# DUOPRIOR-CFFLOW: DUAL-PRIOR COARSE-TO-FINE FLOW MATCHING FOR 3D OCT-TO-OCTA TRANSLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Optical coherence tomography angiography (OCTA) provides clinically valuable retinal vascular information but remains more difficult to acquire than structural optical coherence tomography (OCT), motivating reliable OCT-to-OCTA translation. Three-dimensional OCT-to-OCTA translation is challenging because it must simultaneously preserve continuous global retinal structure and recover sparse, fine vascular details. We propose DuoPrior-CFFlow, a dual-prior coarse-to-fine flow-matching framework for paired three-dimensional OCT-to-OCTA translation. Its CF-Flow path decomposes volumetric representations into low- and high-frequency components, modeling global structural transport with Coarse Flow and vascular-detail transport with Fine Flow. The multi-scale Z-aware (MSZ) prior provides global depth, local depth, and depth-gradient cues for structural guidance, while VesselAux promotes three-dimensional vascular consistency through auxiliary vessel supervision. DuoPrior-CFFlow outperforms the directly comparable baselines on most quantitative metrics and better preserves retinal vascular structures. Structural preservation and downstream utility evaluations show that our translated 3D OCTA volumes achieve the highest 3D vessel-mask consistency, the lowest vascular occupancy in the foveal avascular zone, and the highest disease classification accuracy. Code is available at `https://anonymous.4open.science/r/duoprior-cfflow-code-6D9F/`.

## 1 INTRODUCTION

Optical coherence tomography (OCT) provides noninvasive, high-resolution visualization of retinal morphology (Huang et al., 1991), whereas optical coherence tomography angiography (OCTA) reveals retinal microvasculature without contrast injection(Foulsham et al., 2022). As vascular alterations are central to diseases such as diabetic retinopathy (DR) and age-related macular degeneration (AMD), OCTA offers distinctive biomarkers for disease assessment, progression monitoring, and treatment planning(De Carlo et al., 2015). However, its acquisition requires repeated scans and is susceptible to motion artifacts and signal degradation, limiting its routine availability. Reliable OCT-to-OCTA translation could therefore provide clinically useful vascular information from routinely acquired structural OCT volumes when OCTA is unavailable (Tan et al., 2018).

Recent studies have increasingly addressed three-dimensional OCT-to-OCTA translation, including TransPro with projection-based constraints (Li et al., 2024b), MuTri with multi-view semantic alignment (Chen et al., 2025), XOCT with cross-view modeling (Khosravi et al., 2025), and FlowOCT with flow-based constraints (Xiong et al., 2026). However, vascular fidelity remains a central challenge. As shown in Figure 1, existing generators can reproduce portions of major vessels but still exhibit missing vessels (orange) and spurious vessels (pink) relative to the ground-truth OCTA. These errors disrupt vascular continuity and topology, even when the overall appearance looks plausible. The proposed DuoPrior-CFFlow retains a larger matched vessel extent (cyan) in this representative region. This observation highlights the need to preserve both the continuous retinal structure and sparse high-frequency vascular details during three-dimensional OCTA translation.

Prior structure-aware methods provide important but incomplete ingredients for this task. FlowOCT (Xiong et al., 2026) demonstrates the value of wavelet modeling and structure-focused objectives,



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.

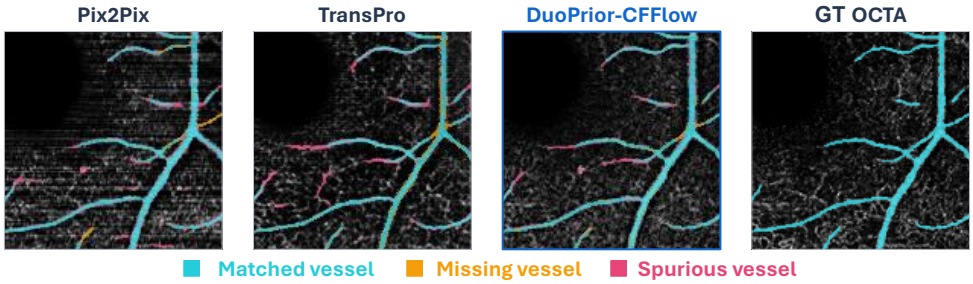



Figure 1: Visual comparison of vascular correspondence . Cyan, orange, and pink indicate matched, missing, and spurious vessel responses with respect to GT OCTA, respectively.

while XOCT (Khosravi et al., 2025) shows that depth and retinal-layer information can improve OCTA generation. TransPro (Li et al., 2024b) further introduces vessel-related supervision through projected vascular representations. However, prior wavelet formulations do not assign separate transport roles to low-frequency retinal structure and high-frequency vessel detail. Depth-aware constraints often rely on limited cues, while projection-level vessel supervision cannot preserve volumetric vascular topology. These limitations motivate a unified formulation that models coarse structure and fine vascular detail separately, while enriching depth-aware guidance and extending vessel supervision from projected representations to three-dimensional vascular predictions.

To this end, we propose **DuoPrior-CFFlow**, a Dual-Prior Coarse-to-Fine Flow Matching framework for three-dimensional OCT-to-OCTA translation. As illustrated in Figure 2, CF-Flow separates the 3D wavelet representation into low- and high-frequency components. Coarse Flow first models low-frequency structural transport, while Fine Flow uses the predicted coarse coefficients as context to recover high-frequency vascular details. Building on this generation path, the Multi-Scale Z-Aware (MSZ) Prior derives global depth, local depth, and depth-gradient cues from the source OCT to construct a volumetric structural weighting map. VesselAux further supervises the predicted high-frequency representation through a three-dimensional vessel prediction target, encouraging consistency of fine vascular detail. Our contributions are summarized as follows:

• We propose **DuoPrior-CFFlow**, a dual-prior coarse-to-fine flow-matching framework that reformulates paired 3D OCT-to-OCTA translation as a structured transport process in the wavelet domain. The coarse-to-fine flow design separates low-frequency retinal structure from sparse high-frequency vascular detail, reducing the tendency of a single generation path to favor one type of information at the expense of the other.

• To provide explicit volumetric structural guidance, we propose the Multi-Scale Z-Aware (MSZ) Prior, which extracts global-depth, local-depth, and depth-gradient cues from OCT to construct a volumetric structural weighting map. It emphasizes informative retinal regions and depth transitions that are easily overwhelmed by the extensive low-intensity background.

• To strengthen high-frequency vascular fidelity, we introduce VesselAux, which applies auxiliary three-dimensional vessel supervision to the predicted high-frequency representation. This guidance encourages vascular continuity and topology, helping reduce missing and spurious vascular responses in the translated OCTA.

• We conduct comprehensive quantitative and qualitative comparisons with representative OCT-to-OCTA translation methods . We further evaluate structural preservation and downstream utility through three-dimensional vessel structure consistency, FAZ-region vascular exclusion, and disease classification of the translated OCTA.

## 2 Related Work

### 2.1 OCT to OCTA Image Translation

Automatic OCT-to-OCTA translation aims to infer flow-sensitive angiographic signals from routinely acquired structural OCT. General paired image-to-image frameworks, such as Pix2Pix (Isola et al., 2017), provide common baselines for conditional translation but were not designed specifically for OCTA synthesis. Early task-specific studies therefore focused on two-dimensional data:

paired OCT/OCTA B-scans were translated using encoder-decoder models, Texture-UNet (Zhang et al., 2021), and AdjacentGAN (Li et al., 2020), while MultiGAN (Pan et al., 2022) and subsequent GAN-based approaches targeted en-face projection maps or coupled two-dimensional vessel segmentation with translation. Although these methods can recover local appearance, their slice-wise or projection-based formulations discard volumetric context and often yield discontinuous vascular responses across adjacent slices.

Recent work has consequently shifted toward three-dimensional OCT-to-OCTA translation. TransPro (Li et al., 2024b) couples a 3D adversarial generator with heuristic contextual guidance to maintain consistency between generated volumes and their projection maps, and applies vessel-promoted guidance on projected vascular regions. XOCT (Khosravi et al., 2025) uses layer-wise en-face projections for cross-dimensional supervision and combines multi-scale feature fusion with channel reweighting to recover layer-aware vessels. MuTri (Chen et al., 2025) learns the mapping in a discrete VQ-VAE space with tri-alignment across 3D OCT, 3D OCTA, and a 2D OCTA projection view. FlowOCT (Xiong et al., 2026) further formulates the task as a noise-free conditional flow in the 3D wavelet domain, initializing from observed OCT coefficients and learning a velocity field along the paired OCT-to-OCTA path. Despite these advances, current methods still struggle to preserve sparse capillaries and multi-scale vascular organization. Projection-level or limited depth-aware constraints cannot fully enforce volumetric vessel continuity and topology, while unified generation paths do not explicitly distinguish continuous retinal structure from sparse high-frequency vascular detail.

### 2.2 Generative Modeling for Medical Images

Generative modeling has become a central tool for medical image synthesis, reconstruction, and cross-modality translation. Conditional GANs, exemplified by Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), and spatially adaptive normalization models (Park et al., 2019), learn direct source-to-target mappings and provide fast inference, but adversarial optimization can be unstable and may fail to preserve fine anatomy. Denoising diffusion probabilistic models (Ho et al., 2020) instead learn iterative denoising processes and have achieved strong fidelity in medical image synthesis. Conditional image-to-image diffusion models (Saharia et al., 2022), latent diffusion models (Rombach et al., 2022), and volumetric diffusion frameworks (Wang et al., 2025) further improve controllability or memory efficiency for high-dimensional medical data. Wavelet-domain diffusion models (Jiang et al., 2023) additionally exploit lossless multi-resolution representations to make three-dimensional generation more tractable. Nevertheless, repeated sampling steps remain costly, and image-level fidelity alone does not ensure preservation of clinically relevant structures.

