POSTERIOR-MEAN RECTIFIED FLOW TO ESTIMATE 3D STRUCTURES FROM 2D MEDICAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

This study addresses the highly ill-posed challenge of 3D volume reconstruction from 2D medical images even in scenarios lacking a known physical forward model. We propose a two-stage framework that first employs a specialized "lifting" network to generate a structurally consistent Minimum Mean Squared Error (MMSE) estimate. This smooth baseline then serves as the foundation for a 2D paired flow-matching process, designed to recover high-frequency details across large 3D volumes efficiently. By decoupling structural translation from perceptual refinement, we leverage the generative power of flow matching in 2D while retaining structural consistency through spherical linear interpolation (slerp) without the need for explicit data-fidelity terms. Our approach operates directly in image space, bypassing the ambiguity of latent space correspondences and enabling test-time control over the distortion-perception tradeoff via adjustable refinement steps. When evaluated on the translation of synthetic X-ray to Computed Tomography (CT) and real Color Fundus Photography (CFP) to Optical Coherence Tomography (OCT), our method demonstrates superior structural fidelity and a significantly lower FID compared to state-of-the-art baselines. Code and example GIFs are available here [1].

1 INTRODUCTION

The task of synthesizing 3D volumes from 2D data is a fundamentally ill-posed inverse problem. Because a single 2D projection lacks the depth information inherent in a 3D volume, the solution space is mathematically vast and underdetermined. However, in ophthalmology, for example, 3D information such as Optical Coherence Tomography (OCT) allows the clinician to assess layer thickness and detect potential retinal detachment (Wojtkowski et al., 2005). These insights are indispensable for the early diagnosis and longitudinal monitoring of vision-threatening pathologies (Lim et al., 2012). In related medical imaging domains, such as reconstructing CT volumes from limited X-ray radiographs (De Wilde et al., 2025), researchers typically exploit data-fidelity terms grounded in known physical forward operators (e.g., the Radon transform) (Chung et al., 2023). However, even in such analytical cases, real-world measurements are noisy and therefore reconstructions require strong regularization to perform well (Knoll et al., 2019).

To circumvent the absence of a physical forward model, prior research has shifted toward direct prediction methods using deep encoder-decoder architectures (Liu et al., 2025b). While these models can learn the mapping statistically, the reliance on classical pixel-based objective functions inevitably results in oversmoothed reconstructions, where fine-grained pathological details are lost (Haase et al., 2025). To restore these details, distributional losses, including adversarial ones, are frequently employed (Isola et al., 2017; Pinetz et al., 2023). While such methods improve visual sharpness, they are notoriously prone to hallucinations (Cohen et al., 2018). The recent emergence of diffusion-based generative models (Song et al., 2021; Karras et al., 2022) offers a promising alternative for high-fidelity reconstruction. However, standard image-space diffusion is computationally prohibitive for high-resolution 3D volumes. Latent Diffusion Models (LDMs) (Rombach et al., 2022; Guo et al., 2025; Liu et al., 2025a; Shi et al., 2025) address this by operating in a compressed latent space. Still, this introduces a new challenge in maintaining data consistency. In latent space, the direct relationship between the original 2D image and the generated volume is often obscured.

[1] https://anonymous.4open.science/r/pmrf3dsynthesis-CEC5

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.

This was recently highlighted in a major benchmark challenge on Fundus-to-OCT translation (Liu et al., 2025b), where LDM-based approaches achieved the highest perceptual but some of the poorest data-fidelity scores. Recently, flow matching models (Lipman et al., 2023; Erbach et al., 2025) have emerged as a powerful alternative, demonstrating superior synthesis capabilities in both natural and medical imaging (Ohayon et al., 2025; Brandstötter & Kobler, 2025; Bogensperger et al., 2025). One advantage of flow models is that their transition can be applied between arbitrary distributions, including empirical ones, instead of starting from a known analytic (typically Gaussian) distribution. We propose to disentangle the reconstruction into two distinct stages, we overcome the computational constraints of 3D modeling while maximizing both structural accuracy and perceptual quality:

1. **MMSE Estimation (Structural Lifting)**: We design a novel neural network that "lifts" the 2D image measurement into a 3D coordinate space. This stage generates a Minimum Mean Squared Error (MMSE) estimate of the target volume, providing a geometrically consistent baseline.
2. **Flow Refinement**: We then apply a flow matching model to transform this smooth MMSE estimate into a sharp, realistic volume on a slice-by-slice basis. By decoupling the "lifting" from the "refinement", we can utilize larger, more expressive networks to capture the high-frequency textures.

Our approach operates directly in image space, bypassing the need for lossy autoencoders and allowing for the adjustment of the distortion-perception tradeoff at test-time without retraining. Hence our contributions are:

- We present the first slice-wise procedural perceptual refinement for the synthesis of 3D structures from 2D images using flow-matching, enabling realistic reconstructions that maintain structural integrity, including a grounding in mathematical theory.
- We introduce a specialized network architecture and augmentation strategy tailored for specific 2D-to-3D synthesis tasks, ensuring scalability to high-resolution data.
- We establish a new state-of-the-art in two benchmarks involving two different modalities and anatomies, demonstrating significant improvements in both structural fidelity and perceptual metrics.

2 Related Work

Our work lies at the intersection of single-view 3D reconstruction, generative medical image-to-image translation, and the navigation of the distortion-perception tradeoff.

Single-View 3D Reconstruction. Estimating 3D structures from 2D natural images is a foundational computer vision problem, historically addressed via monocular depth estimation (Zhou et al., 2021). While monocular approaches represent the depth coordinate as continuous intensity projections, newer single-view reconstruction paradigms leverage 3D morphable models, diffusion priors (Wu et al., 2024), or 3D Gaussian Splatting (Szymanowicz et al., 2024) to synthesize realistic assets. However, these natural scene methods are fundamentally constrained to estimating the *outer surface boundary* (i.e., 2.5D or boundary meshes). In contrast, medical modalities such as Computed Tomography (CT) and Optical Coherence Tomography (OCT) require the reconstruction of highly complex, heterogeneous, and dense *internal volumetric anatomy*, rendering surface-based approaches inapplicable.

Medical 2D-to-3D Lifting. In the medical domain, reconstructing 3D volumes from limited projections is a well-studied inverse problem, most commonly seen in sparse-view or limited-angle CT reconstruction (Gong et al., 2019). The extreme boundary of this problem, reconstructing dense 3D volumes from a single 2D projection, remains a major challenge. To recover the fine-grained pathological details lost in deterministic mean-squared error (MSE) objectives, prior works have employed Generative Adversarial Networks (GANs) (Ying et al., 2019) or Latent Diffusion Models (LDMs) (Bai et al., 2024; Liu et al., 2024; 2025a). However, adversarial losses are highly susceptible to introducing clinically dangerous hallucinations (Cohen et al., 2018). Meanwhile, LDMs compress volumes into low-dimensional latent spaces to fit within GPU hardware memory constraints, which

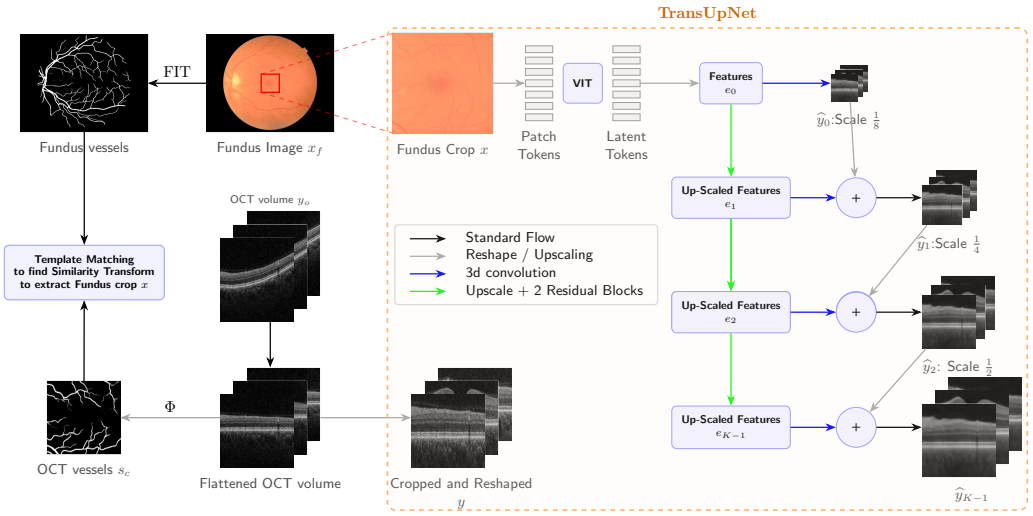

Figure 1: Schematic overview of our image alignment procedure and network design (TransUpNet) to produce a Minimum Mean Squared Error (MMSE) estimate $\widehat{y}_{K-1}$.

breaks explicit spatial correlation and compromises downstream data fidelity (Liu et al., 2025b). Specialized projection modules have been invented to eliviate this issue Liu et al. (2024). Others have synthesized additional views to provide more guidance Xie et al. (2026). Our framework addresses this bottleneck by decoupling deterministic structural "lifting" from image-space perceptual refinement, avoiding lossy latent-space compressions altogether.

