.7$ pp FAZ over U-Net, $+3.7/+1.6$ over Attention U-Net, $+4.0/+1.8$ over Swin-Unet. Against the published joint systems (RV 0.9113–0.9165, FAZ 0.9824–0.9848), RV leads the field and surpasses the vessel-specialized OCT2Former (0.9193), and FAZ (0.9820) operates in the same band as the best systems (within 0.3 pp). Those sys-

Table 1: Joint RV+FAZ segmentation on the official OCTA-500 3 mm split. †: published official-split values (U-Net, IPN, Joint-Seg as compiled by Chen et al., 2025; IPN's original split is private; SAM-OCTA self-reported; OCT2Former vessel-only; PL-Joint-Seg self-reported). ‡: retrained under our protocol (3 seeds, mean±std, best checkpoint; Swin-Unet seed 42); ours: last checkpoint.

Method	Input	RV Dice	RV IoU	FAZ Dice	FAZ IoU
U-Net †	2D	0.9068	0.8301	0.9747	0.9585
IPN †	2D+3D	0.9062	0.8325	0.9505	0.9091
Joint-Seg †	2D	0.9113	0.8378	0.9843	0.9693
OCT2Former †	2D	0.9193	0.8513	–	–
SAM-OCTA †	2D+prompt	0.9157	0.8452	0.9824	0.9659
PL-Joint-Seg †	2D+3D	0.9165	0.8498	0.9848	0.9701
U-Net ‡	2D	0.9020±0.0013	0.8216±0.0022	0.9753±0.0001	0.9517±0.0002
Attn U-Net ‡	2D	0.8854±0.0053	0.7943±0.0085	0.9664±0.0031	0.9349±0.0058
Swin-Unet ‡	2D	0.8825	0.7897	0.9636	0.9297
WaveJoint (ours)	**2D**	**0.9220±0.0002**	**0.8553±0.0003**	**0.9820±0.0013**	**0.9647±0.0026**

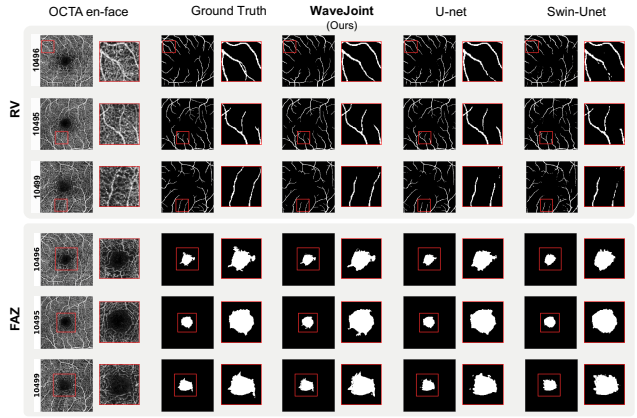

Figure 4: Qualitative results on OCTA-500 3 mm test cases (top three rows: RV; bottom three: FAZ): input with region of interest (red), ground truth, and predictions of WaveJoint, U-Net, and Swin-Unet, with zoomed insets. RV insets highlight capillary-endpoint continuity; FAZ insets boundary delineation.

tems rely on dual-decoder constraints, 3-D projection, or a prompted foundation model; WaveJoint attains this level from a single 2-D input with no pretraining or prompts. We therefore emphasize structural metrics (Section 4.5); Figure 4 previews these gains.

4.5 Structural evaluation: what Dice masks

Table 2 is the paper's core empirical result. The aggregate Dice gaps over the protocol-matched U-Net are $+2.0$ pp (RV) and $+0.7$ pp (FAZ); the structural protocol is sharper. **First**, the capillary-recall gap ($+\mathbf{3.25}$ pp) is the clinically relevant difference and the one Dice expresses worst: capillaries contribute only a few percent of vessel pixels, so their loss barely moves Dice while dominating the biomarker's clinical use. **Second**, fragmentation exposes the mechanism: U-Net's tree breaks into $2.88\times$ as many components as the ground truth (Swin/Attention U-Net: 4.2–$4.7\times$); WaveJoint's $\mathbf{1.50}$ is 48% lower, and density maps from a broken tree underestimate every connectivity-dependent index. RV boundary distances and FAZ HD95 are statistically tied with U-Net under Holm correction, whereas WaveJoint's FAZ ASSD remains significantly lower ($p_{\mathrm{Holm}} = 3.9\times10^{-3}$); against the other baselines every boundary distance improves ($p_{\mathrm{Holm}} < 10^{-4}$). A Dice-comparable U-Net thus matches boundary smoothness while delivering markedly worse capillary recall and connectivity.