Flow matching (Lipman et al., 2023) offers a simulation-free alternative that learns a time-dependent velocity field along a continuous transport path. Flow Matching and its optimal-transport formulation provide general continuous-normalizing-flow objectives; Rectified Flow encourages straighter trajectories, while Conditional Flow Matching extends the formulation to controlled generation. Recent latent-space flow matching models (Geffner et al., 2026; Dao et al., 2023) have shown that performing flow transport in a learned latent space can improve the efficiency and scalability of high-resolution image synthesis. In medical imaging, flow matching has been explored for 2D/3D synthesis, reconstruction, and cross-modality translation (Yazdani et al., 2025; Moschetto et al., 2026), especially when the observed source can serve as an informative starting point rather than random noise. However, existing approaches typically transport the full image or latent representation through a unified trajectory. This can entangle smooth anatomical structure with sparse high-frequency details, limiting structural fidelity in challenging volumetric translation tasks. These observations motivate a structure-aware flow formulation that explicitly separates global anatomical transport from fine vascular-detail reconstruction for paired 3D OCT-to-OCTA translation.

## 3 Method

### 3.1 Problem Setup

Let $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{Y} \in \mathbb{R}^{H\times W\times D}$ denote a paired OCT volume and its corresponding OCTA volume, respectively. Our objective is to learn a conditional transformation $\mathcal{G} : \mathbf{X} \mapsto \widehat{\mathbf{Y}}$ that generates an OCTA volume while preserving the anatomical arrangement of the source OCT. Unlike

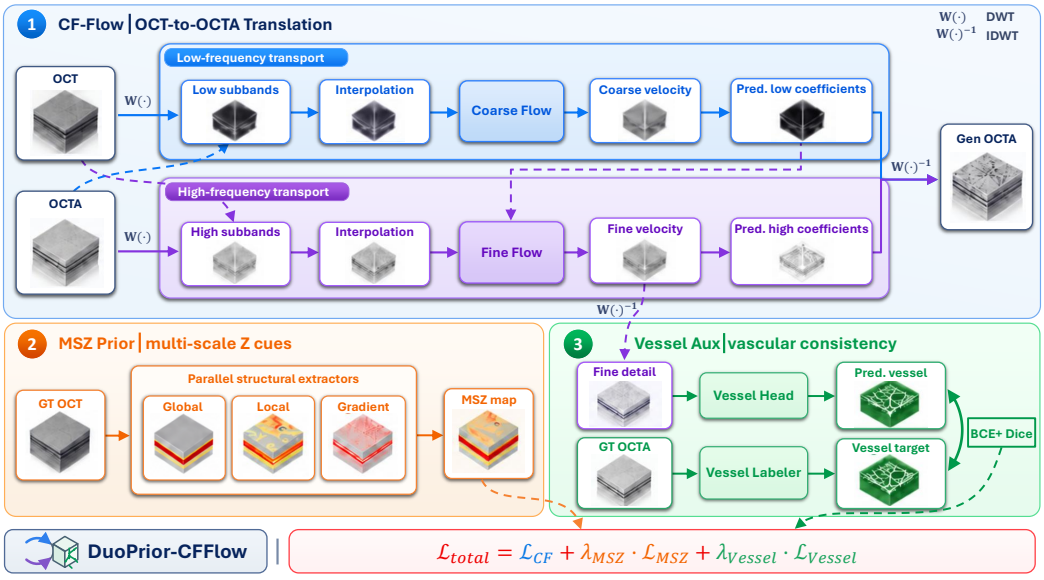



Figure 2: Overview of DuoPrior-CFFlow with three modules: 1) CF-Flow for low-frequency structural transport and high-frequency vascular-detail refinement in the 3D wavelet domain; 2) MSZ Prior for depth-aware structural guidance from OCT; and 3) VesselAux for auxiliary supervision of vascular detail and topology.

unconditional generation from random noise, this task starts from the observed source volume and learns a deterministic transport from OCT to OCTA.

### 3.2 Coarse-to-Fine Flow Matching

We define the conditional transport process directly in the three-dimensional wavelet domain. Let

$$\mathcal{W}(\mathbf{X}) = (\mathbf{x}_{\mathrm{L}}, \mathbf{x}_{\mathrm{H}}), \qquad \mathcal{W}(\mathbf{Y}) = (\mathbf{y}_{\mathrm{L}}, \mathbf{y}_{\mathrm{H}}) \tag{1}$$

denote the wavelet representations of paired OCT and OCTA volumes, where $\mathbf{x}_{\mathrm{L}}$ and $\mathbf{y}_{\mathrm{L}}$ are the low-frequency LLL coefficients, while $\mathbf{x}_{\mathrm{H}}$ and $\mathbf{y}_{\mathrm{H}}$ concatenate the remaining seven high-frequency subbands. Coarse-to-Fine Flow learns a low-frequency structural path and a high-frequency vascular-detail path.

**Coarse Flow.** For $t \in [0, 1]$, we construct the low-frequency interpolation path and its target velocity field as

$$\mathbf{z}_{\mathrm{L}}(t) = (1-t)\mathbf{x}_{\mathrm{L}} + t\mathbf{y}_{\mathrm{L}}, \qquad \mathbf{v}^{*}_{\mathrm{L}} = \mathbf{y}_{\mathrm{L}} - \mathbf{x}_{\mathrm{L}}. \tag{2}$$

The Coarse Flow network predicts the low-frequency velocity and target coefficients by

$$\widehat{\mathbf{v}}_{\mathrm{L}} = \mathcal{F}_{\mathrm{L}}\big(\mathbf{z}_{\mathrm{L}}(t), t\big), \qquad \widehat{\mathbf{y}}_{\mathrm{L}} = \mathbf{z}_{\mathrm{L}}(t) + (1-t)\widehat{\mathbf{v}}_{\mathrm{L}}. \tag{3}$$

**Fine Flow.** We construct the high-frequency interpolation path and its target velocity field as

$$\mathbf{z}_{\mathrm{H}}(t) = (1-t)\mathbf{x}_{\mathrm{H}} + t\mathbf{y}_{\mathrm{H}}, \qquad \mathbf{v}^{*}_{\mathrm{H}} = \mathbf{y}_{\mathrm{H}} - \mathbf{x}_{\mathrm{H}}. \tag{4}$$

Using the predicted low-frequency target coefficients as structural context, the Fine Flow network predicts the high-frequency velocity and target coefficients by

$$\widehat{\mathbf{v}}_{\mathrm{H}} = \mathcal{F}_{\mathrm{H}}\big(\mathbf{z}_{\mathrm{H}}(t), \mathcal{C}(\widehat{\mathbf{y}}_{\mathrm{L}}), t\big), \qquad \widehat{\mathbf{y}}_{\mathrm{H}} = \mathbf{z}_{\mathrm{H}}(t) + (1-t)\widehat{\mathbf{v}}_{\mathrm{H}}. \tag{5}$$

where $\mathcal{C}(\cdot)$ is a lightweight projection that aligns the coarse representation with the high-frequency branch.

The complete OCTA volume is reconstructed by inverse wavelet transform:

$$\widehat{\mathbf{Y}} = \mathcal{W}^{-1}\big(\widehat{\mathbf{y}}_{\mathrm{L}}, \widehat{\mathbf{y}}_{\mathrm{H}}\big). \tag{6}$$

**Flow Matching Objective.** Given paired wavelet coefficients, the CF-Flow loss supervises the two frequency-specific velocity fields:

$$\mathcal{L}_{\mathrm{CF}} = \mathbb{E}_{t,\mathbf{X},\mathbf{Y}} \left[ \|\widehat{\mathbf{v}}_{\mathrm{L}} - \mathbf{v}_{\mathrm{L}}^{*}\|_2^2 + \lambda_{\mathrm{H}} \|\widehat{\mathbf{v}}_{\mathrm{H}} - \mathbf{v}_{\mathrm{H}}^{*}\|_2^2 \right]. \tag{7}$$

where $\lambda_{\mathrm{H}}$ controls the relative contribution of high-frequency flow supervision.

### 3.3 Multi-Scale Z-Aware Structural Prior

Paired OCT-to-OCTA translation must preserve the depth-dependent retinal structure that provides the spatial scaffold for vascular appearance. However, a uniform voxel-wise reconstruction objective is dominated by extensive low-intensity background and treats all depth locations equally. It therefore provides limited emphasis on retinal layers and local structural transitions that are informative for OCTA synthesis. To address this issue, we introduce a Multi-Scale Z-Aware (MSZ) structural prior that derives a dense structural weighting map directly from the input OCT volume.

**Multi-Scale Z-Aware Cues.** Let $\mathbf{X} \in \mathbb{R}^{H\times W\times Z}$ denote an OCT volume, where $z$ indexes the retinal-depth direction. We characterize its structural saliency from three complementary perspectives. First, the global depth cue summarizes the average OCT response at each depth location and broadcasts it to the volume:

$$\mathbf{G}(i,j,z) = \mathcal{N}\left( \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \mathbf{X}(u,v,z) \right), \tag{8}$$

where $\mathcal{N}(\cdot)$ denotes normalization. This cue captures the global retinal-depth profile and suppresses depth ranges dominated by background. Second, a local cue preserves spatially varying structural responses through local aggregation:

$$\mathbf{L} = \mathcal{N}\left(\mathrm{AvgPool}_{r}(\mathbf{X})\right). \tag{9}$$

Third, the Z-gradient cue highlights depth-wise structural transitions:

$$\mathbf{D}(i,j,z) = \mathcal{N}\left(|\mathbf{X}(i,j,z) - \mathbf{X}(i,j,z-1)|\right). \tag{10}$$

The three cues respectively encode global depth localization, local retinal structure, and inter-layer variation. We fuse them into an MSZ structural map,

$$\mathbf{M}_{\mathrm{MSZ}} = \mathcal{N}\left(\mathbf{G} + \mathbf{L} + \mathbf{D}\right), \tag{11}$$

which assigns larger responses to structurally informative retinal regions.

**MSZ-Guided Reconstruction.** We use $\mathbf{M}_{\mathrm{MSZ}}$ to emphasize the reconstruction of depth-aware structural regions in the generated OCTA volume. Given the generated volume $\widehat{\mathbf{Y}}$ and its paired target $\mathbf{Y}$, the MSZ objective is defined as

$$\mathcal{L}_{\mathrm{MSZ}} = \left\| (1 + \eta \mathbf{M}_{\mathrm{MSZ}}) \odot \left( \widehat{\mathbf{Y}} - \mathbf{Y} \right) \right\|_1, \tag{12}$$

We set $\eta = 0.5$ to control the strength of voxel-wise structural emphasis. This objective complements CF-Flow by directing reconstruction capacity toward anatomically relevant depth regions while retaining voxel-wise supervision over the complete volume.