The Distortion-Perception Tradeoff in Generative Models. A fundamental mathematical limit governs image reconstruction: the Pareto frontier between pixel-wise distortion (e.g., PSNR, SSIM) and perceptual quality (e.g., FID) (Blau & Michaeli, 2018). Recently, posterior-mean rectified flow (PMRF) models (Ohayon et al., 2025) were introduced to elegantly traverse this tradeoff by modifying the ODE integration trajectory at inference time. While training-free diffusion-based variants have been proposed to navigate this Pareto frontier (Zhang et al., 2026; Wang et al., 2026), they require full 3D model executions that are computationally prohibitive for high-resolution volumetric medical scans. To make this traversal computationally tractable, we adapt flow-matching refinement as a slice-wise 2D operator, grounding the generative sampling path in a structurally consistent 3D MMSE spatial anchor as well as a specialized noise structure.

3 Method

Our proposed framework decomposes the translation of a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$, where $x^i \in \mathbb{R}^{h_x \times w_x}$ are 2D images into $y^i \in \mathbb{R}^{h_y \times w_y \times d_y}$ 3D volumes, into two distinct phases. First, we use a novel "lifting" architecture, TransUpNet, to generate a structurally consistent Minimum Mean Squared Error (MMSE) estimate (Fig. 1). Second, we apply a flow-matching refinement process in 2D to recover high-frequency textures and anatomical details (Fig. 2). We highlight the methodology based on the application of OCT from CFP generation, however, to ensure reproducibility in the absence of available datasets, we also evaluate on X-ray-to-CT.

TransUpNet for 3D MMSE Estimation The TransUpNet architecture is designed to "lift" 2D features into 3D space. It begins with a patch extraction layer (2D convolution, $k = 8, s = 8$ and $c \cdot k^2$ channels) that feeds into a pretrained ViT-B/16 encoder. We remove the [CLS] token and interpolate the positional encodings to match the input resolution. The resulting latent representation is reshaped into a 4D feature block $e \in \mathbb{R}^{\widehat{c} \times \frac{h_y}{8} \times \frac{w_y}{8} \times \frac{d_y}{8}}$. The decoder reconstructs the volume hierarchically. An initial 3D convolution produces a low-resolution prediction $\widehat{y}_0$. Subsequent decoder blocks b perform upsampling and residual processing, generating refined features $e_{k+1} = b(e_k)$.

A 3D convolutional head h then predicts a residual update, which is added to the upscaled prior: $\widehat{y}_{k+1} = \text{up}(\widehat{y}_k) + h(e_{k+1})$. We optimize the network using an ℓ^2 loss at every scale k, effectively training the model as an MMSE estimator:

$$\omega = \arg\min_{\omega} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \|\text{down}(y^i, 2^{K-k-1}) - \widehat{y}_k\|_2^2. \tag{1}$$

This multi-scale supervision ensures that structural fidelity is maintained from the coarse global geometry to the finer layers. The final prediction is $\widehat{y}_{K-1}$.

In particular, the 3D MMSE additionally acts as a slice-wise 2D MMSE using a slice operator $\mathcal{P}$, which is ensured by the following Proposition:

Proposition 1. *Let $\mathcal{P} : \mathbb{R}^{h\times w\times d} \to \mathbb{R}^{h\times w}$ denote the linear slice-wise projection operator. The projection of the 3D Minimum Mean Square Error (MMSE) estimate, denoted by $\widehat{y}_{K-1}$, is equivalent to the 2D MMSE estimate of the corresponding slice-wise distribution under the assumption that the neural network converges to the true conditional expectation.*

Proof. By definition, the 3D MMSE estimate is given by the conditional expectation $\widehat{y}_{K-1} \coloneqq \mathbb{E}[y \mid x]$. Given that the conditional expectation is linear and the projection is a matrix. It therefore follows that:

$$\mathcal{P}(\widehat{y}_{K-1}) = \mathcal{P}(\mathbb{E}[y \mid x]) = \mathbb{E}[\mathcal{P}y \mid x].$$

Recognizing that $\mathbb{E}[\mathcal{P}y \mid x]$ constitutes the 2D MMSE estimate of the projected slice-wise distribution, denoted as y_{2DMMSE}, we conclude that $\mathcal{P}(\widehat{y}_{K-1}) = y_{\text{2DMMSE}}$. □

We emphasize that this result naturally generalizes to any linear projection operator, such as the slice selector adopted in this work, which is the natural choice in OCT.

Flow Matching for Perceptual Refinement While TransUpNet provides structural consistency, the MMSE loss objective inherently produces the conditional mean of the target distribution, which manifests as oversmoothed reconstructions. To recover the high-frequency textural details of the 3D distribution, we employ posterior-mean rectified flow (Ohayon et al., 2025). We define the intermediate variable $z_j = \mathcal{P}_j\widehat{y}_{K-1} + n$, where $n \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$ represents the starting noise. This $z_j \in \mathbb{R}^{h\times w}$ serves as the starting point for our flow. We then construct a probability path between this noisy MMSE estimate z_j and the ground truth slice y_j. For any time $t \in [0, 1]$, the intermediate state is defined by the linear interpolation $u_t = ty_j + (1-t)z_j$. We train a 2D U-Net (Karras et al., 2022) $g_\theta : \mathbb{R}^{h_x\times w_x\times[0,1]} \to \mathbb{R}^{h_x\times w_x}$ with time conditioning to estimate the velocity field $v(y_j, u_t, t) \coloneqq \frac{y_j - u_t}{max(1-t, t_{\min})}$ that drives the transformation from the structural baseline toward the data manifold. Similarly, we define the predicted velocity $\widehat{v}_\theta(u_t, t) = \frac{g_\theta(u_t,t) - u_t}{\max(1-t, t_{\min})}$. We optimize g_θ using the velocity loss proposed by Li & He (2026):

$$\mathcal{L}(u_t, y_j, t) = \|v(y_j, u_t, t) - \widehat{v}_\theta(u_t, t)\|_2^2. \tag{2}$$

This formulation allows the model to leverage the generative capabilities of flow matching to "fill in" the missing anatomical textures while being strictly conditioned on the 3D geometry established by the first stage.

Inference Process During inference, the TransUpNet lifts the 2D input x into 3D space. To retain a consistent estimate along the trajectory and reduce flickering artifacts, we also adopt a spherical linear interpolation (slerp) to correlate the 3D noise structure, as was recently proposed for diffusion models by Du et al. (2026). We sample two noise vectors for the first and last slices, respectively $\epsilon_0, \epsilon_1 \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$. Then, for each slice $j \in [0, d-1]$, we interpolate along the great-circle arc spanned by these vectors:

$$\mathbf{n}_j = \frac{\sin((1-\alpha_j)\Omega)}{\sin(\Omega)}\epsilon_0 + \frac{\sin(\alpha_j\Omega)}{\sin(\Omega)}\epsilon_1 \tag{3}$$

using $\alpha = [0, \frac{1}{d-1}, \cdots, 1]$ and $\Omega = \arccos(\langle\epsilon_0, \epsilon_1\rangle / (\|\epsilon_0\| \cdot \|\epsilon_1\|))$ is the angle between the noise vectors. This ensures that the noise is correlated across the slice dimension and furthermore, that