FAZ morphology. The structural metrics above are RV-defined; on the same 50 cases we compared FAZ area, perimeter, circularity ($4\pi A/P^2$), and solidity (A/convex-hull area) with GT. All three runs show no Holm-significant bias, with area within 1% and perimeter showing the largest cross-case spread (Table 7, Appendix F).

What the Dice row does and does not say. Against the retrained U-Net, the RV Dice difference ($+2.0$ pp) is itself significant ($p_{\mathrm{Holm}} = 7.5\times10^{-10}$); at this magnitude Dice is not blind. The claim defended is narrower: (i) vs. the *published* systems of Table 1, routine 0.1–0.6 pp Dice gaps are protocol noise; (ii) even a significant Dice gap understates structure—$+3.25$ pp capillary recall,

Table 2: Structural comparison on the 50 test cases (RV unless noted; paired Wilcoxon, Holm-corrected p as in Section 4.3). Baselines: retrained, best checkpoint, 3 seeds (Swin-Unet: seed 42); WaveJoint: last checkpoint, 3 seeds. Fragmentation: predicted/ground-truth skeleton components (1.0 = no spurious breaks).

Metric	WaveJoint	U-Net	Attn U-Net	Swin-Unet	Δ vs. U-Net	p_{Holm}
Cap. recall (≤ 2 px)	$\mathbf{0.8354 \pm 0.0035}$	0.8029 ± 0.0058	0.7469 ± 0.0336	0.7525	$\mathbf{+3.25}$ pp	$\mathbf{1.1 \times 10^{-7}}$
RV HD95 (px) ↓	$\mathbf{1.74 \pm 0.10}$	2.09 ± 0.09	3.45 ± 0.46	2.76	-0.36 px	0.70 (n.s.)
RV ASSD (px) ↓	$\mathbf{0.355 \pm 0.006}$	0.368 ± 0.007	0.486 ± 0.029	0.462	-0.013 px	0.90 (n.s.)
FAZ HD95 (px) ↓	$\mathbf{4.15 \pm 0.11}$	4.89 ± 0.09	6.89 ± 0.21	5.88	-0.74 px	0.76 (n.s.)
FAZ ASSD (px) ↓	$\mathbf{0.470 \pm 0.014}$	0.691 ± 0.019	0.992 ± 0.075	0.931	-0.221 px	$\mathbf{3.9 \times 10^{-3}}$
Fragmentation ↓	$\mathbf{1.50 \pm 0.04}$	2.88 ± 0.10	4.19 ± 0.46	4.67	$\mathbf{-48\%}$	$\mathbf{1.6 \times 10^{-8}}$
RV Dice (reference)	0.9220 ± 0.0002	0.9020 ± 0.0013	0.8854 ± 0.0053	0.8825	$+2.0$ pp	$\mathbf{7.5 \times 10^{-10}}$

Table 3: 2×2 factorial ablation for three seeds: capillary-scale (≤ 2 px) recall and mean Dice (RV/FAZ average; learnable filters hold the filter variable constant; PU = pixel unshuffle). The WaveDown cap-recall main effect is $+1.26$/$+3.30$/$+4.20$ pp (mean $+2.9$, significant in every seed), while Dice cells span only 0.18 pp. Effect formulas are in Appendix C.

	Cap. recall (≤ 2 px)				Mean Dice			
Cell	s42	s43	s44	mean	s42	s43	s44	mean
a: WACGA+WaveDown	0.8311	0.8529	0.8456	0.8432	0.9463	0.9485	0.9483	0.9477
b: WACGA+PU	0.8317	0.8273	0.8280	0.8290	0.9457	0.9471	0.9463	0.9464
c: MDTA+WaveDown	0.8494	0.8330	0.8642	0.8489	0.9452	0.9468	0.9473	0.9464
d: MDTA+PU	0.8361	0.8256	0.8397	0.8338	0.9465	0.9447	0.9465	0.9459

48% lower fragmentation. The issue is Dice *sufficiency*, not *significance*; we suggest reporting skeleton-stratified recall and fragmentation alongside Dice and release our scripts.

4.6 Factorial ablation: localizing the load-bearing mechanism

We now ask which component carries the gain. The two wavelet-carrying modules (attention: WACGA vs. standard MDTA; downsampling: WaveDown vs. pixel unshuffle) form a complete 2×2 design over three seeds (42/43/44; learnable filters hold the filter variable constant; the MDTA+pixel-unshuffle cell contains no wavelet). Table 3 reports capillary recall—the metric the modules are meant to protect.