### 3.4 Vessel Auxiliary Supervision

Although CF-Flow separates low-frequency structure from high-frequency detail, the high-frequency objective alone does not explicitly distinguish clinically meaningful vascular patterns from nonvascular texture. This is challenging for OCTA, where vessel signals are sparse and thin branches contribute weakly to volume reconstruction loss. We therefore introduce VesselAux to provide explicit vessel-level supervision to the Fine Flow branch. The architecture of the Vessel Head and the construction of the vessel target by the vessel labeler (SAM-OCTA (Wang et al., 2024)) are described in Appendix A.

**Fine-Detail Vessel Prediction.** The predicted high-frequency coefficients $\widehat{\mathbf{y}}_{\mathrm{H}}$ encode the detail component synthesized by Fine Flow. We first reconstruct a high-frequency detail volume by setting the low-frequency coefficients to zero:

$$\widehat{\mathbf{Y}}_{\mathrm{H}} = \mathcal{W}^{-1}\left(\mathbf{0}, \widehat{\mathbf{y}}_{\mathrm{H}}\right). \tag{13}$$

A lightweight Vessel Head $\mathcal{H}_{\mathrm{v}}(\cdot)$ then maps this detail volume to a dense vessel probability map:

$$\widehat{\mathbf{M}}_{\mathrm{v}} = \mathcal{H}_{\mathrm{v}}\big(\widehat{\mathbf{Y}}_{\mathrm{H}}\big). \tag{14}$$

By predicting vessels from the fine-detail representation rather than the full translated OCTA, VesselAux directly couples vascular supervision with the branch responsible for high-frequency synthesis.

**Vessel-Level Supervision.** For each real OCTA target $\mathbf{Y}$, a pretrained vessel labeler $\mathcal{S}_{\mathrm{v}}(\cdot)$ provides the target vessel map,

$$\mathbf{M}_{\mathrm{v}} = \mathcal{S}_{\mathrm{v}}(\mathbf{Y}). \tag{15}$$

We supervise the predicted vessel map with a joint binary cross-entropy and Dice objective:

$$\mathcal{L}_{\mathrm{Vessel}} = \mathcal{L}_{\mathrm{BCE}}\big(\widehat{\mathbf{M}}_{\mathrm{v}}, \mathbf{M}_{\mathrm{v}}\big) + \mathcal{L}_{\mathrm{Dice}}\big(\widehat{\mathbf{M}}_{\mathrm{v}}, \mathbf{M}_{\mathrm{v}}\big). \tag{16}$$

This objective encourages the predicted vascular structures to maintain better continuity, preserve subtle vessel branches, and reduce missing or spurious vascular responses.

### 3.5 Training Objective and Inference

The complete training objective combines frequency-specific flow matching, MSZ-guided reconstruction, and vessel-level supervision:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CF}} + \lambda_{\mathrm{MSZ}}\mathcal{L}_{\mathrm{MSZ}} + \lambda_{\mathrm{Vessel}}\mathcal{L}_{\mathrm{Vessel}}, \tag{17}$$

where $\lambda_{\mathrm{MSZ}}$ and $\lambda_{\mathrm{Vessel}}$ balance the contributions of the two structural priors. The three terms jointly promote coherent frequency-domain transport, depth-aware reconstruction, and vascular-detail preservation.Starting from $\mathbf{z}_{\mathrm{L}}^0 = \mathbf{x}_{\mathrm{L}}$ and $\mathbf{z}_{\mathrm{H}}^0 = \mathbf{x}_{\mathrm{H}}$, we integrate the learned fields from $t = 0$ to 1 using $K$ Euler steps. We use $K = 3$ in all experiments, as supported by the sampling analysis in Section J. With step size $\Delta t = 1/K$, each step first predicts the low-frequency velocity and its endpoint condition:

$$\widehat{\mathbf{v}}_{\mathrm{L}}^k = \mathcal{F}_{\mathrm{L}}(\mathbf{z}_{\mathrm{L}}^k, t_k), \qquad \widehat{\mathbf{y}}_{\mathrm{L}}^k = \mathbf{z}_{\mathrm{L}}^k + (1 - t_k)\widehat{\mathbf{v}}_{\mathrm{L}}^k, \tag{18}$$

$$\widehat{\mathbf{v}}_{\mathrm{H}}^k = \mathcal{F}_{\mathrm{H}}\big(\mathbf{z}_{\mathrm{H}}^k, \mathcal{C}(\widehat{\mathbf{y}}_{\mathrm{L}}^k), t_k\big). \tag{19}$$

The two states are then updated by Euler integration:

$$\mathbf{z}_{\mathrm{L}}^{k+1} = \mathbf{z}_{\mathrm{L}}^k + \Delta t\widehat{\mathbf{v}}_{\mathrm{L}}^k, \qquad \mathbf{z}_{\mathrm{H}}^{k+1} = \mathbf{z}_{\mathrm{H}}^k + \Delta t\widehat{\mathbf{v}}_{\mathrm{H}}^k. \tag{20}$$

The translated OCTA volume is recovered by

$$\widehat{\mathbf{Y}} = \mathcal{W}^{-1}\big(\mathbf{z}_{\mathrm{L}}^K, \mathbf{z}_{\mathrm{H}}^K\big). \tag{21}$$

## 4 Experiments

### 4.1 Experiment Setup

**Dataset.** We evaluate DuoPrior-CFFlow on the OCTA-500 dataset (Li et al., 2024a), which provides paired three-dimensional OCT and OCTA volumes at two fields of view: OCTA-6M and OCTA-3M. The OCTA-6M subset contains 300 paired volumes and is partitioned into 180 training, 20 validation, and 100 test cases. The OCTA-3M subset contains 200 paired volumes and is partitioned into 120 training, 16 validation, and 64 test cases, providing an additional acquisition scale for assessing robustness. All methods use the same fixed disjoint data partitions within each subset. Further details on the OCTA-500 dataset (Li et al., 2024a) and its annotations are provided in Appendix B, and details of the data preprocessing are provided in Appendix C.

**Implementation Details.** We implement DuoPrior-CFFlow in PyTorch using MONAI's `DiffusionModelUNet`. All models are trained on a single 48 GB NVIDIA GPU with a batch size of one. For each method, we select the best checkpoint according to validation-set image quality and report its performance on the held-out test set. We optimize the network using Adam with $\beta_1 = 0.5$ and a learning rate of $5 \times 10^{-5}$. We set $\lambda_{\mathrm{H}} = 2$, $\lambda_{\mathrm{MSZ}} = 1.0$, and $\lambda_{\mathrm{Vessel}} = 0.02$, and assign equal weights to the BCE and Dice components of $\mathcal{L}_{\mathrm{Vessel}}$. Following FlowOCT (Xiong et al., 2026), we pre-specify **three Euler steps** ($K = 3$) for all main experiments (see Appendix J for

| Method | OCTA-3M | | | | | OCTA-6M | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ | MAE ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
| Pix2Pix 2D (Isola et al., 2017) | 1.37 | 28.38 | 86.19 | 5.85 | 36.78 | 1.29 | 28.43 | 85.96 | 5.56 | **15.21** |
| Pix2Pix 3D (Isola et al., 2017) | 1.35 | 28.56 | 86.22 | 7.62 | 85.96 | 1.37 | 28.40 | 85.41 | 7.07 | 75.50 |
| TransPro (Li et al., 2024b) | 1.33 | 28.47 | 86.30 | 8.83 | 96.67 | 1.26 | 28.73 | 86.40 | 8.99 | 92.30 |
| BBDM 2D (Li et al., 2023) | 1.53 | 27.28 | 82.17 | 11.74 | 129.54 | 1.55 | 27.34 | 82.54 | 9.24 | 113.81 |
| BBDM 3D (Li et al., 2023) | 1.36 | 28.22 | 84.66 | 16.37 | 183.24 | 1.39 | 28.59 | 84.78 | 15.39 | 143.03 |
| MuTri (Chen et al., 2025) | 1.29 | 28.33 | 86.15 | 16.31 | 118.37 | 1.26 | 28.69 | 86.23 | 10.43 | 107.45 |
| FlowOCT (Xiong et al., 2026) | <u>1.27</u> | <u>28.75</u> | <u>86.39</u> | <u>5.54</u> | **16.66** | <u>1.22</u> | <u>28.80</u> | <u>86.82</u> | <u>5.36</u> | 18.76 |
| **DuoPrior-CFFlow (ours)** | **1.19** | **29.07** | **87.21** | **5.31** | <u>19.47</u> | **1.13** | **29.38** | **87.65** | **5.12** | <u>17.48</u> |

Table 1: Quantitative comparison on the OCTA-500 dataset. MAE, SSIM, and LPIPS are reported as percentages. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

the fixed-checkpoint post-hoc analysis). Detailed architecture , branch configurations, and training parameters are provided in Appendix F.

**Evaluation Metrics.** We evaluate volumetric reconstruction using MAE, PSNR and SSIM (Horé & Ziou, 2010). We also report learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018) and Fréchet inception distance (FID) (Heusel et al., 2017) to assess perceptual similarity and distributional fidelity. Definitions and computation procedures for the reconstruction metrics and the Section 4.3 structural preservation and downstream utility evaluations are given in Appendix D.

### 4.2 Comparison with Baseline Methods

We compare DuoPrior-CFFlow with Pix2Pix2D (Isola et al., 2017), Pix2Pix3D (Isola et al., 2017), TransPro (Li et al., 2024b), BBDM2D and BBDM3D (Li et al., 2023), MuTri (Chen et al., 2025), and FlowOCT (Xiong et al., 2026). These public methods share the same fixed OCTA-500 partitions and unified preprocessing protocol for training and evaluation. Detailed descriptions and implementation configurations are provided in Appendix E. Efficiency statistics are reported in Appendix I.