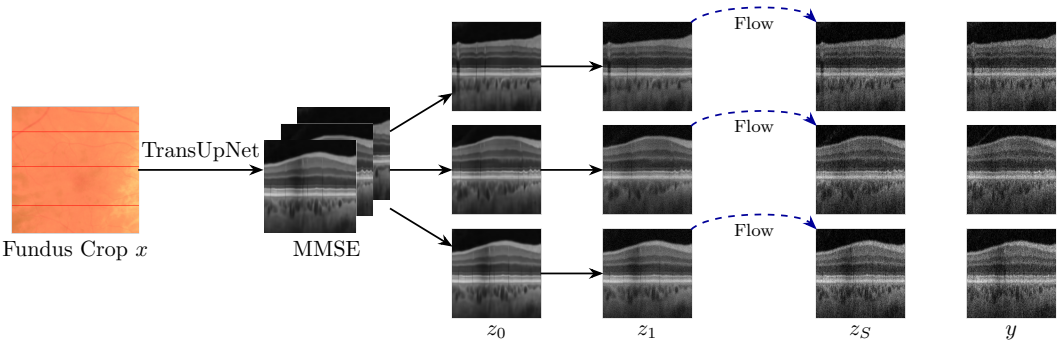

Figure 2: Schematic overview of the proposed flow matching. The red lines in the original fundus crop correspond to the actual three B-scans shown in y. The flow retains the geometry, even though it is not aware of any neighboring slices.

each component follows a Gaussian distribution. To initialize the refinement, the noise is added according to the pre-defined strength: $z_0 = \mathcal{P}_j f_\omega(x) + \mathbf{n}_j$. We then traverse the learned trajectory using a simple Euler solver over S discrete steps:

$$z_{s+1} = z_s + \frac{v_\theta(z_s, \frac{s}{S})}{S}. \tag{4}$$

The final output z_S represents the reconstructed volume. Moreover, theoretical guarantees from posterior-mean rectified flow naturally transfer to our volumetric setting. Specifically, Proposition 1 of Ohayon et al. (2025) establishes that the solution of the underlying ODE of Eq. 4 achieves optimal perceptual quality while maintaining a non-decreasing MSE distortion relative to the baseline MMSE estimate as the noise term $\mathbf{n} \to 0$. Because mean squared error is additive across slices, this monotonicity extends directly to the entire 3D volume. While the perceptual quality also holds slice-wise, this is not necessarily the case for the volume, as the volume wise distribution has inter-slice dependencies, which are modeled by our choice of noise term. Moreover, in practical applications, as observed by Ohayon et al. (2025), a small non-zero noise level ($\mathbf{n} \neq 0$) is required for numerical stability during trajectory integration, yielding a minor tradeoff in pixel-wise data fidelity.

4 Applications to Medical Imaging

Image Alignment and Preprocessing in Fundus to OCT To minimize the ambiguity of the 2D-to-3D mapping, we align the input fundus image $x_f \in \mathbb{R}^{h_f \times w_f \times 3}$ and the target volume $y_o \in \mathbb{R}^{h_o \times d_o \times w_o}$. We first flatten the OCT volume by segmenting the Bruch's membrane using SD-RetinaNet[2] (Fazekas et al., 2025). This process removes height-related axial ambiguity not present in the 2D fundus image. The flattened volumes are then cropped in height and resized in width for each bscan to $y \in \mathbb{R}^{h_o \times d \times d}$ as is visualized in Fig.1.

To align the fundus image x_f with the OCT y, we utilize retinal vessel maps as stable landmarks. We extract the fundus vessel map using the Fundus Image Toolbox[3] (Gervelmeyer et al., 2025) (FIT), while the OCT vessel maps are generated by a U-Net with residual blocks (precise network architecture is shown in the appendix A.4) Φ trained to produce an en-face probability map from the 3D volume. Using template matching with normalized cross-correlation, we estimate a similarity transform (constrained to rotations $\pm 10°$ and scaling 90%–110%). The final fundus crop is resized to $x \in \mathbb{R}^{h_o \times d \times 3}$, resulting in the paired dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$, which again is shown in Figure 1.

Image Alignment and Preprocessing in X-ray to CT This application closely follows the work of Bai et al. (2024). Also the DiffDrr framework[4] is used to construct synthetic X-rays from the CT volumes, which are subsequently upsampled using our algorithm. For simplicity, we opted to stick

[2] https://github.com/ABotond/SD-LayerNet
[3] https://github.com/berenslab/fundus_image_toolbox
[4] https://github.com/eigenvivek/DiffDRR

with the synthetic X-rays as this serves as a synthetic benchmark environment to pit various methods against each other and neglects all kinds of auxiliary sources of error, introduced by modalities acquired using different scanners, such as registration and alignment of intensity values. We stress that, the method is easily extendable to multi-view CT reconstruction by adding the additional views as additional channels to the input.

5 Experimental results

Experimental setup: Our study utilizes a multi-center clinical dataset of 365 patients exhibiting early-stage AMD (Lim et al., 2012) in at least one eye, recruited from 6 different hospitals. From this cohort, we collected 527 eyes with paired color fundus photography and OCT volumes acquired with a Maestro2 (Topcon Healthcare, Inc.) device. To ensure consistent data quality, we excluded eyes with poor imaging quality or registration artifacts. This resulted in a final dataset of 403 eyes, which was partitioned into an 80/20 patient-wise split for training and test. We rely on a pre-registered paired modality to test our framework, however we have added an inference script, which locates the fovea and crops a region around it to produce an example OCT without knowledge of the paired scan. As a second modality, we also evaluate on a publicly available X-ray to CT dataset (Armato III et al., 2011). To simplify this task and to keep inline with existing literature, we use synthetic X-ray generated by the DiffDRR library. Training and validation splits are taken from Bai et al. (2024). Other implementation details for both OCT and CT are listed in the Appendix A.1. To showcase the performance in relation to established architectures, we also trained a model on the publicly available and fully processed dataset used in X2CT-GAN (Ying et al., 2019) in the supplement.

We evaluated performance using PSNR, 3D Structural Similarity Index (3D-SSIM) in percent, and Fréchet Inception Distance (FID). As advocated by Wu et al. (2026), we also calculated the Dice Score (Dice) based on retinal layer segmentations provided by SD-LayerNet (Fazekas et al., 2025) and retinal vessel segmentation (Lachinov et al., 2021). The FID was computed across the entire dataset using RETFound embeddings (Zhou et al., 2023) for OCT and a pre-trained ResNet50 on RadImageNet (Mei et al., 2022) for CT as feature extractors. Statistical significance was determined using the Wilcoxon signed-rank test with a threshold of $p < 0.05$ to closest competitor.

Baselines: We benchmarked our framework against several established 3D generation and reconstruction strategies, while implementation details are supplied in the Appendix:

- **Encoder-Decoder (ED):** We use the implementation from stable diffusion autoencoder[5] (Rombach et al., 2022), where we replaced 2D convolutions with 3D variants and removed self-attention to accommodate GPU memory constraints. We use a learned projection kernel and a permutation strategy to lift into 3D space.
- **Adversarial Methods (AM):** We implemented Pix2Pix (Isola et al., 2017) and a Wasserstein Loss-based model without pixel loss (Pinetz et al., 2023), both utilizing our TransUpNet as the generator. The adversarial loss weights were optimized via grid search.
- **Latent Diffusion (LDM):** We adapted the Stable Diffusion framework (Rombach et al., 2022) for 3D synthesis with a $\times 8$ compression ratio. The VAE implementation replaced 2D convolutions with 3D variants; self-attention layers were removed to accommodate GPU memory constraints. We also tested a video VAE variant as recently proposed in Chen et al. (2026).
- **Paired Flow Matching (PFM):** Based on Li & He (2026), we learn a flow model between the paired learned lifting and the target OCT directly in 3D using an encoder-decoder network for flow estimation.

Experimental Results in Fundus to OCT In Table 1, we compare our framework against current baseline architectures. When compared to traditional ED models, the TransUpNet stage maintains PSNR values, but achieves significantly higher SSIM and Dice scores. This suggests that while the intensity range is similar, our structural alignment is superior. Crucially, our model demonstrates better synthesis of clinically relevant biomarkers, as visualized in Fig. 3. For instance, our method

[5] https://github.com/CompVis/stable-diffusion

Table 1: Results on the in-house early/intermediate AMD dataset. PSNR and SSIM showcase the consistency with the measurement, while FID showcases the plausibility of the resulting synthesis.* denotes statistical significance to the closest competitor marked by underline.