Three findings. **(i) WaveDown is the load-bearing module.** Its capillary-recall main effect is positive in every seed—$+1.26$, $+3.30$, and $+4.20$ pp (mean $+2.9$ pp)—and significant each time ($p = 0.020$, 1.8×10^{-10}, 1.9×10^{-13}); it is the only factor whose effect is both significant and consistent in direction. **(ii) The WACGA effect is seed-context-dependent.** Its main effect is -2.27, $+2.15$, and -3.03 pp across seeds (mean -1.1 pp): significant per seed but conditional on the training trajectory, rather than a stable standalone gain—motivating the deployed-setting replication below. **(iii) Interactions and Dice are flat.** Cap-recall interaction estimates sign-flip (-1.38/$+1.82$/-0.69 pp) and the pooled estimate over 150 seed-cases is null (-0.09 pp, $p = 0.41$): no consistent redundancy. The four configurations' 3-seed mean Dice spans only 0.18 pp (0.9459–0.9477; contrasts on those means stay at or below 0.18 pp, while per-seed Dice contrasts reach 0.4 pp, against up to 4.2 pp for capillary recall), whereas the WaveDown mean cap-recall effect is $+2.9$ pp: a Dice-only ablation would conclude, incorrectly, that no module is important—metric insensitivity, not an interaction effect.

Replication under the deployed fixed-Haar setting. Two fixed-bank rerun cells test whether the seed-dependent WACGA effect transfers to the deployed setting—the released WACGA+WaveDown model (capillary recall 0.8513, RV Dice 0.9156, FAZ Dice 0.9816, mean 0.9486) and its MDTA counterpart (0.8508, 0.9133, 0.9817, 0.9475)—show the WACGA variant matching or exceeding the MDTA variant on capillary recall, RV Dice, and mean Dice, with FAZ Dice tied within 0.01 pp (0.9816 vs. 0.9817): the per-seed learnable-setting effects do not transfer. The leaner MDTA+WaveDown variant is equally valid; the gain localizes to WaveDown (finding (i)), and we retain WACGA because the fixed-filter cells show no cost and its gate matches the design intent. The fixed-bank WaveDown cell (0.8508) likewise exceeds the available pixel-unshuffle cell (0.8361, learnable suite) by 1.5 pp in seed 42. Against fixed filters, learnable filters in the full architecture cost 2.02 pp capillary recall in the matched seed-42 comparison (0.8513 vs. 0.8311, $p_{\mathrm{Holm}} = 4.1 \times 10^{-5}$); the learnable full-model cells average 0.8432 across seeds, below

the fixed-bank 0.8513, and a multi-seed fixed-filter 2×2 remains future work (Limitation (i)). The joint-vs.-single-task comparison is in Appendix B.

4.7 Generalization: two shift axes

Field-of-view shift: OCTA-500 6 mm. Zero-shot (the frozen 3 mm model on the official 100-case test set), the ordering is preserved and amplified: WaveJoint exceeds U-Net by $+2.2$ pp RV Dice and $+\mathbf{12.5}$ pp capillary recall; U-Net's fragmentation and RV HD95 nearly double (8.35 vs. 4.75 px zero-shot). FAZ runs the other way zero-shot (0.6434 vs. 0.6690, Table 5); plausibly the frozen head misreads a contour whose support shifts coarser—we leave this open. In-domain, three seeds (pooled Dice) favor WaveJoint on every Dice mean—RV 0.8841 vs. 0.8719 ($+1.2$ pp), FAZ 0.8854 vs. 0.8801 ($+0.5$ pp)—alongside $+\mathbf{7.8}$ pp capillary recall and 57% lower fragmentation; the RV edge holds in all seeds; the FAZ edge reverses in seed 42 (0.8648 vs. 0.8796; Section 4.2). 6 mm capillary-recall/fragmentation rows are single-seed (Limitation (i)); this probes FOV/resolution, not external validation (FAROS serves that role).

Device and population shift: FAROS. Zero-shot on all 40 FAROS cases (Table 5) yields low RV Dice for *both* models—the annotation gap (vessel area ≈25–37% vs. ≈6–8%) penalizes capillary beds OCTA-500 labels never covered. Capillary recall floors at 0.015 for both, so we claim no zero-shot FAROS RV structural advantage; the surviving edges are FAZ Dice ($+1.4$ pp) and fragmentation (45% lower). In-domain training on the official 24/8/8 split, which removes the protocol gap, shows the clear WaveJoint advantage: $+4.8$ pp RV Dice, $+1.4$ pp FAZ Dice, $+\mathbf{8.1}$ pp capillary recall, and $2.5\times$ lower fragmentation. The 8-case test limits power; we make no per-pathology claims.