**Experimental Results on OCTA-500 Dataset.** Table 1 summarizes the quantitative comparison on the OCTA-3M and OCTA-6M subsets. On the 3M subset, DuoPrior-CFFlow obtains the best MAE, PSNR, SSIM, and LPIPS, while FlowOCT (Xiong et al., 2026) achieves the lowest FID. DuoPrior-CFFlow improves over FlowOCT (Xiong et al., 2026) by $0.8‰$ in MAE, $0.32$ dB in PSNR, $0.82\%$ in SSIM, and $0.23\%$ in LPIPS. On the 6M subset, DuoPrior-CFFlow again achieves the best MAE, PSNR, SSIM, and LPIPS. DuoPrior-CFFlow improves over FlowOCT (Xiong et al., 2026) by $0.9‰$ in MAE, $0.58$ dB in PSNR, $0.83\%$ in SSIM, and $0.24\%$ in LPIPS. Pix2Pix 2D (Isola et al., 2017) obtains the lowest FID on this subset, while DuoPrior-CFFlow remains the second-best method. Across both fields of view, the results show that the proposed coarse-to-fine transport improves reconstruction and structural similarity while remaining competitive in distributional fidelity. This trend supports the use of separate low-frequency structural and high-frequency vascular-detail transport paths for paired volumetric OCT-to-OCTA translation.

**Visualization Results on OCTA-500 Dataset.** For qualitative evaluation, Figure 3 shows OCTA slices and enlarged ROIs from the OCTA-500 dataset (Li et al., 2024a). GAN-based methods may produce stripe-like artifacts, unstable vessel brightness, or incomplete fine structures. Diffusion baselines generate smoother outputs but tend to lose thin vessels and local vascular contrast. TransPro (Li et al., 2024b) and FlowOCT (Xiong et al., 2026) preserve the main vascular layout more effectively, yet still show local vessel discontinuities, spurious responses, or blurred fine branches. In contrast, DuoPrior-CFFlow produces clearer and more continuous vascular structures, with fewer false responses and better preserves fine branches in the enlarged regions. Its visual appearance is closest to the ground truth, consistent with the quantitative improvements reported above.

### 4.3 Structural Preservation and Downstream Utility

We further assess structural fidelity and downstream utility through three complementary tasks: three-dimensional vessel structure consistency, FAZ-region vascular exclusion, and disease classification. These analyses evaluate whether the translated OCTA preserves vascular anatomy, avoids

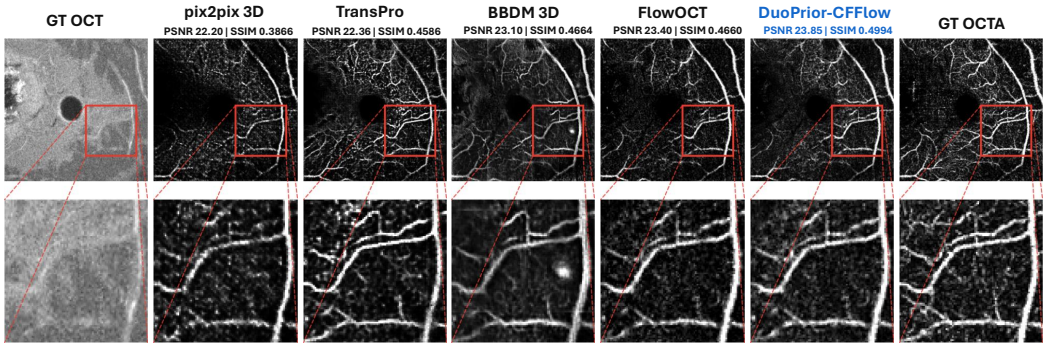



Figure 3: The visualization results of OCT-to-OCTA translation methods on the OCTA-500 dataset. The figure shows a representative slice and the corresponding enlarged region of interest.

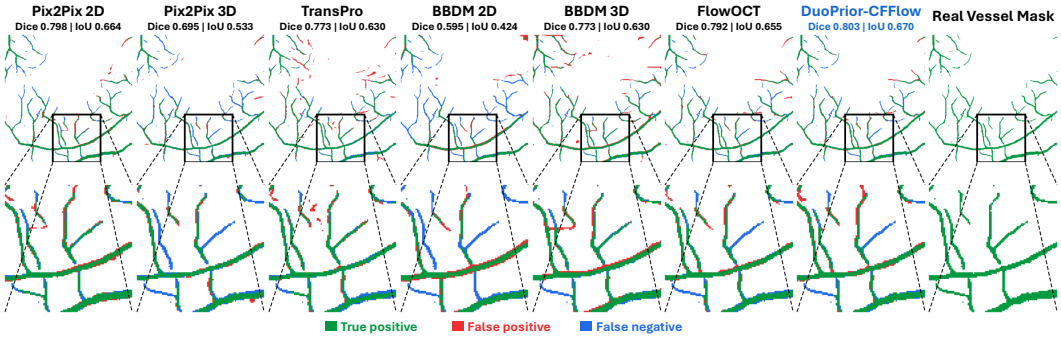



Figure 4: Representative slice-level visualization of vessel correspondence on the OCTA-6M. Colored overlays indicate matched, missing, and spurious responses against the real 3D vessel mask.

implausible vessels in the foveal avascular zone, and retains clinically meaningful disease information.

| Method | Dice ↑ (A / B) | IoU ↑ (A / B) | Precision ↑ | Recall ↑ |
|---|---|---|---|---|
| Pix2Pix 2D | 0.520 / 0.583 | 0.360 / 0.426 | 0.421 | 0.745 |
| Pix2Pix 3D | 0.522 / 0.579 | 0.359 / 0.419 | **0.490** | 0.585 |
| TransPro | 0.464 / 0.573 | 0.309 / 0.413 | 0.338 | 0.784 |
| BBDM 2D | 0.309 / 0.429 | 0.220 / 0.297 | 0.235 | **0.804** |
| BBDM 3D | 0.339 / 0.445 | 0.239 / 0.318 | 0.240 | 0.614 |
| MuTri | 0.438 / 0.521 | 0.287 / 0.356 | 0.382 | 0.714 |
| FlowOCT | 0.533 / 0.630 | 0.373 / 0.474 | 0.437 | 0.732 |
| **DuoPrior-CFFlow** | **0.564 / 0.672** | **0.406 / 0.506** | 0.445 | 0.770 |

Table 2: Three-dimensional vessel structure consistency on the OCTA-6M dataset. The best results are highlighted in **bold**.

| Method | FAZ Occupancy ↓ | FAZ Internal Occupancy ↓ | Excess Leakage ↓ | FAZ / Outside Ratio ↓ |
|---|---|---|---|---|
| Real OCTA | 0.34% | 0.25% | 0.00% | 0.09 |
| Pix2Pix 2D | 1.04% | 0.89% | 0.70% | 0.19 |
| Pix2Pix 3D | 0.53% | 0.40% | 0.22% | 0.16 |
| TransPro | 0.99% | 1.01% | 0.80% | 0.20 |
| BBDM 2D | 3.52% | 3.54% | 3.34% | 0.24 |
| BBDM 3D | 2.46% | 2.82% | 2.30% | 0.43 |
| MuTri | 1.89% | 1.70% | 1.58% | 0.25 |
| FlowOCT | 1.00% | 0.74% | 0.68% | 0.22 |
| **DuoPrior-CFFlow** | **0.40%** | **0.29%** | **0.14%** | **0.12** |

Table 3: FAZ vascular exclusion analysis on the OCTA-6M. Best results are highlighted in **bold**.

#### 4.3.1 Three-Dimensional Vessel Structure Consistency

We evaluate volumetric vessel fidelity by applying the same pretrained vessel labeler described in Appendix A to translated and real OCTA volumes and comparing the corresponding 3D vessel masks. Table 2 reports Dice, IoU, precision, and recall. We report two complementary Dice/IoU values: A is computed over the complete volume, while B is averaged over vessel-dense slices containing at least 3500 vessel pixels in the real 3D vessel mask. DuoPrior-CFFlow achieves the best Dice and IoU in both settings, with 0.564/0.672 Dice and 0.406/0.506 IoU. These results indicate strong volumetric vessel overlap and effective preservation of dense vascular structures.

Figure 4 provides a representative slice-level visualization of vessel correspondence. The colored overlays identify matched, missing, and spurious vessel responses relative to the real 3D vessel mask. Although this view is limited to one slice, it qualitatively illustrates the local error patterns summarized by the volumetric metrics. DuoPrior-CFFlow shows more continuous main vessels and fine branches, with fewer isolated false responses and missing segments, supporting the quantitative evidence of improved three-dimensional vascular consistency.

#### 4.3.2 FAZ Region Vascular Exclusion

We evaluate vascular responses inside the three-dimensional foveal avascular zone (FAZ). Table 3 reports the FAZ-exclusion results. Together, the four metrics characterize predicted vessel signal inside the FAZ, leakage within its internal region, and the relative vessel-density contrast between the FAZ and surrounding tissue. DuoPrior-CFFlow achieves the best results among generated methods on all four metrics, indicating fewer spurious responses and more faithful preservation of the avascular region. Metric definitions and computation details are provided in Appendices D and G.

#### 4.3.3 Disease Classification

We assess the downstream utility of translated OCTA via three-class NORMAL/AMD/DR classification. Table 4 reports OCTA-6M test results. Although its AUC is lower than FlowOCT's, DuoPrior-CFFlow achieves the highest accuracy (66.04%), balanced accuracy (62.37%), and macro-F1 (62.18%) among generated methods, indicating better preservation of disease-discriminative information. Details are in Appendix H.

| Method | Acc. ↑ (%) | Bal. Acc. ↑ (%) | F1 ↑ (%) | AUC ↑ (%) |
|---|---|---|---|---|
| Real OCTA | 83.02 | 80.26 | 81.41 | 95.48 |
| Pix2Pix 2D | 54.72 | 50.51 | 49.00 | 75.09 |
| Pix2Pix 3D | 50.94 | 45.83 | 37.14 | 68.15 |
| TransPro | 49.06 | 56.69 | 49.98 | 78.00 |
| BBDM 2D | 45.28 | 33.33 | 20.78 | 64.28 |
| BBDM 3D | 56.60 | 55.68 | 54.75 | 70.03 |
| MuTri | 50.94 | 51.52 | 50.14 | 66.84 |
| FlowOCT | 56.60 | 56.10 | 52.73 | **78.70** |
| **DuoPrior-CFFlow** | **66.04** | **62.37** | **62.18** | 76.78 |

Table 4: Disease classification results on the OCTA-6M dataset. Real OCTA is for reference only and not ranked. Best results are in **bold**.