	PSNR↑	SSIM↑	FID↓	Dice Layers↑	Dice Vessels ↑
ED (Liu et al., 2025b)	**19.4** $\pm$ 1.0	25.5 $\pm$ 1.9	448.3	23.0 $\pm$ 3.1	16.8 $\pm$ 9.3
Pix-2-Pix (Isola et al., 2017)	18.6 $\pm$ 1.0	22.3 $\pm$ 2.3	116.9	<u>47.8</u> $\pm$ 5.3	<u>44.4</u> $\pm$ 12.6
Wasserstein (Pinetz et al., 2023)	**19.4** $\pm$ 1.2	<u>26.1</u> $\pm$ 2.3	405.8	24.4 $\pm$ 2.6	28.6 $\pm$ 10.5
LDM (Rombach et al., 2022)	15.8 $\pm$ 0.7	15.5 $\pm$ 1.3	<u>25.7</u>	40.3 $\pm$ 5.5	8.9 $\pm$ 3.1
$\text{LDM}_{\text{WF-VAE}}$ (Chen et al., 2026)	14.1 $\pm$ 1.2	11.5 $\pm$ 1.7	118.4	22.8 $\pm$ 11.1	5.1 $\pm$ 3.8
PFM (Li & He, 2026)	17.9 $\pm$ 0.7	20.2 $\pm$ 1.4	236.0	28.7 $\pm$ 4.1	0.5 $\pm$ 1.2
Ours (no flow)	**19.4** $\pm$ 1.2	**26.6*** $\pm$ 2.7	435.0	32.6 $\pm$ 4.9	45.9 $\pm$ 10.9
Ours (linear flow)	18.0 $\pm$ 0.9	18.8 $\pm$ 2.0	**17.8**	**48.3** $\pm$ 5.5	46.7 $\pm$ 11.9
Ours (slerp Flow)	18.0 $\pm$ 0.9	19.1 $\pm$ 2.0	18.0	**48.3*** $\pm$ 5.5	**48.0*** $\pm$ 11.1

Table 2: Results on the public CT dataset (Armato III et al., 2011). PSNR and SSIM showcase consistency with the measurement, while FID showcases the plausibility of the resulting synthesis.* denotes statistical significance to the closest competitor.

	PSNR↑	SSIM↑	LPIPS↓	FID↓	Runtime in s↓
ED (Liu et al., 2025b)	16.2 $\pm$ 1.2	40.9 $\pm$ 9.7	0.60 $\pm$ 0.06	9.7	1.0
Pix-2-Pix (Isola et al., 2017)	16.5 $\pm$ 1.5	36.5 $\pm$ 11.5	<u>0.23</u> $\pm$ 0.05	11.5	0.3
Wasserstein (Pinetz et al., 2023)	<u>16.9</u> $\pm$ 1.3	**60.6** $\pm$ 8.8	0.31 $\pm$ 0.08	<u>1.9</u>	0.3
LDM (Rombach et al., 2022)	9.5 $\pm$ 1.2	26.6 $\pm$ 9.0	0.43 $\pm$ 0.07	6.0	83.8
$\text{LDM}_{\text{WF-VAE}}$ (Chen et al., 2026)	9.8 $\pm$ 1.2	25.6 $\pm$ 11.8	0.44 $\pm$ 0.07	7.4	83.8
PFM (Li & He, 2026)	10.9 $\pm$ 1.1	17.7 $\pm$ 6.5	0.71 $\pm$ 0.03	33.5	51.1
Ours (no flow)	**18.0*** $\pm$ 1.5	<u>58.3</u> $\pm$ 9.3	0.53 $\pm$ 0.08	1.4	0.3
Ours (linear flow)	17.2 $\pm$ 1.5	56.1 $\pm$ 9.0	0.19 $\pm$ 0.06	0.28	55.0
Ours (slerp Flow)	17.0 $\pm$ 1.4	56.1 $\pm$ 9.1	**0.19*** $\pm$ 0.07	**0.22**	55.0

preserves Drusen bumps, an important biomarker for this cohort (Row 2) and retinal vessel structures (Row 3) with high fidelity. In contrast, while LDM produce perceptually sharp images, they exhibit poor spatial correlation with the ground truth, reflected in their lower PSNR scores. The closest baseline, Pix2Pix (utilizing TransUpNet as a backbone), suffers from common adversarial artifacts. For example, Row 3 of Fig. 3 shows synthesized vessels that do not exist in the ground truth, which in turn result in lower Dice scores for the vessel segmentation compared to our method in Tab.1. Our flow-based refinement avoids these pitfalls by grounding the generative process in the MMSE structural estimate. Thereby, our method is also the only one that actually improves upon the MMSE estimator in structural fidelity on both layer and vessel segmentation by two different unbiased downstream network.

Experimental Results in X-ray to CT The same general trends in principle hold for our X-ray to CT experiments, shown in Table 2 and Figure 3. One metric that behaves differently on the CT benchmark is the FID metric. Due to small shifts in the mean intensities, some models like the adversarial loss and the latent diffusion exhibit larger FID scores, despite generating visually appealing images. Our MMSE estimator better PSNR values than all competing methods. However, this time the SSIM values are slightly lower than using the Wasserstein loss, which follows the original reported results by Pinetz et al. (2023). Refining the prediction yields better fidelity at the cost of a minor drop in distortion metrics. In the absence of downstream segmentation models, we also include an LPIPS comparison, which is significantly better than competing methods.

In general, the usage of our refinement approach significantly improves the perceptual and downstream results at the cost of additional inference time, which for clinical usability is still okay as the clinical examinations themselves take a lot longer.

Ablation Study We conducted a series of ablation studies to evaluate the impact of our augmentation pipeline and our hyperparameters (see Table 3 for OCT and Table 4 for CT). Further results, such as the choice of flow-matching network or augmentation strategies, are shown in the Appendix.

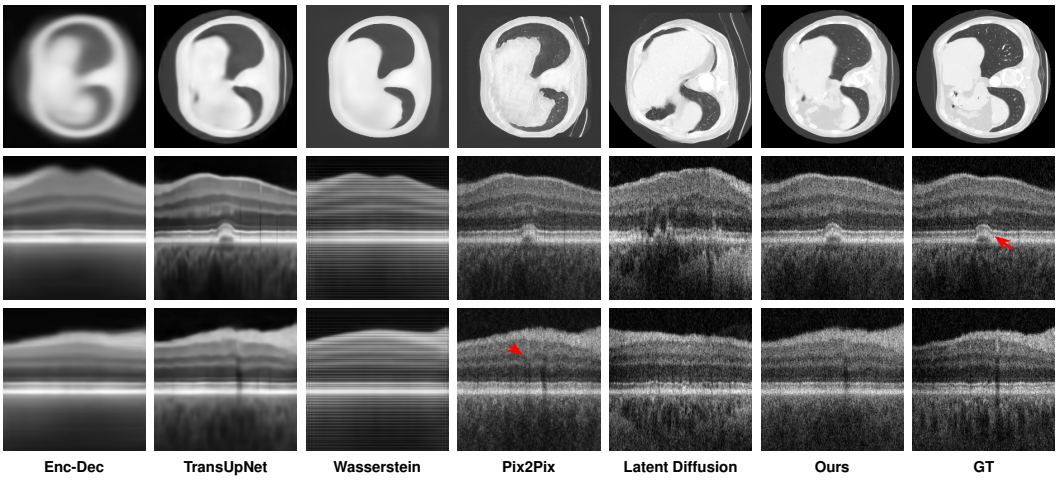

Figure 3: Qualitative comparison to the state-of-the-art. In the first row, we showcase the general behavior on CT. Methods based on our architecture retain the general structure of the CT, and the flow-matching recovers high-frequency information. In the second row, we showcase (red arrows) that an important biomarker (Drusen) is retained through the synthesis with TransUpNet. In the third row, the adversarial loss leads to a hallucinated vessel.

Flow Steps (S): The number of inference steps S directly controls the trade-off between computational cost and perceptual fidelity. As shown in Table 3 and Fig. 4, increasing S from 1 to 20 dramatically improves the FID (from 438.4 to 18.0), moving from oversmoothed MMSE outputs to realistic OCT textures. However, this increase in realism results in a decline in PSNR and SSIM. However, increasing the flow steps incurs a cost in terms of runtime, which compared to an OCT appointment, is still negligible (Bosche et al., 2019). Following refinement the layer dice worsens after more than two steps. A similar trend has been documented in physics by Han et al. (2026).

Noise Level (σ_ϵ): We also investigated the impact of the starting noise level. As is noted in Ohayon et al. (2025), small noise is needed to prevent singularities in mapping from one image to its counterpart, however the PSNR degrades with rising noise level. As shown in the Tables 3, 4, higher noise levels have the same impact in general as increasing the number of flow steps and therefore improves perceptual quality, while decreasing the data-fidelity fit, the impact is rather small even after 20 steps. Hence, also the highest noise level ($\sigma_\epsilon = 0.1$) yields the best perceptual quality (FID) across both tasks.