5 Discussion and limitations

Takeaways for the field. First, comparable Dice does not mean comparable structure: models within 0.5 pp Dice can differ by several points of capillary recall and a factor of two in fragmentation; we release the evaluation scripts and suggest structural reporting become standard. Second, mechanisms should be localized rather than presumed: the benefit concentrates in a single path. Third, generalization claims need scope: the structural advantage holds in-domain on all three datasets and zero-shot at 6 mm ($+12.5$ pp); on FAROS zero-shot recall floors for both models.

Scope and limitations. The factorial's pixel-unshuffle arm applies no low-pass, so the WaveDown effect is measured against lossless downsampling, not BlurPool-style anti-aliasing (Zhang, 2019); a direct BlurPool arm remains open. The structural edges are *expected*, not measured (no downstream biomarker-error experiments), to improve vessel-density, connectivity, and FAZ quantification. Three further limitations bound the evidence. **(i) Seed and aggregation coverage.** The primary contrasts—3 mm main results, factorial capillary recall and mean Dice, 6 mm Dice, and FAROS in-domain—use three seeds; the fixed-Haar replication cells, fixed-vs.-learnable filter contrasts, 6 mm capillary-recall/fragmentation rows, and Swin-Unet per-case values remain single-seed, with two further dataset-level seeds (Table 8) confirming stability. **(ii) Metric scope.** Connectivity and caliber metrics are defined on RV; FAZ is covered by boundary distances and the morphology audit (Table 7, one WaveJoint seed), and caliber buckets are pixel-defined and subset-specific, with μm conversion awaiting device calibration. **(iii) Statistical power.** The 10-case validation and 8-case FAROS test splits limit checkpoint selection (last for us, best for baselines, agreeing within 0.2 pp) and pathology stratification, respectively.

6 Conclusion

WaveJoint delivers competitive joint RV–FAZ segmentation from a single OCTA projection by placing an explicit frequency decomposition at the scale transitions where capillaries otherwise die. Its broader message concerns evaluation and design: structural metrics reveal what Dice masks, and factorial ablations with interactions identify the module carrying the benefit—here, the wavelet-guided downsampling path. Validated across field-of-view and device-population shifts, we expect the protocol and localization methodology to outlive the architecture.

REPRODUCIBILITY STATEMENT

All three datasets are public. We will release code, trained checkpoints, and the full structural evaluation pipeline (skeletonization, caliber stratification, fragmentation, and the Holm-corrected paired Wilcoxon tests) upon publication. Training uses the official splits with the exact hyperparameters of Appendix B; every number in the paper states its seeds and checkpoint policy explicitly.

AI USE STATEMENT

In this work, we used generative AI tools for language editing of the manuscript and for assistance with LaTeX formatting; we did not use generative AI tools for generating research ideas, experimental code, data analysis, or the scientific conclusions, which were conducted entirely by the authors. We have reviewed all AI-assisted text for correctness and for consistency with our experimental records, and all reported numbers were produced by our own experiments. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available, de-identified datasets (OCTA-500 and FAROS); no new human-subject data was collected. The method is a research prototype for quantitative biomarker extraction and is not intended for clinical diagnosis without prospective validation.

REFERENCES

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280.

Xinrun Chen, Chengliang Wang, Haojian Ning, Shiying Li, and Mei Shen. SAM-OCTA: Prompting segment-anything for OCTA image segmentation. *Biomedical Signal Processing and Control*, 106:107698, 2025. doi: 10.1016/j.bspc.2025.107698.

Helya Goharbavang, Artem T. Ashitkov, Athira Pillai, Joshua D. Wythe, Guoning Chen, and David Mayerich. Segmentation and modeling of large-scale microvascular networks: A survey. *Frontiers in Bioinformatics*, 5:1645520, 2025. doi: 10.3389/fbinf.2025.1645520.

Kai Hu, Shuai Jiang, Yuan Zhang, Xuanya Li, and Xieping Gao. Joint-Seg: Treat foveal avascular zone and retinal vessel segmentation in OCTA images as a joint task. *IEEE Transactions on Instrumentation and Measurement*, 71:1–13, 2022. doi: 10.1109/TIM.2022.3193188.