### 4.4 Ablation and Sensitivity Analysis

We first compare the single-path full-frequency FlowOCT (Xiong et al., 2026) with CF-Flow only. CF-Flow only outperforms the complete FlowOCT baseline, supporting the dual-path configuration in this comparison. CF-Flow also outperforms Coarse Flow, with gains across all metrics after adding the Fine Flow branch. Adding the MSZ Prior and VesselAux individually (Table 5) further improves performance. The full configuration achieves the best MAE, PSNR, SSIM, and LPIPS, supporting the complementary roles of the two priors.

We further analyze the sensitivity to the MSZ and VesselAux weights in Table 6. Moderate weights maintain stable reconstruction and perceptual quality, whereas stronger VesselAux weighting degrades LPIPS and FID. The main experiments use $\lambda_{\mathrm{MSZ}} = 1.00$ and $\lambda_{\mathrm{Vessel}} = 0.02$; the sensitivity results show that this setting provides a balanced operating point for the two priors.

| Configuration | MAE ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|---|
| Single-Path (FlowOCT) | 1.22 | 28.80 | 86.82 | 5.36 | 18.76 |
| Coarse Flow | 1.34 | 28.61 | 86.25 | 8.79 | 25.63 |
| CF-Flow only | 1.18 | 28.97 | 87.03 | 5.34 | 21.11 |
| CF-Flow + MSZ prior | 1.20 | 29.04 | 87.15 | <u>5.20</u> | 19.99 |
| CF-Flow + VesselAux | <u>1.16</u> | <u>29.11</u> | <u>87.26</u> | 5.24 | **15.42** |
| **DuoPrior-CFFlow** | **1.13** | **29.38** | **87.65** | **5.12** | <u>17.48</u> |

Table 5: Ablation results on the OCTA-6M dataset. MAE, SSIM, and LPIPS are reported as percentages. Best and second-best results are in **bold** and <u>underlined</u>, respectively.

| $\lambda_{\mathrm{MSZ}}$ | $\lambda_{\mathrm{Vessel}}$ | MAE ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|---|---|
| 0.25 | 0.02 | 1.19 | 29.27 | 87.38 | 5.40 | **17.39** |
| 0.50 | 0.02 | <u>1.15</u> | <u>29.34</u> | <u>87.45</u> | <u>5.34</u> | 17.60 |
| 1.00 | 0.02 | **1.13** | **29.38** | **87.65** | **5.12** | <u>17.48</u> |
| 1.00 | 0.10 | 1.18 | 29.21 | 87.33 | 5.90 | 23.34 |
| 1.50 | 0.50 | 1.29 | 28.70 | 86.98 | 7.56 | 25.62 |

Table 6: Sensitivity analysis on the OCTA-6M dataset. MAE, SSIM, and LPIPS are reported as percentages. The gray row denotes the main configuration. Best and second-best results are in **bold** and <u>underlined</u>, respectively.

## 5 Conclusion

In this work, we presented DuoPrior-CFFlow, a dual-prior coarse-to-fine flow-matching framework for paired three-dimensional OCT-to-OCTA translation. CF-Flow models low-frequency retinal structure and high-frequency vascular detail through dedicated wavelet-domain transport paths, while the MSZ Prior provides multi-scale depth-aware guidance and VesselAux reinforces three-dimensional vessel consistency. Experiments on the OCTA-500 dataset (Li et al., 2024a) show that DuoPrior-CFFlow achieves higher reconstruction quality than the compared state-of-the-art methods and better preserves retinal vascular structures. Structural preservation and downstream utility evaluations further show the highest 3D vessel-mask consistency, the lowest vascular occupancy in the FAZ, and the highest disease classification accuracy, demonstrating the potential of reliable OCT-to-OCTA translation from routinely acquired structural OCT volumes.

## Reproducibility Statement

To support reproducibility, we report the dataset subsets, preprocessing procedures, data splits, and evaluation protocols used in our experiments. The implementation details and training model configurations not included in the main paper are described in detail in the appendix. Code is available at `https://anonymous.4open.science/r/duoprior-cfflow-code-6D9F/`.

## AI Use Statement

In this work, we used generative AI tools only to aid or polish the writing and to assist with creating and formatting related tables and visual materials. We did not use generative AI tools to generate research data, conduct experiments, train or evaluate models, or make scientific decisions; the remaining disclosure categories are not applicable to this work. We reviewed all AI-assisted text, tables, and visual materials for accuracy, consistency with the experiments, and potential errors. The authors take responsibility for the final content of this work, including all text, claims, tables, figures, and other artifacts produced with the aid of generative AI.

## References

Zhuangzhuang Chen, Hualiang Wang, Chubin Ou, and Xiaomeng Li. Mutri: Multi-view tri-alignment for oct to octa 3d image translation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20885–20894, 2025. doi: 10.1109/CVPR52734.2025.01945.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space, 2023.

Talisa E De Carlo, Andre Romano, Nadia K Waheed, and Jay S Duker. A review of optical coherence tomography angiography (octa). *International journal of retina and vitreous*, 1(1):5, 2015.

William Foulsham, Jason Chien, Tamara Lee Lenis, and Thanos D. Papakostas. Optical coherence tomography angiography: Clinical utility and future directions. *Journal of VitreoRetinal Diseases*, 6(3):229–242, 2022. PMID: 37008547.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *The Fourteenth International Conference on Learning Representations*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

David Huang, Eric A. Swanson, Charles P. Lin, Joel S. Schuman, William G. Stinson, Warren Chang, Michael R. Hee, Thomas Flotte, Kenton Gregory, Carmen A. Puliafito, and James G. Fujimoto. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Trans. Graph.*, 42(6), December 2023. ISSN 0730-0301. doi: 10.1145/3618373.

Pooya Khosravi, Kun Han, Anthony T. Wu, Arghavan Rezvani, Zexin Feng, and Xiaohui Xie. Xoct: Enhancing oct to octa translation via cross-dimensional supervised multi-scale feature learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part IV*, pp. 695–705, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-04964-3. doi: 10.1007/978-3-032-04965-0_65.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961, 2023. doi: 10.1109/CVPR52729.2023.00194.

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, 2024a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103092.

Pei Lin Li, Céline O'Neil, Samin Saberi, Kenneth Sinder, Kathleen Wang, Bingyao Tan, Zohreh Hosseinaee, Kostadinka Bizhevat, and Vasudevan Lakshminarayanan. Deep learning algorithm for generating optical coherence tomography angiography (OCTA) maps of the retinal vasculature. In Michael E. Zelinski, Tarek M. Taha, Jonathan Howe, Abdul A. S. Awwal, and Khan M. Iftekharuddin (eds.), *Applications of Machine Learning 2020*, volume 11511, pp. 1151109. International Society for Optics and Photonics, SPIE, 2020. doi: 10.1117/12.2568629.

Shuhan Li, Dong Zhang, Xiaomeng Li, Chubin Ou, Lin An, Yanwu Xu, Weihua Yang, Yanchun Zhang, and Kwang-Ting Cheng. Vessel-promoted oct to octa image translation by heuristic contextual constraints. *Medical Image Analysis*, 98:103311, 2024b. ISSN 1361-8415.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023.

Andrea Moschetto, Lemuel Puglisi, Alec Sargood, Pierluigi Dell'Acqua, Francesco Guarnera, Sebastiano Battiato, and Daniele Ravì. Benchmarking gans, diffusion models, and flow matching for t1w-to-t2w mri translation. In Emanuele Rodolà, Fabio Galasso, and Iacopo Masi (eds.), *Image Analysis and Processing - ICIAP 2025 Workshops*, pp. 429–440, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-11317-7.

Bing Pan, Zexuan Ji, and Qiang Chen. Multigan: Multi-domain image translation from oct to octa. In Shiqi Yu, Zhaoxiang Zhang, Pong C. Yuen, Junwei Han, Tieniu Tan, Yike Guo, Jianhuang Lai, and Jianguo Zhang (eds.), *Pattern Recognition and Computer Vision*, pp. 336–347, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-18910-4.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 Real-Time Live!*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363150. doi: 10.1145/3306305.3332370.

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, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393379. doi: 10.1145/3528233.3530757.

Anna CS Tan, Gavin S Tan, Alastair K Denniston, Pearse A Keane, Marcus Ang, Dan Milea, Usha Chakravarthy, and Chui Ming Gemmy Cheung. An overview of the clinical applications of optical coherence tomography angiography. *Eye*, 32(2):262–286, 2018.

Chengliang Wang, Xinrun Chen, Haojian Ning, and Shiying Li. SAM-OCTA: A Fine-Tuning Strategy for Applying Foundation Model OCTA Image Segmentation Tasks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 363. IEEE, April 2024. doi: 10.1109/ICASSP48485.2024.10446904.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 44(12):4960–4972, 2025. doi: 10.1109/TMI.2025.3585372.

Ze Xiong, Dehui Qiu, Liguo Deng, Longfei Zhou, Zhetao Xu, Fa Zhang, and Xiaohua Wan. Flowoct: Wavelet flow matching for oct-to-octa translation. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 1911–1919. International Joint Conferences on Artificial Intelligence Organization, 8 2026. Main Track.