Noise type: We further ablated the difference of our inference noise in Table 1 and Table 2. In principle, the results are very similar, with linear noise having a slight advantage in raw FID, slerp noise has the advantage in SSIM and downstream metrics. The main differences are shown by looking at the 3D reconstructions, where the slerp noise generates smooth transitions, which is not the case with the linear noise[6].

6 Conclusions

To overcome the severe computational bottlenecks and spatial distortions of 3D generative models, we introduce a principled 2D-to-3D flow-matching framework that achieves state-of-the-art volumetric reconstruction from single 2D projections. Additionally, a novel architecture is proposed improving on the encoder-decoder architectures, commonly used in 3D imaging, with improved reconstruction performance as well as decreased inference time. We envision flow matching as an alternative to adversarial losses, which minimizes their instabilities and tendencies to hallucinate (Cohen et al., 2018). We also demonstrate that this improves crucial downstream clinical tasks, such as vessel and layer segmentation. Because standard fundus scans cannot accurately capture all layers, our more realistic scans make layer separation significantly easier. Furthermore, adjusting

[6] https://anonymous.4open.science/r/pmrf3dsynthesis-CEC5

Table 3: Ablation studies on flow refinement parameters: Noise level (σ_ϵ) and number of refinement steps (S). For σ_ϵ ablation ($S = 20$) and for S ablation $\sigma_\epsilon = 0.1$ is used.

Ablation	**Value**	PSNR↑	SSIM↑	FID↓	Dice Layers↑	Dice Vessels↑
Flow Steps S	1	19.4 ± 1.2	26.6 ± 2.7	438.4	34.7 ± 4.8	47.0 ± 11.0
	2	19.1 ± 1.1	24.9 ± 2.6	243.2	$\mathbf{49.1} \pm 5.2$	47.6 ± 11.5
	5	18.6 ± 1.0	21.9 ± 2.3	83.7	48.0 ± 5.3	47.3 ± 11.8
	10	18.2 ± 0.9	20.0 ± 2.1	29.0	48.1 ± 5.4	47.9 ± 11.2
	20	18.0 ± 0.9	19.1 ± 2.0	**18.0**	$\underline{48.3} \pm 5.5$	48.0 ± 11.1
Noise Level σ_ϵ	0.1	18.0 ± 0.9	19.1 ± 2.0	**18.0**	$\underline{48.3} \pm 5.5$	48.0 ± 11.1
	0.05	18.1 ± 0.9	19.1 ± 2.0	20.8	47.7 ± 5.3	47.6 ± 11.7
	0.01	18.1 ± 0.8	19.5 ± 2.1	26.6	48.2 ± 5.4	47.4 ± 11.5

Table 4: Ablation studies on flow refinement parameters for the CT dataset: Noise level (σ_ϵ) and number of refinement steps (S). For σ_ϵ ablation ($S = 20$) and for S ablation $\sigma_\epsilon = 0.1$ is used.

Ablation	**Value**	PSNR↑	SSIM↑	LPIPS↓	FID↓	Runtime in s
Flow Steps S	1	$\mathbf{17.8} \pm 1.5$	$\mathbf{61.8} \pm 9.3$	0.27 ± 0.07	1.09	3.0
	2	17.6 ± 1.5	61.6 ± 9.4	0.26 ± 0.07	1.08	5.7
	5	17.5 ± 1.5	59.4 ± 8.8	0.23 ± 0.06	0.72	13.8
	10	17.3 ± 1.4	58.4 ± 8.9	0.21 ± 0.06	0.48	27.4
	20	17.0 ± 1.4	56.1 ± 9.1	$\mathbf{0.19} \pm 0.07$	**0.22**	55.0
Noise Level σ_ϵ	0.1	17.0 ± 1.4	56.1 ± 9.1	0.19 ± 0.07	**0.22**	55.0
	0.05	17.1 ± 1.4	56.1 ± 9.2	$\mathbf{0.19} \pm 0.07$	0.24	55.0
	0.01	17.2 ± 1.4	56.1 ± 9.0	0.19 ± 0.06	0.28	55.0

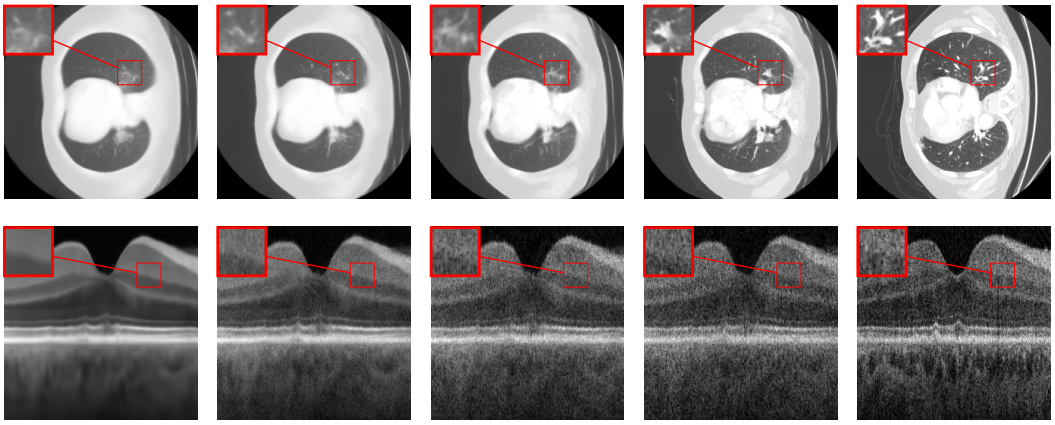

Figure 4: Visual comparison of the number of steps used during stage 2. As the number of steps grows, so does the local noise, increasing the realism in the generated images. There is subtle Drusen in the image, which is not well represented, but each image shows differing behavior at this part of the image, making the standard deviation across reconstructions in this pathological region high.

the number of refinement steps at test-time allows for dynamic control over the tradeoff between data fidelity and perceptual quality, enabling the generation of various outputs across the distortion-perception plane. In the future, we plan to investigate the utility of the inherent uncertainty in the reconstruction process for downstream performance and uncertainty estimation in various downstream applications such as drusen volume detection (Pinetz et al., 2026) as well as exploiting the framework for data augmentation in multi-modal foundation models such as Morano et al. (2025).

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs and designing or providing feedback on research methodology or experiments. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, implementing methods, assisting with translation, cleaning and reformat dataset, supporting qualitative and thematic data analysis, as well as interpreting results and generating synthetic data sets is not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, sourcing/searching for information, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. The identified literature was read and manually added to bibtex. We used LLM's as advanced code completion, and the resulting lines of code were checked for correctness. We also made use of the LLM based review provided by the ICLR conference and made changes to the manuscript in response to this feedback. The paper was proofread by all authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All subjects included in this study did so with ethical consent. We acknowledge that the tasks considered in this study are very challenging and our proposed method has the potential to hallucinate harmful imaging artifacts. Further research is required to ensure no harm is done by applying this methodology in clinical practice.

REPRODUCIBILITY STATEMENT

We uploaded an autonomous code repository with the code and model checkpoints necessary to produce the values in our tables for our method. The dataset for the CT experiments is publicly available[7] and the splits are taken from Bai et al. (2024) and are included in the repository. We did not test against the model as they did not provide any model checkpoints. We did not get access to a comparable dataset for Fundus to OCT, as our requests for the APTOS challenge were denied (Liu et al., 2025b). All necessary implementation details and pre-processing steps are detailed in the Appendix in their corresponding section.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Qingze Bai, Tiange Liu, Zhi Liu, Yubing Tong, Drew Torigian, and Jayaram Udupa. Xctdiff: Reconstruction of ct images with consistent anatomical structures from a single radiographic projection image. *arXiv preprint arXiv:2406.04679*, 2024.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

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

Fabian Bosche, Jil Andresen, Daniel Li, Frank Holz, and Christian Brinkmann. Spectralis oct1 versus oct2: time efficiency and image quality of retinal nerve fiber layer thickness and bruch's membrane opening analysis for glaucoma patients. *Journal of Current Glaucoma Practice*, 13 (1):16, 2019.

[7] https://www.cancerimagingarchive.net/collection/lidc-idri/

Bastian Brandstötter and Erich Kobler. Synthesizing Accurate and Realistic T1-Weighted Contrast-Enhanced MR Images Using Posterior-Mean Rectified Flow. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 159–169. Springer, 2025.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

Qi Chen, Shuhan Ding, Yu Gu, Nan Liu, Jiang Bian, Alan Yuille, Zongwei Zhou, and Jingjing Fu. Foundation vae for ct reconstruction, augmentation, and generation. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=f51haqkwRQ.

Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3D inverse problems using pre-trained 2D diffusion models. In *CVPR*, pp. 22542–22551, 2023.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *MICCAI*, pp. 529–536. Springer, 2018.

Daniel De Wilde, Olivier Zanier, Raffaele Da Mutten, Michael Jin, Luca Regli, Carlo Serra, and Victor E Staartjes. Strategies for generating synthetic computed tomography-like imaging from radiographs: A scoping review. *Medical Image Analysis*, pp. 103454, 2025.

Chenhe Du, Qing Wu, Xuanyu Tian, Jingyi Yu, Hongjiang Wei, and Yuyao Zhang. Improving 2d diffusion models for 3D medical imaging with Inter-Slice consistent stochasticity. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=R5ETdN6ifA.

Julius Erbach, Dominik Narnhofer, Andreas Robert Dombos, Bernt Schiele, Jan Eric Lenssen, and Konrad Schindler. Solving inverse problems with FLAIR. In *NeuRIPS*, 2025. URL https://openreview.net/forum?id=w9xETx7HT1.

Botond Fazekas, Guilherme Aresta, Dmitrii Lachinov, Sophie Riedl, Julia Mai, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. SD-LayerNet: Robust and label-efficient retinal layer segmentation via anatomical priors. *Computer Methods and Programs in Biomedicine*, 261:108586, 2025.

Julius Gervelmeyer, Sarah Müller, Ziwei Huang, and Philipp Berens. Fundus Image Toolbox: A Python package for fundus image processing. *Journal of Open Source Software*, 10(108):7101, April 2025. doi: https://doi.org/10.21105/joss.07101.

Changcheng Gong, Li Zeng, and Chengxiang Wang. Image reconstruction model for limited-angle ct based on prior image induced relative total variation. *Applied Mathematical Modelling*, 74: 586–605, 2019.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *WACV*, pp. 4430–4441. IEEE, 2025.

Robert Haase, Thomas Pinetz, Erich Kobler, Zeynep Bendella, Christian Gronemann, Daniel Paech, Alexander Radbruch, Alexander Effland, and Katerina Deike. Artificial t1-weighted postcontrast brain mri: a deep learning method for contrast signal extraction. *Investigative radiology*, 60(2): 105–113, 2025.

Woojung Han, Seil Kang, Youngjun Jun, Min-Hung Chen, Fu-En Yang, and Seong Jae Hwang. Physics in 2-steps: Locking motion priors before visual refinement erases them. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5f2hxtNqpM.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeuRIPS*, 35:26565–26577, 2022.

Florian Knoll, Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P Recht, and Daniel K Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic resonance in medicine*, 81(1):116–128, 2019.

Dmitrii Lachinov, Philipp Seeböck, Julia Mai, Felix Goldbach, Ursula Schmidt-Erfurth, and Hrvoje Bogunovic. Projective skip-connections for segmentation along a subset of dimensions in retinal oct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 431–441. Springer, 2021.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Laurence S. Lim, Paul Mitchell, Johanna M Seddon, Frank G. Holz, and Tien Y. Wong. Age-related macular degeneration. *The Lancet*, 379(9827):1728–1738, 2012.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Bowen Liu, Yue Wu, Ruoyu Chen, Pusheng Xu, Peng Xiao, Zhen Tian, Binwei Huang, Mingguang He, and Danli Shi. Generating 3D Optical Coherence Tomography from 2D Fundus Images via Diffusion Models, 2025a. Preprint.

Bowen Liu, Weiyi Zhang, Peranut Chotcomwongse, Xiaolan Chen, Ruoyu Chen, Pawin Pakaymaskul, Niracha Arjkongharn, Nattaporn Vongsa, Xuelian Cheng, Zongyuan Ge, et al. APTOS-2024 challenge report: Generation of synthetic 3D OCT images from fundus photographs. *arXiv preprint arXiv:2506.07542*, 2025b.

Xuhui Liu, Zhi Qiao, Runkun Liu, Hong Li, Juan Zhang, Xiantong Zhen, Zhen Qian, and Baochang Zhang. DiffuX2CT: diffusion learning to reconstruct ct images from biplanar x-rays. In *European conference on computer vision*, pp. 458–476. Springer, 2024.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4 (5):e210315, 2022.

José Morano, Botond Fazekas, Emese Sükei, Ronald Fecso, Taha Emre, Markus Gumpinger, Georg Faustmann, Marzieh Oghbaie, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Multimodal foundation model and benchmark for comprehensive retinal oct image analysis. *NPJ Digital Medicine*, 8(1):576, 2025.

Andrew Murphy. Windowing (CT). Radiopaedia.org, 2017. URL https://doi.org/10.53347/rID-52108. Accessed: 2026-08-24.

Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hPOt3yUXii.

Thomas Pinetz, Erich Kobler, Robert Haase, Katerina Deike-Hofmann, Alexander Radbruch, and Alexander Effland. Faithful synthesis of low-dose contrast-enhanced brain MRI scans using noise-preserving conditional GANs. In *MICCAI*, pp. 607–617. Springer, 2023.

Thomas Pinetz, Veit Hucke, and Hrvoje Bogunović. Exploiting intermediate reconstructions in optical coherence tomography for test-time adaption of medical image segmentation. In *Medical Imaging with Deep Learning*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Danli Shi, Bowen Liu, Zhen Tian, Yue Wu, Jiancheng Yang, Ruoyu Chen, Bo Yang, Ou Xiao, and Mingguang He. Fundus2globe: Generative ai-driven 3d digital twins for personalized myopia management. *arXiv preprint arXiv:2502.13182*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024.

Yuhan Wang, Suzhi Bi, and Ying-Jun Angela Zhang. Training-free rate-distortion-perception traversal with diffusion. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DwtVUwEwTF.

Maciej Wojtkowski, Vivek Srinivasan, James G Fujimoto, Tony Ko, Joel S Schuman, Andrzej Kowalczyk, and Jay S Duker. Three-dimensional retinal imaging with high-speed ultrahigh-resolution optical coherence tomography. *Ophthalmology*, 112(10):1734–1746, 2005.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

Yuli Wu, Fucheng Liu, Rüveyda Yilmaz, Henning Konermann, Peter Walter, and Johannes Stegmaier. A pragmatic note on evaluating generative models with fréchet inception distance for retinal image synthesis. In *Medical Imaging with Deep Learning*, 2026.

Xing Xie, Jiawei Liu, Huijie Fan, Zhi Han, Yandong Tang, and Liangqiong Qu. Dvg-diffusion: Dual-view guided diffusion model for ct reconstruction from x-rays. *IEEE Transactions on Image Processing*, 2026.

Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10611–10620. IEEE, 2019.

Jiawei Zhang, Ziyuan Liu, Leon Yan, Zhenyu Xiao, and Yuantao Gu. Stage-wise distortion–perception traversal in zero-shot inverse problems with diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wJeR4HQQWy.

Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

A Appendix

A.1 Modality specific pre-processing

We are given raw fundus images (2576×1958) and OCT volumes ($128 \times 885 \times 512$). In the first step, we find the fovea using the FIT (Gervelmeyer et al., 2025) toolbox, and then we extract the vessels from both the OCT and the fundus image. Then we extract a 700 pixel square around the fovea as this is a potential location for the OCT image. Next, we perform template matching, where a grid search using a pre-defined range of angles and scales is used to fit the images together. Afterwards, we aligned the images to a standardized resolution of 128×256 and $128 \times 256 \times 256$, respectively to make the prediction easier. We filtered the eyes using the FIT (Gervelmeyer et al., 2025) toolbox for quality assessment and used the vessel foreground Dice of 0.1 as a threshold for registration success. Both images were normalized between [-1, 1] using their respective max values. These values were manually chosen after visual inspection.

We follow standard CT literature and clip the Hounsfield units between -1350 and 150 for the use-case of lung screening (Murphy, 2017), resample to 1 mm^3, with a resolution of $128 \times 256 \times 256$.