Shuai Jiang, Kai Hu, Yuan Zhang, and Xieping Gao. Joint segmentation of FAZ and RVs in OCTA images with auxiliary 3-D image projection learning. *IEEE Transactions on Instrumentation and Measurement*, 74:2512512, 2025. doi: 10.1109/TIM.2025.3545862.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Mingchao Li, Yerui Chen, Zexuan Ji, Keren Xie, Songtao Yuan, Qiang Chen, and Shuo Li. Image projection network: 3D to 2D image segmentation in OCTA images. *IEEE Transactions on Medical Imaging*, 39(11):3343–3354, 2020. doi: 10.1109/TMI.2020.2992244.

Mingchao Li, Kun Huang, Qiuzhuo Xu, Jiadong Yang, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. OCTA-500: A retinal dataset for optical coherence tomography angiography study. *Medical Image Analysis*, 93:103092, 2024. doi: 10.1016/j.media.2024.103092.

Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. WaveCNet: Wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification. *IEEE Transactions on Image Processing*, 30: 7074–7089, 2021. doi: 10.1109/TIP.2021.3101395.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Ke Ma, Jinze Zhang, Xiaodong Hu, Yupei Chen, Gengyuan Wang, Yuancong Huang, Jin Yuan, and Peng Xiao. Deep learning-based segmentation and quantitative analysis of retinal microstructures in optical coherence tomography angiography images using RS_Unet3+. *Intelligent Medicine*, 6 (1):12–23, 2026. doi: 10.1016/j.imed.2025.05.008.

Martin J. Menten, Johannes C. Paetzold, Veronika A. Zimmer, Suprosanna Shit, Ivan Ezhov, Robbie Holland, Monika Probst, Julia A. Schnabel, and Daniel Rueckert. A skeletonization algorithm for gradient-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21337–21346, 2023.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015.

Wentao Ruan, Xinhui Li, Zhan Cheng, Cunhang Fan, Libao Tian, and Zhao Lv. LWTformer: A detail-aware, learnable wavelet-transformer for ancient Chinese character image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dai Shi. TransNeXt: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17773–17783, 2024.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. 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 (CVPR)*, pp. 16560–16569, 2021.

Richard F. Spaide, James G. Fujimoto, Nadia K. Waheed, Srinivas R. Sadda, and David Sarraf. Optical coherence tomography angiography. *Progress in Retinal and Eye Research*, 64:1–55, 2018.

Xiao Tan, Xinjian Chen, Qingquan Meng, Fei Shi, Dehui Xiang, Zhongyue Chen, Lingjiao Pan, and Weifang Zhu. OCT2Former: A retinal OCT-angiography vessel segmentation transformer. *Computer Methods and Programs in Biomedicine*, 233:107454, 2023. doi: 10.1016/j.cmpb.2023.107454.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3614–3633, 2022. doi: 10.1109/TPAMI.2021.3054719.

Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-ViT: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 328–345, 2022.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

A Dataset and preprocessing details

OCTA-500 3 mm. Images are resized from 303×303 to 304×304 (bilinear for images, nearest-neighbor for masks), normalized to $[0, 1]$, and masks binarized at 127. Training augmentation (random horizontal/vertical flips and 90° rotations) is sampled per case and applied consistently to the image and both masks.

OCTA-500 6 mm. Same pipeline as 3 mm, at 400×400 native resolution with the official 180/20/100 split. The network is fully convolutional and 400 is divisible by the $8\times$ total downsampling factor, so inputs require no modification.

FAROS. 304×304 native, official 24/8/8 split, mask semantics 0 = background / 100 = FAZ / 255 = RV. Vessel annotations are semi-automatic and capillary-inclusive (vessel area ≈25–37% vs. ≈6–8% for OCTA-500); this protocol difference is disclosed wherever cross-dataset numbers are interpreted (Sections 4.1, 4.7, 5).

Baseline input handling. Since 304 is not divisible by 32, U-Net-family inputs are padded to 320 during training and center-cropped back to 304 at prediction, preserving per-case pairing for all paired statistics.

B Implementation and training details

Main configuration: 12.4M parameters, base width $C = 32$; encoder blocks 4/6/6, bottleneck 8, decoder 6/6/4, refinement 4; attention heads 1/2/4/8/4/2/1/4; FFN expansion 2.66; bias-free LayerNorm; fixed Haar wavelet filters ($\mathbf{l} = \frac{1}{\sqrt{2}}(1, 1)$, $\mathbf{h} = \frac{1}{\sqrt{2}}(1, -1)$), frozen. Training: AdamW, learning rate 10^{-4}, weight decay 10^{-2}, 3-epoch linear warmup with cosine decay, batch size 2, mixed precision, 150 epochs, single NVIDIA RTX 3090, ∼29 s/epoch; inference ∼50 ms per 304×304 input. Seeds: main results, 6 mm in-domain Dice, and FAROS in-domain use $\{42, 43, 44\}$; 6 mm capillary recall/fragmentation, the factorial design, and Swin-Unet use seed 42. Loss weights and the Dice/BCE mixture are given in Equation (4) and Appendix E. All experiments use 150 epochs; the 300-epoch control changes mean Dice by -3×10^{-5}.