Milad Yazdani, Yasamin Medghalchi, Pooria Ashrafian, Ilker Hacihaliloglu, and Dena Shahriari. Flow matching for medical image synthesis: Bridging the gap between speed and quality. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part XVI*, pp. 216–226, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-05324-4. doi: 10.1007/978-3-032-05325-1_21.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Ziyue Zhang, Zexuan Ji, Qiang Chen, Songtao Yuan, and Wen Fan. Texture-guided u-net for oct-to-octa generation. In Huimin Ma, Liang Wang, Changshui Zhang, Fei Wu, Tieniu Tan, Yaonan Wang, Jianhuang Lai, and Yao Zhao (eds.), *Pattern Recognition and Computer Vision*, pp. 42–52, Cham, 2021. Springer International Publishing. ISBN 978-3-030-88013-2.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

# APPENDIX

## A VESSELAUX SUPPLEMENTARY DETAILS

**Vessel Head.** The Vessel Head is implemented as a compact three-dimensional convolutional prediction head attached to the Fine Flow branch. Importantly, it does not take the final reconstructed OCTA volume as input. During training, the predicted high-frequency wavelet representation $\widehat{\mathbf{y}}_{\mathrm{H}}$ is converted into a high-frequency-only image-space response by setting the low-frequency component to zero and applying the inverse three-dimensional Haar wavelet transform:

$$\widehat{\mathbf{Y}}_{\mathrm{H}} = \mathrm{IDWT}_{3\mathrm{D}}(0, \widehat{\mathbf{y}}_{\mathrm{H}}). \tag{22}$$

This isolates the information carried by the high-frequency branch before vessel supervision is applied. The resulting volume has $C_{\mathrm{out}}$ channels, where $C_{\mathrm{out}}$ is the output channel count of the generation task. The convolutional logit predictor $h_v$ is

$$h_v = \mathrm{Conv3D}(C_{\mathrm{out}}, C_h, 3, \mathrm{padding} = 1) \;\rightarrow\; \mathrm{SiLU} \;\rightarrow\; \mathrm{Conv3D}(C_h, 1, 1), \tag{23}$$

where the intermediate width is $C_h = 8$ in the main configuration. Thus, the first layer uses a $3\times3\times3$ kernel with unit padding to aggregate local three-dimensional context, the SiLU nonlinearity provides the only intermediate activation, and the final $1 \times 1 \times 1$ convolution produces one vessel logit per voxel. The output is denoted by

$$S_v = h_v(\widehat{\mathbf{Y}}_{\mathrm{H}}), \qquad \widehat{\mathbf{M}}_{\mathrm{v}} = \mathcal{H}_{\mathrm{v}}(\widehat{\mathbf{Y}}_{\mathrm{H}}) = \sigma(S_v), \tag{24}$$

where $S_v$ is the logit volume and $\widehat{\mathbf{M}}_{\mathrm{v}}$ is the corresponding vessel probability volume. Consistent with the main text, $\mathcal{H}_{\mathrm{v}} = \sigma \circ h_v$ denotes the complete Vessel Head including the final sigmoid. The spatial resolution is unchanged throughout the head, so the prediction remains aligned with the original three-dimensional vessel target.

**SAM-OCTA.** SAM-OCTA (Wang et al., 2024) is a publicly available PyTorch implementation that adapts the Segment Anything Model (SAM) to OCTA segmentation through LoRA fine-tuning. Its repository explicitly supports OCTA-500 segmentation tasks and provides code for targets including large vessels, capillaries, arteries, veins, and the foveal avascular zone, with global and local prediction modes. Here, we use its OCTA-500-compatible vessel-segmentation implementation to obtain slice-wise masks.

**3D vessel mask.** We apply the publicly available SAM-OCTA (Wang et al., 2024) model adapted to OCTA-500 vessel labels to OCTA slices. We stack the resulting 2D masks in their original depth order and manually assemble and align them to obtain a 3D vessel mask in the OCTA volume coordinates. All slice-wise masks and the assembled 3D masks were reviewed by clinicians at our collaborating hospital for vessel delineation and consistency across depth, and were confirmed as suitable 3D vessel annotations. The reviewed 3D masks serve as the VesselAux target and the reference for structural and FAZ evaluations.

**Auxiliary objective and gradient routing.** Vessel supervision is applied to the logits $S_v$ and the aligned binary target $M_v$. For numerical stability, the implementation uses `BCEWithLogits` on $S_v$, which is mathematically equivalent to the probability-based BCE in the main text:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[M_{v,i}\log\sigma(S_{v,i}) + (1-M_{v,i})\log\left(1-\sigma(S_{v,i})\right)\right], \tag{25}$$

where $N$ is the number of voxels. The Dice term uses the sigmoid probabilities rather than logits, and is computed as a batch-averaged soft Dice loss. For sample $b$, let $p_b = \sigma(S_{v,b})$ and $m_b = M_{v,b}$; then

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{1}{B}\sum_{b=1}^{B}\frac{2\langle p_b, m_b\rangle + 1}{\|p_b\|_1 + \|m_b\|_1 + 1}, \tag{26}$$

where $B$ is the batch size and the additive constant one is the smoothing term used in the implementation. The auxiliary loss is

$$\mathcal{L}_{\mathrm{VesselAux}} = w_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}} + w_{\mathrm{Dice}}\mathcal{L}_{\mathrm{Dice}}. \tag{27}$$

The main configuration uses $w_{\mathrm{BCE}} = 1$ and $w_{\mathrm{Dice}} = 1$. The total objective multiplies this auxiliary term by $\lambda_{\mathrm{Vessel}}$; the main experiments use $\lambda_{\mathrm{Vessel}} = 0.02$. Because $\widehat{\mathbf{Y}}_{\mathrm{H}}$ is reconstructed directly from the predicted high-frequency coefficients, gradients from both BCE and Dice pass through the Vessel Head and the inverse wavelet reconstruction into the Fine Flow parameters. The auxiliary branch therefore emphasizes vessel-sensitive high-frequency responses without changing the coarse transport objective or adding a separate segmentation network to the inference pipeline.

## B OCTA-500 Dataset Supplementary Description

**Dataset contents.** The OCTA-500 dataset (Li et al., 2024a) provides paired structural OCT and OCTA data together with complementary annotations and clinical labels. The paired volumes are released at two fields of view, OCTA-3M and OCTA-6M, and preserve voxel-wise correspondence between the structural and angiographic modalities. In addition to volumetric data, the dataset provides slice-level 2D data that can be used to construct and evaluate slice-wise image translation models. The dataset also contains three-dimensional foveal avascular zone (FAZ) annotations and disease categories associated with the subjects.

**Use of SAM-OCTA vessel masks.** We use the OCTA-500-adapted SAM-OCTA (Wang et al., 2024) implementation to obtain the 3D vessel masks for VesselAux supervision and structural and FAZ-region evaluation. Details of the mask-generation procedure are provided in Appendix A.

**Use of three-dimensional FAZ annotations.** The official three-dimensional FAZ annotation is used to evaluate whether a translated OCTA volume introduces vascular responses inside the anatomically avascular region. The FAZ annotation is aligned with the corresponding OCTA volume and is used only for downstream structural evaluation, not as an input or supervision signal for the generator. For each translated volume, we measure vascular occupancy inside the FAZ, internal FAZ occupancy, excess leakage, and the FAZ-to-outside vessel-density ratio. These measurements quantify whether the generated volume maintains the expected vessel exclusion of the FAZ.

**Disease labels and selected categories.** The OCTA-500 dataset (Li et al., 2024a) provides multiple disease categories, but several categories contain relatively few cases. To obtain a sufficiently stable and comparable downstream evaluation, we select three categories with adequate representation: NORMAL, age-related macular degeneration (AMD), and diabetic retinopathy (DR). We formulate the task as three-class disease classification. A classifier is trained using real OCTA volumes from the training partition with stratified five-fold cross-validation, and the resulting classifier is applied to real and translated OCTA test volumes under the same protocol. We report accuracy, balanced accuracy, macro-F1, and macro-AUC. The disease labels are used only for this downstream evaluation and do not participate in OCT-to-OCTA generator training.

## C OCTA-500 Dataset Preprocessing

The paired OCT and OCTA volumes are intensity-normalized to $[0, 1]$ while preserving their voxel-wise correspondence. For OCTA-6M, each $400 \times 400 \times 640$ volume is linearly resampled to $400 \times 400 \times 256$, followed by an in-plane center crop with indices $72{:}328$ along both spatial axes, yielding a $256^3$ input. For OCTA-3M, each $304 \times 304 \times 640$ volume is linearly resampled to $304 \times 304 \times 256$. We retain the native in-plane field of view and draw paired $256^3$ spatial crops for training. The same preprocessing protocol is applied consistently to the paired OCT and OCTA volumes. During training, spatial crops are sampled from the paired volumes, with the OCT and OCTA crop locations kept identical to preserve voxel-wise correspondence. During validation and testing, cropping is deterministic: OCTA-6M uses the in-plane center crop $72{:}328$, and OCTA-3M uses $24{:}280$; no random crop sampling is applied. All test predictions and real OCTA volumes remain in these processed coordinates for volumetric metrics and downstream analyses.

## D Evaluation Metrics

We use the following metrics throughout the paper. Unless otherwise specified, metric values are computed for each test volume and then averaged over the corresponding test set. For reconstruction metrics, the paired OCTA volumes are normalized to $[0, 1]$, and all voxel-wise comparisons use the same spatially aligned field of view.

**Volumetric reconstruction metrics.** Mean absolute error (MAE) measures the average absolute voxel difference between the translated volume $\hat{Y}$ and the real OCTA volume $Y$:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{Y}_i - Y_i|. \tag{28}$$

Here, $N$ is the number of voxels in the evaluated volume. Peak signal-to-noise ratio (PSNR) is computed from the mean squared error (MSE):

$$\mathrm{PSNR} = 10 \log_{10} \left( \frac{I_{\max}^2}{\mathrm{MSE}} \right), \qquad \mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2, \tag{29}$$

where $I_{\max} = 1$ for the normalized volumes. Structural similarity (SSIM) compares local luminance, contrast, and structural statistics between the translated and real volumes. We compute SSIM on spatially aligned slices or local windows using the standard formulation and average the resulting values over the evaluated volume.

**Perceptual and distributional metrics.** LPIPS is computed slice-wise with the fixed pretrained AlexNet-based LPIPS network. Each normalized grayscale axial OCTA slice is replicated to three channels before the corresponding translated and real slices are passed to LPIPS, and the distances are averaged over all slices and test volumes. For FID, each axial slice of every test volume is treated as one sample. The translated and real grayscale slices are exported as 8-bit PNG images and evaluated with CleanFID using its Inception-based feature extractor. Lower LPIPS and FID indicate closer perceptual similarity and feature-distribution fidelity, respectively. If the feature distributions have means $\mu_g$, $\mu_r$ and covariances $\Sigma_g$, $\Sigma_r$, FID is computed as

$$\mathrm{FID} = \|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}\left( \Sigma_g + \Sigma_r - 2(\Sigma_g \Sigma_r)^{1/2} \right). \tag{30}$$

Lower FID indicates a smaller distributional discrepancy. For LPIPS, we use the pretrained AlexNet-based network provided by the official LPIPS package, including its learned perceptual calibration weights. For FID, we use the default pretrained Inception-v3 feature extractor provided by clean-fid. These feature networks remain fixed during evaluation and are not fine-tuned on OCTA-500 dataset (Li et al., 2024a).