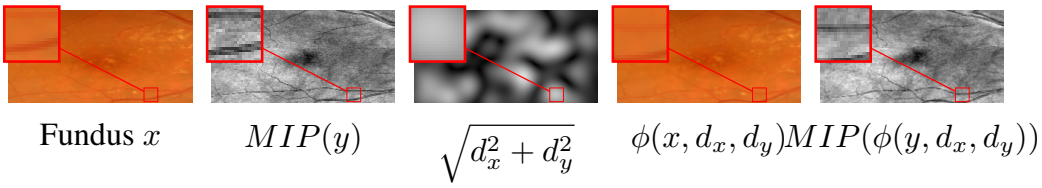

Figure 5: Visualization of the elastic field (d_x, d_y) and its result on the input fundus crop x and the Mean-Intensity Projection (MIP) of the corresponding OCT y.

The synthetically generated X-ray has a resolution of 256×256, and we used the standard values proposed by the GitHub repository example case. Both the X-ray and the CT volumes are rescaled to $[-1,\ 1]$ before being fed into the neural network.

For the flow matching networks, we sampled a random patient and then a random slice from this patient using the raw model output and the corresponding ground truth from stage one. We used standard flow matching augmentation scheme including the following transformations with standard parameters using Albumentations (Buslaev et al., 2020) and listed probability:

- 20% RandomBrightnessContrast changes
- 50% horizontal flips
- 50% vertical flips
- 50% random rotations by 90 degrees
- 50% transpose
- 10% elastic transform
- 10% grid distortion

A.2 Hyperparameters for neural networks

For the flow matching framework, we set $\sigma_\epsilon = 0.1$ and $t_{\text{min}} = 0.05$. In the augmentation pipeline, the elastic transformation was applied to 20% of the samples with $\alpha_\epsilon = 200$. All networks were optimized with a learning rate of $1e-4$, utilizing a cosine annealing schedule decaying to 10^{-6}.

The ED and PFM baselines utilize the same 2D-to-3D lifting strategy and network architecture. It is a classical 1×1 convolutional kernel that is used to upsample to the correct final shape. Then, we adapted the Song U-Net (Karras et al., 2022) for 3D by replacing all the convolutions with 3D counterparts and removing the attention layers for memory reasons. For all methods except LDM, we applied the same augmentation pipeline, as it is unclear how the augmentation would affect the features of the ground truth. We still applied all the augmentations that do not alter the final image, such as Gaussian noise. For LDM, the fundus image was provided as an additional input channel for the diffusion model. We both tried to train a VAE on the training data and use a foundational VAE according to Chen et al. (2026). For CT, the foundational VAE works better than the newly trained one, but for OCT it is the other way around. We attribute this to the underlying acquisition mechanisms. Unlike OCT, which records individual B-scans independently, CT relies on projection rays that span across slices, resulting in highly correlated inter-slice noise. Every model was trained for 3 days on a single NVIDIA A6000 GPU.

A.3 OCT specific augmentation

To improve the model's robustness to morphological variations, we introduce a 2D-to-3D elastic transformation. We generate a random distortion field $[-\alpha_\epsilon, \alpha_\epsilon]^{h_x \times w_x \times 2}$, smoothed via a Gaussian filter ($\sigma_f = 15$), to create displacement maps d_x, d_y. This deformation is applied to the fundus image x via linear interpolation. Critically, to maintain anatomical correspondence, the same transformation is applied slice-wise to the A-scans of the OCT volume y (Fig. 5).

In Table 5, an ablation is shown on the addition of the OCT-specific augmentation. The inclusion of our 2D-to-3D elastic deformation significantly improves data-fidelity metrics. The PSNR and

SSIM increased significantly ($p < 0.05$), while the standard deviation across samples decreased, indicating more stable training. We observed only a marginal increase in FID and dice scores in vessel segmentation, suggesting that this increase primarily reinforces structural consistency. However, there is a slight drop in layer segmentation, which might be because the warping sometimes introduces unrealistic layer structure.

Table 5: Results on the in-house early/intermediate AMD dataset between our model with and without augmentation. PSNR and SSIM showcase the consistency with the measurement, while FID showcases the plausibility of the resulting synthesis.* denotes statistical significance to the closest competitor.

	PSNR↑	SSIM↑	FID↓	Dice Layers↑	Dice Vessels ↑
TransUpNet (no augmentation)	19.2 ± 1.2	25.6 ± 3.2	413.4	35.5 ± 4.9	34.0 ± 12.5
TransUpNet	$\mathbf{19.4} \pm 1.2$	$\mathbf{26.6}^{*} \pm 2.7$	435.0	32.6 ± 4.9	45.9 ± 10.9

A.4 Layer & Vessel Segmentation Networks

For the detection of the retinal layers, we employed SD-RetinaNet Fazekas et al. (2025), a state-of-the-art architecture optimized specifically for OCT data. We trained this model on a proprietary internal dataset featuring expert-level annotations for both retinal layers and lesions. A core advantage of SD-RetinaNet is its differentiable topology engine, which guarantees that the resulting segmentations strictly adhere to the established anatomical structure of the retina.

The dataset consists of 45,037 B-scans sourced from 1,105 volumes (corresponding to 764 eyes across 625 patients). Image acquisition spanned three distinct devices: Heidelberg Spectralis (30,967 B-scans; 68.8%), Heidelberg Spectralis High-Res (7,853 B-scans; 17.4%), and Zeiss Cirrus (6,217 B-scans; 13.8%), where Zeiss Cirrus is very similar to Topcon Maestro 2. The patient cohort is diverse, capturing healthy retinas as well as a wide array of pathologies, including iAMD, GA, nAMD, DME, RVO, and high myopia.

A key characteristic of the ground truth data is its multidimensional sparsity. While lesion masks are extensively available (44,016 B-scans, or 97.7%), layer boundaries are provided for only 13.5% of the dataset (6,100 B-scans), with a mere 5,102 B-scans containing both. Within any given volume, layer delineations are generally limited to a small fraction of the B-scans (median 20%). Lesion class coverage is similarly uneven at the volume level, ranging from 8,892 B-scans annotated for PED to 29,826 for IRF, and certain lesion types were exclusively annotated for specific scanner brands. Finally, because a large portion of the lesion annotations are negative cases (empty masks indicating the absence of a lesion), we undersampled these negative examples to 10% to prevent severe class imbalance during training.

For the vessel segmentation algorithm, we use an in-house model, trained on a large collection of OCTs with corresponding ground truth. A schematic drawing is shown in Fig. 6. It closely mimics a standard U-Net architecture with residual blocks, where average pooling is used for the skip connections. The residual blocks use two convolutions with fixed feature size, instance norm, and ReLU. The dataset is described by Lachinov et al. (2021).

To prevent evaluation bias and target circularity, both downstream segmentation networks were trained on an independent, disjoint patient cohort. Crucially, all comparative baselines and our framework were subjected to the exact same standardized pre-processing and evaluation protocol.

A.5 Comparison to dedicated architectures

Note that the only dedicated architecture to the best of our knowledge for the CFP to OCT task, namely Fundus2OCT (Liu et al., 2025a) was omitted from quantitative comparison as no public code or trained checkpoints are available. Furthermore, their primary contribution, post-hoc elastic deformation to account for retinal curvature, is rendered redundant under our planar Bruch's membrane flattening protocol.

For the other task, the dataset split provided by Ying et al. (2019) also provides a processed training and testing dataset. In their work, they employed a different CT-based clipping, between $[0, 2500]$

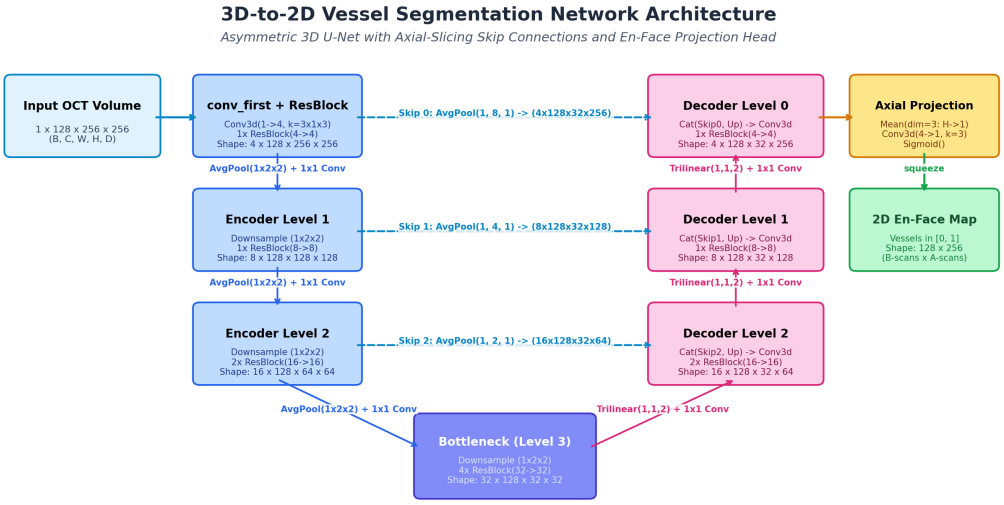

Figure 6: Network architecture of the vessel segmentation algorithm.