Checkpoint-protocol example. The instability motivating last-checkpoint evaluation is documented directly: one fixed-Haar seed's FAZ best checkpoint (epoch 37) departed from its validation trajectory, with a test-set FAZ Dice of 0.9626 at that checkpoint, while its last checkpoint recovered to 0.9831.

Cross-dataset protocol. Zero-shot transfer evaluates the frozen 3 mm model directly on the target test set (FAROS: all 40 cases; 6 mm: the official 100-case test set) with training-time normalization and no target-domain tuning. In-domain training uses identical hyperparameters; FAROS in-domain runs three seeds (test set: 8 cases), and 6 mm Dice runs three seeds while its capillary-recall/fragmentation rows use seed 42.

Ablation-suite aggregation and joint-vs.-single-task. The factorial cells (Section 4.6; capillary recall and mean Dice over three seeds) and the joint-vs.-single-task comparison (single seed 42) report dataset-level pooled scores (one score over the test set) rather than the per-case means with standard deviations used in Tables 1–2; their absolute values are therefore not comparable across the two aggregations, and the contrasts of interest are within-suite only. Additional s43/s44 runs under this aggregation, including best/last checkpoints, are reported in Table 8 (Appendix F). Under this within-suite comparison, RV and FAZ Dice match the dedicated single-task models within noise—RV Dice 0.9156 vs. 0.9138 (RV-only), FAZ Dice 0.9816 vs. 0.9786 (FAZ-only)—while capillary recall in fact favors the joint model by 2.3 pp (0.8513 vs. 0.8284, RV-only): joint training halves the cost of two dedicated models with no Dice penalty and a capillary-recall edge in its favor.

C STRUCTURAL METRIC DEFINITIONS AND STATISTICAL PROTOCOL

Caliber-stratified skeleton recall. The ground-truth RV mask is thinned to its one-pixel-wide medial skeleton by morphological thinning, as reviewed for microvascular networks by Goharbavang et al. (2025); a modern topology-preserving, gradient-compatible skeletonization is given by Menten et al. (2023). Each skeleton pixel is assigned a local caliber via the distance transform of the ground-truth mask and bucketed: $\leq$ 2 px (capillary scale), 2–4 px, 4–8 px, $>$ 8 px. Recall of the binarized prediction is computed per bucket over all skeleton pixels in the bucket. The $\leq$ 2 px bucket contains 56,755 skeleton pixels across the 50 test cases—the population where fine-vessel failure actually manifests; the $>$ 8 px bucket contains 151 pixels across 18 cases and is excluded as statistically meaningless. Buckets are pixel-defined and therefore subset-specific: $\leq$ 2 px corresponds to $\leq\approx$20 μm at 3 mm but $\leq\approx$30 μm at 6 mm, so recall values are comparable only within a subset.

Fragmentation ratio. The number of 8-connected components of the predicted skeleton divided by the number for the ground-truth skeleton, per case; 1.0 indicates no spurious breaks.

Boundary distances. HD95: the 95th percentile of directed surface distances between prediction and ground-truth boundaries; ASSD: the average symmetric surface distance. Both are computed for RV and FAZ.

FAZ morphology. For each predicted and ground-truth FAZ mask we report area A (mm^2: pixel count times the squared pixel pitch, 0.009868 mm/px at 3 mm), perimeter P (mm, boundary-corrected), circularity $4\pi A/P^2$ (1 for a perfect circle), and solidity A/A_{hull} with A_{hull} the convex-hull area. Area and perimeter errors are relative errors (%); circularity and solidity errors are signed differences; bias is tested per case with paired Wilcoxon tests, Holm-corrected over the four metrics (Table 7).

Statistics. All structural comparisons are per-case paired Wilcoxon signed-rank tests; each model's per-case value is the mean over its training seeds, the pair couples the two models on the same test case, and Holm correction is applied within each metric family (the set of baseline comparisons for one metric). For Table 2 this yields seven families of three tests each (21 tests); the ablation contrasts of Section 4.6 pair single-seed (42) runs case by case. Only corrected survivors are marked significant. For the three-seed factorial, effects are high-vs-low factor contrasts: $E_{WaveDown} = (a+c)-(b+d)$, $E_{WACGA} = (a+b)-(c+d)$, interaction $= (a-b)-(c-d)$; significance stars denote $^{***}p < 10^{-4}$, $^{**}p < 10^{-3}$, $^{*}p < 0.05$.