**Three-dimensional vessel structure consistency.** The translated and real OCTA volumes are converted to binary vessel maps using the same vessel-segmentation protocol. Let $P$ denote the predicted vessel set and $G$ the real vessel set. True positives, false positives, and false negatives are defined voxel-wise from $P$ and $G$. Dice and intersection-over-union (IoU) are computed as

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \qquad \mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|}. \tag{31}$$

Precision and recall are computed from TP, FP, and FN as

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \qquad \mathrm{Recall} = \frac{TP}{TP + FN}. \tag{32}$$

We report two complementary aggregation settings. A is the complete-volume metric computed over the full three-dimensional vessel maps. B is the vessel-dense-slice metric: only slices whose real 3D vessel mask contains at least 3500 vessel pixels are retained, and the metric is averaged over the retained slices. This filtering reduces the influence of nearly empty slices when assessing local vessel agreement.

**FAZ-region vascular exclusion.** Let $F$ be the official three-dimensional FAZ mask, $F_{\mathrm{int}}$ its defined interior region, and $O$ the voxels outside the FAZ. Let $V$ denote the binary vessel map obtained from a translated or real OCTA volume. FAZ occupancy measures the fraction of FAZ voxels occupied by vessels:

$$\mathrm{FAZ\ Occupancy} = \frac{|V \cap F|}{|F|}. \tag{33}$$

Internal FAZ occupancy applies the same calculation to $F_{\mathrm{int}}$, reducing sensitivity to the FAZ boundary. Excess leakage measures the additional vascular occupancy inside the FAZ relative to real OCTA:

$$\mathrm{Excess\ Leakage} = \max\left(\mathrm{Occ}_{\mathrm{pred}}(F) - \mathrm{Occ}_{\mathrm{real}}(F), 0\right). \tag{34}$$

The FAZ-to-outside vessel-density ratio is

$$\mathrm{FAZ/Outside\ Ratio} = \frac{|V \cap F|/|F|}{|V \cap O|/|O|}. \tag{35}$$

Lower values indicate better exclusion of vascular responses from the avascular region. The same official FAZ coordinate system is used for all methods.

**Downstream disease classification.** For the downstream task, we use the three selected categories NORMAL, AMD, and DR. A classifier is trained only with real OCTA volumes from the training partition under stratified five-fold cross-validation, and is then applied to real and translated OCTA volumes in the held-out test partition. Accuracy is the fraction of correctly classified volumes. Balanced accuracy is the mean of the three per-class recalls. Macro-F1 is the arithmetic mean of the class-wise F1 scores. Macro-AUC is the mean one-vs-rest area under the receiver operating characteristic curve across the three classes. These metrics are computed using the same labels and test cases for all methods.

## E Baseline Methods

This section describes the baseline methods and the implementation choices used for comparison. All baselines use the same disjoint partitions and paired OCT–OCTA preprocessing. Two-dimensional methods process individual aligned slices, while volumetric methods process three-dimensional crops. Slice-wise outputs are restored to the original order before volume-level evaluation.

**Pix2Pix 2D and Pix2Pix 3D.** Pix2Pix 2D (Isola et al., 2017) is a conditional adversarial image-translation framework. The generator receives one OCT slice and predicts the paired OCTA slice,

while the discriminator distinguishes real and translated pairs. Slices are processed independently without neighboring-slice context. For Pix2Pix 3D, we apply the same paired conditional adversarial formulation to aligned three-dimensional crops. The generator predicts an OCTA crop from the corresponding OCT crop, and the discriminator operates on the volumetric pair. Both variants are trained from scratch without pretrained weights.

**TransPro.** TransPro (Li et al., 2024b) is a three-dimensional adversarial OCT-to-OCTA framework that combines volumetric image translation with projection-guided contextual constraints. Its generator predicts an OCTA volume from the paired OCT volume, while the discriminator evaluates the realism of the generated volumetric pair. The projection-guided module aggregates the generated and target volumes along multiple directions and uses these projections to constrain global vascular layout and contextual consistency. This design helps the model preserve large-scale vessel organization while retaining three-dimensional spatial correspondence. Following the method README, we use its volumetric generator and discriminator, load the provided frozen module weights during training, and train the remaining components on the paired OCT–OCTA data.

**BBDM 2D and BBDM 3D.** Brownian Bridge Diffusion Models (Li et al., 2023) formulate paired translation as diffusion along a bridge between source and target domains. BBDM 2D is trained and sampled independently on slices, whereas BBDM 3D is applied to aligned volumetric crops. The 3D implementation uses memory-compatible patches and overlapping sliding-window inference, followed by patch merging in common volume coordinates.

**MuTri.** MuTri (Chen et al., 2025) is a three-dimensional OCT-to-OCTA translation framework that learns modality-specific volumetric representations and uses multi-view tri-alignment to connect the generated OCTA volume with cross-view projection features. Its alignment objective encourages voxel-level correspondence and consistent structural semantics across views. We follow its staged training strategy: the first stage learns modality representations, and the second stage trains OCT-to-OCTA translation with multi-view alignment.

**FlowOCT.** FlowOCT (Xiong et al., 2026) is a wavelet-domain flow-matching framework for paired three-dimensional OCT-to-OCTA translation. It applies a three-dimensional discrete wavelet transform to paired volumes and learns deterministic transport from OCT coefficients to OCTA coefficients, using the observed OCT representation instead of random initialization. Its Noise-Free Wavelet Flow Matching models transport in the wavelet domain, while RDCA enforces retinal-depth alignment. The Spatial Retinal Layer Projection Constraint combines voxel-wise and layer-projection supervision, and the Structure-Focused Objective emphasizes vascular regions to improve fine vessel reconstruction. We train FlowOCT from scratch without pretrained weights.

**Comparison protocol.** All methods are trained for $400$ epochs without early stopping. After each epoch, the checkpoint with the best validation MAE is retained for test generation. The random seed for each run is generated from the system time. Test volumes and test labels are not used for training or model selection. The test set does not participate in checkpoint, model configuration, or hyperparameter selection and is used only for final evaluation. All methods use the same disjoint data partitions and unified preprocessing pipeline. Data normalization, spatial correspondence, volume reconstruction, and all evaluation metrics are implemented through our common code framework and kept identical across methods. No data augmentation is used; details of the shared data preprocessing are provided in Appendix C.

## F DuoPrior-CFFlow Architecture and Parameters

**Input representation and tensor dimensions.** The implementation receives a normalized OCT volume as a tensor of shape $[B, 1, D, H, W]$. For the reported experiments, the spatial input is $256^3$ and the training batch size is one. The three-dimensional discrete wavelet transform is applied to the spatial dimensions and reduces the coefficient resolution by a factor of two along each axis. The low-frequency LLL coefficient therefore has spatial size $128^3$, while the seven high-frequency subbands have the same spatial size and are concatenated along the channel dimension. The paired OCTA target is transformed with the same operation, so the source and target coefficients remain spatially aligned.

**Flow-matching data path.** During training, the current coefficient state and its paired OCTA target are represented at the same reduced spatial resolution. The interpolation time is sampled from

the unit interval, and the training target is the constant displacement between the source and target coefficient representations. Coarse Flow receives the low-frequency state and predicts a velocity tensor with the same spatial dimensions as the low-frequency coefficients. Fine Flow receives the high-frequency state together with a projected coarse representation, concatenated along the feature-channel dimension, and predicts a velocity tensor for the concatenated high-frequency subbands. The coarse context is detached before it is provided to Fine Flow, matching the training implementation.

**Network implementation.** Both branches use MONAI's `DiffusionModelUNet` with `spatial_dims` $= 3$. The Coarse Flow constructor uses `in_channels` $= 1$ and `out_channels` $= 1$ for the LLL coefficient. The Fine Flow constructor uses `in_channels` $= 8$ and `out_channels` $= 7$: seven high-frequency subbands plus one projected coarse-context channel are provided as input, and seven high-frequency velocity channels are predicted. For both branches, the feature widths are `channels` $= (24, 48, 96)$ and the number of residual blocks at the three levels is `num_res_blocks` $= (1, 1, 1)$. The normalization uses `norm_num_groups` $= 8$, and the attention head width is set to `num_head_channels` $= 8$. The two branches are instantiated independently with these settings and preserve the reduced wavelet resolution until the inverse three-dimensional discrete wavelet transform reconstructs the translated OCTA volume.

**MSZ prior implementation.** For a tensor $x$, min–max normalization is computed as $(x - \min x)/(\max x - \min x + 10^{-6})$. The global cue is obtained by averaging OCT intensity over the two in-plane axes, normalizing along depth, and broadcasting the result over the in-plane dimensions. The local cue uses $9 \times 9 \times 9$ average pooling (stride one, padding four), followed by min–max normalization over the full volume. The depth-gradient cue is computed as the absolute forward difference along depth; the first depth plane is set to zero, and the resulting cue is normalized over the full volume. The three cues are combined as in Eq. (11), where $\mathcal{N}(\cdot)$ denotes normalization.

**Optimization and random seeds.** We optimize DuoPrior-CFFlow using Adam with $(\beta_1, \beta_2) = (0.5, 0.999)$ and zero weight decay. Random seeds are fixed per run and saved with the corresponding experiment configurations.

## G FAZ Region Vascular Exclusion

**Official 3D FAZ annotation and alignment.** We use the official three-dimensional FAZ annotation provided with OCTA-500. For each test case, the label is read from the `FAZlabel` field of the corresponding MAT file in the `GT_FAZ3D` label set. The original annotation has shape $400 \times 400 \times 640$, matching the original OCTA-6M volume coordinates. To align it with the processed OCTA-6M volume, we resample the FAZ mask to $400 \times 400 \times 256$ using nearest-neighbor interpolation and then apply the same in-plane center crop, $72{:}328$ along both axes, yielding a $256^3$ binary mask. Generated and real OCTA volumes are evaluated directly in these processed coordinates; no generated volume is upsampled to the original resolution. The same fixed transformation is applied for every method, and no generated image is used to define or move the FAZ region.