and use a larger FOV with 256 slices, which were subsequently resized to 128. However, in their PSNR calculations they used the maximal pixel value of 4095, which is not actually the maximum observed value. Nevertheless, we also trained the TransUpNet on this dataset and added the results in Table 6 using their evaluation protocol. Due to the larger field of view, which includes many empty slices and higher range of values that are typically not present in lung CT and furthermore clipped, the resulting PSNR values are significantly higher than in our setup. One thing to note is that the values shown here slightly change compared to the reported paper values (PSNR: 22.2 vs 22.3 and SSIM: 52.8 vs 52.5). We also added the values for XctDiff directly from the paper, as no model checkpoint is available, where X2CT-GAN performs approximately the same as here (22.66 vs 22.2 PSNR). The values however are not directly applicable as the pre-processing in XctDiff scales the iamges between [-1,1] instead of clipping. However, according to the repository the pre-processing is different again, as they are using actual images and a different clip window. Nevertheless, we also outperform specifically designed architectures for this task. Moreover, we value the clinical applicability and designed our pre-processing to closely mimics the real clinical practice in radiology, even though the values are substantially lower. Three example images are shown in Fig.7.

Table 6: Results on the public CT dataset (Armato III et al., 2011), where the pre-processing follows Ying et al. (2019). PSNR and SSIM showcase the consistency with the measurement, while LPIPS showcases the plausibility of the resulting synthesis.* denotes statistical significance to the closest competitor, where we computed the results. Values for methods using† are taken from the paper itself in the absence of an official checkpoint.

	PSNR↑	SSIM↑	LPIPS↓	Runtime in s↓
X2CT-GAN (Ying et al., 2019)	<u>22.2</u> $\pm$ 1.5	<u>52.8</u> $\pm$ 6.9	<u>0.21</u> $\pm$ 0.05	**0.2**
XctDiff† (Bai et al., 2024)	23.6 $\pm$ 1.2	52.4 $\pm$ 7.5	**0.12**	-
TransUpNet	**23.6*** $\pm$ 1.6	**57.2*** $\pm$ 6.6	0.24 $\pm$ 0.04	0.3
Ours (TransUpNet + Flow)	22.8 $\pm$ 1.4	55.6 $\pm$ 7.0	0.17* $\pm$ 0.04	55.0

A.6 Flow matching network architecture

We also compared a transformer architecture to the classical CNN-based architecture, which we adopted for this task in Tab.7 for the OCT reconstruction task. The inference speed is quicker for the JiT (Li & He, 2026) model, however, the eventual performance is inferior, especially in FID. As we deem the inference speed to be less important in clinical applications and we reached a lower FID score using the classical U-Net and the entire point is that we can interpolate between the extrema,

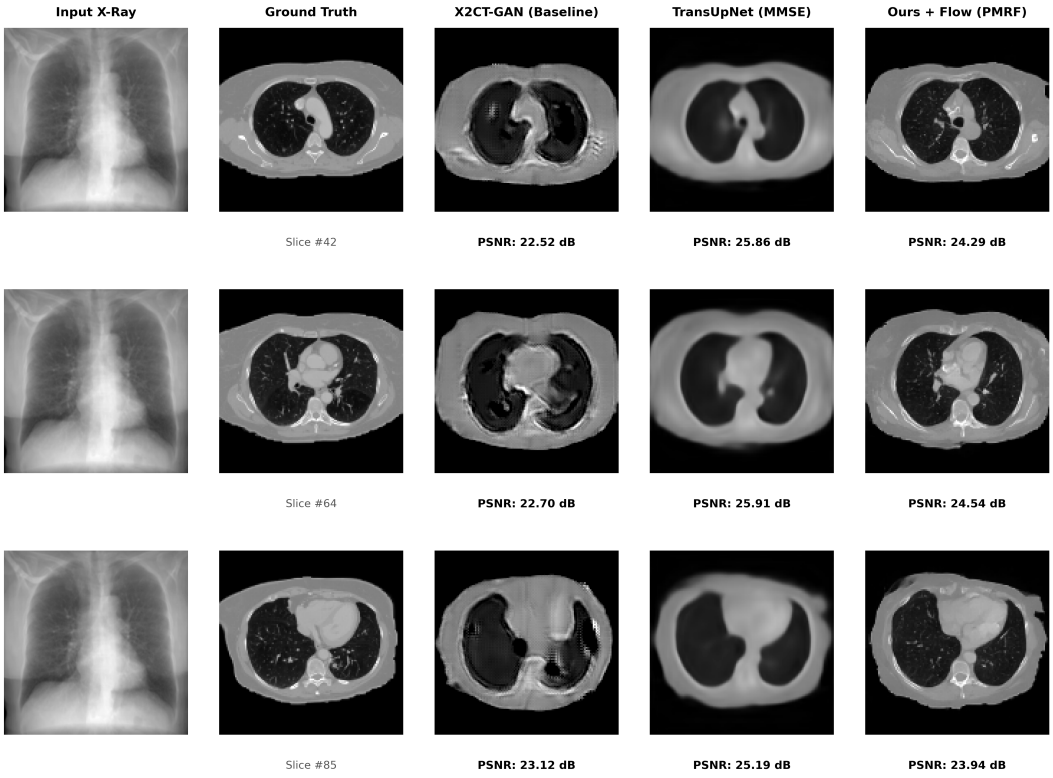

Figure 7: Comparison with the model from Ying et al. (2019), showing that differences in pre-processing setups yield vastly different PSNR values, without noticeable differences in image quality compared to our pre-processing.

we advocate for the usage of 2D U-Nets over diffusion transformers, at least for the medium sized datasets considered in this work.

Table 7: Results on the intermediate AMD dataset to compare Song U-Net to a JiT model.

	PSNR↑	SSIM↑	FID↓	Dice Layer↑	Dice Vessel↑
TransUpNet + JiT	$\mathbf{18.3} \pm 0.9$	$\mathbf{20.4} \pm 2.2$	47.3	$\mathbf{49.6} \pm 6.0$	45.7 ± 12.3
TransUpNet + Song	18.0 ± 0.9	19.1 ± 2.0	**18.0**	48.3 ± 5.5	$\mathbf{48.0}^{*} \pm 11.1$

A.7 Additional Examples

We show some additional examples and an example of the layer segmentation for all scans in Figure 8.

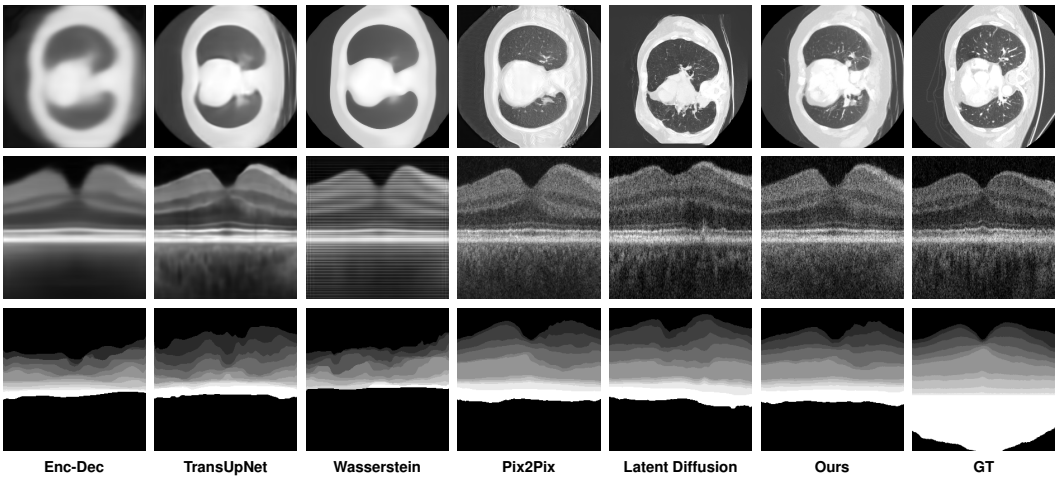

Figure 8: Further example with a segmentation given by our downstream model. Only in Pix2Pix and our model the layers are at least mostly correct.