D DWT DETAILS AND SUBBAND CORRESPONDENCE

One level of 2-D DWT with the kernels Φ_s of Section 3.2 is a strided convolution of the boundary-padded map $\tilde{X}$:

$$Y_s^{(c)}[m,n] = \sum_{i,j} \Phi_s[i,j]\, \tilde{X}^{(c)}[2m+i,\, 2n+j], \qquad s \in \{a,h,v,d\}, \tag{5}$$

implemented as a single grouped convolution (stride 2), yielding four subbands $Y_s \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$. Subbands are named by their (row, column) filter pair in Φ_s: Y_h (kernel $\mathbf{hl}^\top$, often written LH) responds to horizontal edges, Y_v ($\mathbf{lh}^\top$, HL) to vertical edges.

Table 4: Frequency–structure–operator correspondence (design level).

Subband	Dominant OCTA content	Operator	Intended effect
Y_a (LL)	lumens, background, FAZ interior	1×1 proj. + 3×3 DW-Conv	preserve large-scale structure; anchor gate where no detail
Y_h (LH)	horizontal vessel edges, FAZ boundary	grouped 1×3 conv. + P_h	direction-matched enhancement of horizontal edges
Y_v (HL)	vertical vessel edges	grouped 3×1 conv. + P_v	analogous enhancement of vertical edges
Y_d (HH)	bifurcations, capillary tips, decorrelation noise	$\tanh$ + 1×1 proj. + P_d	bound heavy tails; isolate point-like detail

Table 4 summarizes how each subband's expected anatomical content is matched to its operator, in both WACGA and WaveDown.

E MODULE OPERATORS AND LOSS DETAILS

Spatially enhanced attention (SEA). SEA is an MDTA variant with value-conditioned channel gating (ConvolutionalGLU (Shi, 2024)):

$$O = \mathrm{softmax}\Big((\hat{Q}\hat{K}^\top)\tau\Big)\,V, \qquad SEA(X) = \mathrm{Proj}\Big(O \odot \mathrm{MLP}\big(V^\top\big)^\top\Big). \tag{6}$$

Wavelet-branch operators. With Z denoting the concatenated branch outputs (each compressing $C \to C/4$):

$$\begin{aligned} Z_a &= \mathrm{DWConv}_{3\times3}(W_a Y_a), \quad & Z_h &= P_h\big(\mathrm{Conv}^g_{1\times3} Y_h\big), \\ Z_v &= P_v\big(\mathrm{Conv}^g_{3\times1} Y_v\big), & Z_d &= P_d(W_d \tanh(Y_d)), \end{aligned} \tag{7}$$

where Conv^g is a 4:1 grouped anisotropic convolution (channel compression and directional filtering in one step) and P_h, P_v, P_d are 3×3 depthwise kernels initialized to Prewitt-type directional gradients and a Laplacian, all trainable:

$$P_h = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \quad P_v = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}, \quad P_d = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{8}$$

Loss terms. With $p = \sigma(\hat{Y})$ and binary target g (Milletari et al., 2016):

$$\begin{aligned} \mathcal{L}_{\mathrm{Dice}} &= 1 - \frac{1}{B}\sum_b \frac{2\sum_i p_{b,i} g_{b,i} + \epsilon}{\sum_i p_{b,i} + \sum_i g_{b,i} + \epsilon}, \\ \mathcal{L}_{\mathrm{BCE}} &= -\frac{1}{B}\sum_{b,i}\Big[g_{b,i}\log p_{b,i} + (1 - g_{b,i})\log(1 - p_{b,i})\Big], \end{aligned} \tag{9}$$

combined per Equation (4) with equal task weights $w_{\mathrm{RV}} = w_{\mathrm{FAZ}} = 1$.

F ADDITIONAL RESULTS

Table 5: Cross-dataset generalization (dual task), referenced from Section 4.7. Zero-shot: frozen 3 mm model on the target test set (6 mm: official 100 cases; FAROS: all 40). In-domain: target-split training (6 mm Dice: 3 seeds, mean±std, per-seed values in Table 6; 6 mm capillary recall/fragmentation: seed 42 only, marked *, Limitation (i); FAROS: 3 seeds, mean±std); ‡: retrained U-Net, same protocol. Caliber buckets are subset-specific (Section 4.3); FAROS fragmentation under dense labels is not absolutely comparable with Table 2; 6 mm RV HD95 is in the Section 4.7 text.