**FAZ interior and vessel maps.** Let $F$ denote the aligned binary three-dimensional FAZ mask. To reduce sensitivity to the annotation boundary, we construct the interior region $F_{\text{int}}$ by applying three-dimensional binary erosion to $F$ with two iterations. The outside region is defined as $O = \neg F$. For each translated OCTA and real OCTA volume, we apply the OCTA-500-adapted SAM-OCTA (Wang et al., 2024) vessel labeler described in Appendix A slice by slice and stack its 2D predictions into a 3D vessel mask. We then use the same thresholding pipeline for both to obtain binary vessel maps. Therefore, the comparison uses identical vessel binarization for generated and real OCTA and does not introduce method-specific postprocessing.

**Metric computation.** The definitions and equations for FAZ occupancy, internal FAZ occupancy, excess leakage, and the FAZ-to-outside vessel-density ratio are provided in Appendix D. In brief, these metrics are computed directly on the aligned three-dimensional binary vessel maps and the fixed official FAZ coordinates. Lower occupancy, internal occupancy, excess leakage, and FAZ-to-outside ratio indicate better exclusion of vascular responses from the avascular region. Real OCTA is used as a reference for contextual comparison.

## H DISEASE CLASSIFICATION

**Task definition and category selection.** OCTA-500 provides disease labels from multiple diagnostic categories. Because several categories contain relatively few cases, we select the three categories with sufficient representation for a stable three-class experiment: NORMAL, AMD, and DR. Each selected case retains its original category label, and the same labels are used for real and translated OCTA test volumes. According to clinicians at our collaborating hospital, two-dimensional en-face OCTA projections are commonly used views for ophthalmic disease assessment. We therefore use depth-projected OCTA as a clinically relevant view for this downstream analysis.

**Cohort size and class distribution.** The classifier-development cohort contains 103 real OCTA-6M training cases: 58 NORMAL, 22 AMD, and 23 DR. The held-out test cohort contains 53 cases: 24 NORMAL, 18 AMD, and 11 DR. The official validation partition is not used for this downstream task. We report macro-averaged metrics alongside accuracy because the test cohort is moderately imbalanced and contains relatively few DR cases.

**Classifier training and data separation.** The classifier is trained only on real OCTA volumes from the original training partition. We use stratified five-fold cross-validation within this partition so that the class proportions are preserved across folds. For each fold, four folds are used for classifier optimization and the remaining fold is used for validation and checkpoint selection. The held-out test partition is never used for classifier training, normalization fitting, or model selection. Generated OCTA volumes are not added to the classifier training set; they are evaluated only after the classifier has been trained on real OCTA. The five folds are used only to select the training duration: we take the median of the five fold-specific best epochs and then train a single final classifier on all 103 training cases. All metrics in Table 4 are computed once from this single frozen classifier on the 53 held-out test cases; no fold-wise probability ensemble or averaging of fold-level metrics is used.

**En-face representation.** Each preprocessed three-dimensional OCTA volume is converted to a two-dimensional en-face image by averaging along the depth axis and transposing the resulting in-plane matrix according to the project coordinate convention:

$$I_{\text{enface}} = \text{transpose}\left(\text{mean}_Z(V)\right).$$

The same projection, spatial sizing, intensity normalization, and tensor conversion are applied to real OCTA and to every translated OCTA volume. Thus, the classifier receives inputs generated by an identical representation pipeline for all methods.

**Classifier and optimization.** We use a two-dimensional ResNet-18 classifier with a three-class output layer. The model is optimized with class-weighted cross-entropy to reduce the effect of class imbalance. For the final classifier, the class weights are computed from all 103 training cases, and training lasts for the median epoch selected by the five-fold validation procedure.

**Test-time evaluation.** The single frozen final classifier is applied once to the same held-out OCTA test cases for real OCTA and every translation method. The test labels and cases are fixed across methods, and no method-specific classifier is retrained on generated data. We report accuracy, balanced accuracy, macro-F1, and macro-AUC.

## I COMPUTATIONAL EFFICIENCY STATISTICS

This section reports the computational cost of the compared translation methods on the OCTA-500 dataset. We summarize the number of trainable parameters in the generation network, peak GPU memory during training, single-volume inference time, and average training time per epoch.

**Measurement protocol.** Parameter counts refer to the generation network used for image synthesis and exclude the auxiliary evaluation classifier. Peak memory is the maximum allocated GPU memory observed during training. Inference time is measured for one complete test volume, including the forward pass and the method-specific sampling procedure. Epoch time is averaged over the corresponding training configuration under the same hardware and data-loading protocol used for the reported experiments.

As shown in Table 7, DuoPrior-CFFlow has the fewest generation-network parameters (7.81 M), compared with 217.60 M for TransPro. Its peak training memory is 22.09 GiB, and its 0.95 s per-

| Method | Generator Parameters | Peak GPU Memory | Inference Time / Volume | Training Time / Epoch |
|---|---|---|---|---|
| Pix2Pix 2D (Isola et al., 2017) | 122.40 M | 28.94 GiB | 0.29 s | 311 s |
| Pix2Pix 3D (Isola et al., 2017) | 116.93 M | 17.90 GiB | 0.08 s | 36 s |
| TransPro (Li et al., 2024b) | 217.60 M | 23.44 GiB | 0.09 s | 188 s |
| BBDM 2D (Li et al., 2023) | 115.91 M | 31.00 GiB | 26.50 s | 7158 s |
| BBDM 3D (Li et al., 2023) | 17.46 M | 20.49 GiB | 14.62 s | 65 s |
| MuTri (Chen et al., 2025) | 191.09 M | 25.99 GiB | 0.12 s | 167 s |
| FlowOCT (Xiong et al., 2026) | 10.84 M | 30.27 GiB | 0.99 s | 146 s |
| **DuoPrior-CFFlow** | **7.81 M** | **22.09 GiB** | 0.95 s | 187 s |

Table 7: Computational efficiency comparison on the OCTA-500 dataset. Parameter counts refer to the generation network. Peak GPU memory, inference time, and epoch time are measured under the protocol described in this appendix.

volume inference time is close to FlowOCT (0.99 s) while substantially shorter than those of BBDM 2D and BBDM 3D (26.50 s and 14.62 s, respectively).

**Interpreting training time.** Training time per epoch denotes the elapsed optimization time for one complete pass through a method's training unit; it is not an inference-time measurement. The unusually large BBDM 2D value arises because the slice-wise implementation processes approximately 5,760 independent 2D slices per epoch, yielding 5,760 optimizer updates. In contrast, the volumetric BBDM 3D setting processes 180 three-dimensional patches per epoch. Thus, epoch time is useful for reporting practical training cost but is not a direct per-sample speed ranking; the per-volume inference column is the appropriate comparison for deployment-time cost. These results indicate that the proposed dual-prior design improves structural guidance without requiring a larger generation network.

## J DuoPrior-CFFlow Inference-Step Analysis

We evaluate the effect of Euler integration steps using the fixed OCTA-6M DuoPrior-CFFlow checkpoint. All settings except the number of inference steps are held fixed. Table 8 reports volumetric reconstruction, perceptual, distributional, and runtime results.

| Euler Steps | MAE ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ | Time / Volume ↓ |
|---|---|---|---|---|---|---|
| **1** | **1.01** | **30.41** | **88.67** | 10.43 | 97.44 | **0.329 s** |
| 2 | 1.08 | 29.78 | 88.11 | 6.39 | 29.71 | 0.873 s |
| 3 | 1.13 | 29.38 | 87.65 | 5.12 | 17.48 | 0.95 s |
| 4 | 1.17 | 29.10 | 87.37 | 4.97 | 13.52 | 1.493 s |
| 5 | 1.20 | 28.92 | 87.19 | 4.81 | 11.69 | 1.566 s |
| 10 | 1.26 | 28.51 | 86.72 | 4.62 | 8.86 | 3.118 s |
| 20 | 1.30 | 28.26 | 86.40 | **4.59** | **7.80** | 6.222 s |

Table 8: Effect of Euler inference steps for DuoPrior-CFFlow on the OCTA-6M dataset. MAE, SSIM, and LPIPS are reported as percentages. The best result in each column is highlighted in **bold**.

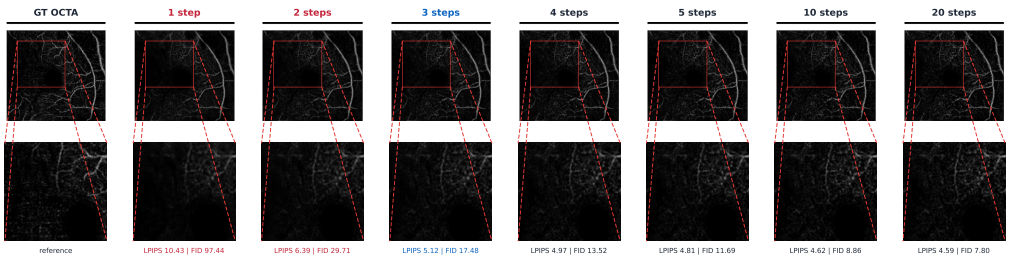



Figure 5: Qualitative comparison of Euler inference steps on a vessel-rich OCTA-6M slice. The lower row enlarges the corresponding red ROI in the upper row.

Figure 5 provides a qualitative complement to Table 8. The one- and two-step outputs exhibit weaker local vascular contrast and incomplete fine branches in the enlarged ROI, whereas three steps recover

the vessel pattern more consistently. One Euler step achieves the best MAE, PSNR, and SSIM, but has substantially worse LPIPS and FID, indicating that this very short trajectory does not adequately recover perceptual vessel details or the target distribution. Increasing the step count progressively improves LPIPS and FID, but also monotonically reduces reconstruction metrics and increases inference time. This behavior suggests that, after a small number of integration steps, additional Euler updates accumulate approximation error in the learned transport field rather than further improving voxel-wise reconstruction. Following FlowOCT (Xiong et al., 2026), we pre-specify **three Euler steps** for all main experiments. A post-hoc analysis with the fixed checkpoint confirms that three steps provide a favorable trade-off between paired reconstruction, perceptual fidelity, distributional similarity, and inference cost.