Dataset	Method	Setting	RV Dice	FAZ Dice	Cap. recall	Frag.
6 mm	WaveJoint	zero-shot	0.8570	0.6434	0.7188	4.71
6 mm	U-Net ‡	zero-shot	0.8351	0.6690	0.5938	10.03
6 mm	WaveJoint	in-domain	0.8841 ± 0.0017	0.8854 ± 0.0171	0.8524^*	4.58^*
6 mm	U-Net ‡	in-domain	0.8719 ± 0.0015	0.8801 ± 0.0015	0.7741^*	10.62^*
FAROS	WaveJoint	zero-shot	0.3710	**0.9757**	0.0150	**5.42**
FAROS	U-Net ‡	zero-shot	0.3790	0.9618	0.0150	9.90
FAROS	WaveJoint	in-domain	$\mathbf{0.8390} \pm 0.0010$	$\mathbf{0.9819} \pm 0.0005$	$\mathbf{0.6975} \pm 0.0136$	$\mathbf{32.90} \pm 6.70$
FAROS	U-Net ‡	in-domain	0.7912 ± 0.0041	0.9679 ± 0.0052	0.6162 ± 0.0315	83.70 ± 3.40

Additional qualitative examples on OCTA-500 3 mm, 6 mm, and FAROS, and—pending case-level metadata—a per-pathology stratification of the FAROS results (healthy / DR / AMD / RVO) under the same three-seed protocol as Section 4.7, will accompany the released code; the present manuscript deliberately makes no per-pathology claims.

Table 6: 6 mm in-domain per-seed dataset-level pooled Dice on the official 100-case test set. RV Dice favors WaveJoint in every seed; the FAZ mean advantage reverses in seed 42. Across seeds, WaveJoint sensitivity ranges 0.88–0.90 (RV) and 0.886–0.899 (FAZ), with specificity above 0.985 for both tasks.

Model	Seed	RV Dice	FAZ Dice	Mean Dice
WaveJoint	42	0.8818	0.8648	0.8733
WaveJoint	43	0.8844	0.8847	0.8845
WaveJoint	44	0.8860	0.9067	0.8964
U-Net ‡	42	0.8714	0.8796	0.8755
U-Net ‡	43	0.8739	0.8822	0.8780
U-Net ‡	44	0.8704	0.8786	0.8745

Table 7: FAZ morphology audit on the 3 mm test set (50 cases), predictions vs. ground truth (GT: area $0.354\pm0.122\,\mathrm{mm}^2$, perimeter 2.929 ± 0.748 mm, circularity 0.5365 ± 0.1293, solidity 0.8690 ± 0.0571; pixel pitch 0.009868 mm/px). Area/perimeter: relative error (%); circularity/solidity: signed mean difference. All three runs pass Holm correction over the four metrics (0/4, no systematic bias); area error is below 1% and perimeter has the largest cross-case spread. Definitions in Appendix C.

Model (seed)	Area rel-err (%)	Peri. rel-err (%)	ΔCircularity	ΔSolidity
WaveJoint (43)	$+0.73\pm2.86$	-1.28 ± 8.88	$+0.019\pm0.082$	$+0.001\pm0.054$
Swin-Unet (43)	$+0.19\pm4.61$	$+0.38\pm14.01$	$+0.008\pm0.120$	-0.001 ± 0.053
Swin-Unet (44)	$+0.36\pm4.71$	$+1.78\pm14.59$	-0.004 ± 0.120	-0.014 ± 0.088

Table 8: Additional seeds in the p01 ablation-suite aggregation (3 mm test, 50 cases; dataset-level pooled Dice/IoU as in the factorial suite—not directly comparable with the per-case means of Tables 1–2; these runs also supply the morphology audit of Table 7). The full-model s43 mean (0.9495 best / 0.9472 last) brackets the s42 fixed-Haar value 0.9486; Swin-Unet s43/s44 means (0.9244/0.9239) match its s42 level. The s43 best–last gap (0.23 pp mean) is within the checkpoint-protocol noise of Section 4.2.

Run	Checkpoint	RV Dice	RV IoU	FAZ Dice	FAZ IoU	Mean Dice
WaveJoint s43	best	0.9160	0.8450	0.9830	0.9665	0.9495
WaveJoint s43	last	0.9117	0.8377	0.9828	0.9661	0.9472
Swin-Unet s43	best	0.8848	0.7934	0.9640	0.9304	0.9244
Swin-Unet s44	best	0.8854	0.7943	0.9624	0.9275	0.9239