Conditioning 3D Diffusion Models with 2D Images: Towards Standardized OCT Volumes through *En Face*-Informed Super-Resolution

Coen de Vente[1,2,3] **Mohammad Mohaiminul Islam**[1,2] **Philippe Valmaggia**[4,5]
Carel Hoyng[6] **Adnan Tufail**[7] **Clara I. Sánchez**[1,2]
on behalf of the MACUSTAR consortium*
[1]qurAI Group, Informatics Institute, University of Amsterdam, The Netherlands
[2]Amsterdam UMC location University of Amsterdam, Biomedical Engineering and Physics, The Netherlands
[3]DIAG, Department of Radiology and Nuclear Medicine, Radboudumc, The Netherlands
[4]Department of Biomedical Engineering, Universität Basel, Switzerland
[5]Department of Ophthalmology, University Hospital Basel, Switzerland
[6]Department of Ophthalmology, Radboudumc, Nijmegen, The Netherlands
[7]Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
{c.w.devente,m.islam,c.i.sanchezgutierrez}@uva.nl
philippe.valmaggia@unibas.ch carel.hoyng@radboudumc.nl
adnan.tufail@nhs.net

Abstract

High anisotropy in volumetric medical images can lead to the inconsistent quantification of anatomical and pathological structures. Particularly in optical coherence tomography (OCT), slice spacing can substantially vary across and within datasets, studies, and clinical practices. We propose to standardize OCT volumes to less anisotropic volumes by conditioning 3D diffusion models with *en face* scanning laser ophthalmoscopy (SLO) imaging data, a 2D modality already commonly available in clinical practice. We trained and evaluated on data from the multicenter and multimodal MACUSTAR study. While upsampling the number of slices by a factor of 8, our method outperforms tricubic interpolation and diffusion models without *en face* conditioning in terms of perceptual similarity metrics. Qualitative results demonstrate improved coherence and structural similarity. Our approach allows for better informed generative decisions, potentially reducing hallucinations. We hope this work will provide the next step towards standardized high-quality volumetric imaging, enabling more consistent quantifications.

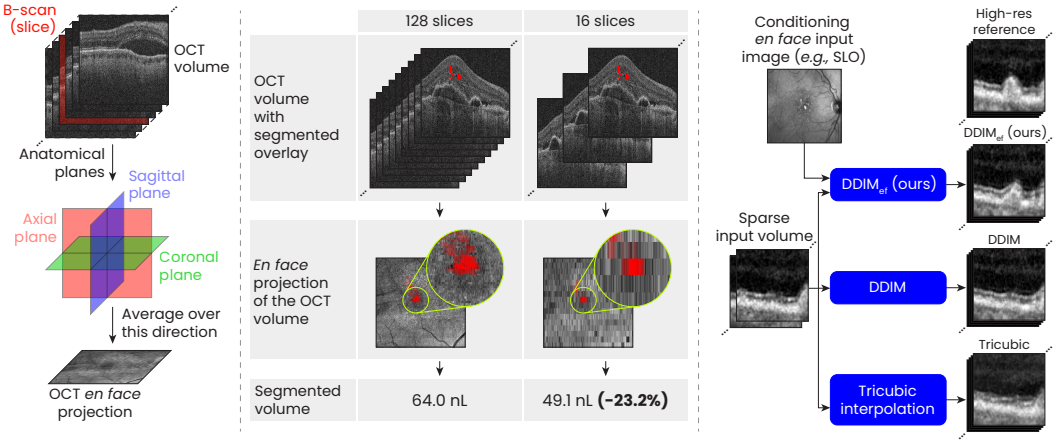
1 Introduction

Volumetric medical images can be highly anisotropic, i.e., having high-resolution slices in one anatomical plane but poor through-plane resolution. This has been shown to lead to imprecise volume and shape measurements of structures of interest [17], potentially resulting in wrong diagnoses and severe negative clinical implications.

A prominent example of a modality often affected by this is optical coherence tomography (OCT). OCT is commonly acquired as a raster, where multiple line scans generate B-scans (slices), and multiple slices generate a volume (see Fig. 1a). The spacing between slices can vary substantially

*The list of MACUSTAR consortium members is in Appendix Section A.7.

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.



(a) Medical terminology. (b) Problem statement. (c) Our solution versus baselines.

Figure 1: Overview of our proposed approach and background. (a) The medical terminology related to OCT volumes. (b) Low slice densities in volumetric images can lead to inaccurate quantifications. (c) We propose to upsample the number of slices by conditioning a diffusion model with *en face* images. In our experiments, we use SLO as conditioning *en face* data but our method could be extended to other types. In the shown example, our model is the only approach that correctly generates the druse (the bright bump in the retina).

between OCT devices and imaging protocols [32, 3, 6, 8]. This can hamper consistent biomarker quantification [30, 2, 27, 19, 18]. Fig. 1b illustrates this issue by demonstrating fluid volume estimations in a single retinal OCT volume for various slice spacings. We can observe a 23.2% drop in estimated fluid volume when the slice spacing increases by a factor of 8.

A possible solution to these imprecise measurements is to standardize volumes with low slice density to high-resolution data through reliable super-resolution methods. Several super-resolution approaches have been proposed for OCT [12, 33, 4, 21] but they all aim to improve the resolution within individual B-scans. Approaches to reduce anisotropy have been proposed for other volumetric medical images, such as computed tomography (CT) [15] and magnetic resonance imaging (MRI) [26]. These methods use deep learning models to upsample the number of slices based solely on low-resolution input data during inference.

A major drawback of these works is their lack of knowledge about anatomical and pathological structures that fall between two adjacent slices. This can lead to hallucinating models that generate incorrect biological features, potentially resulting in misdiagnoses or otherwise harmful clinical outcomes. We hypothesize that including information about regions between slices as input to a super-resolution model helps make better-informed generative decisions that correctly reflect the biological truth.

Therefore, we propose a method based on 3D diffusion models to increase slice density by utilizing additional imaging modalities as conditioning information (see Fig. 1c). We use diffusion models, as they have been shown to outperform other popular generative models such as generative adversarial networks at generating high-quality images [5], super-resolution [24], and leveraging multimodal data as conditioning information [22]. These capabilities align well with the objectives of our study.

We evaluate our developed method on OCT data while conditioning on scanning laser ophthalmoscopy (SLO) fundus images. SLO is a 2D *en face* (i.e., parallel to the coronal plane) imaging modality that is commonly acquired alongside OCT scans. OCT devices internally use SLO images as a reference to position the OCT acquisition at the desired anatomical location [1]. Our method can be extended to include other modalities, such as color fundus photography (CFP) and fundus autofluorescence (FAF), potentially resulting in even better-informed models.

We hope this approach is a valuable step towards more isotropic, high-quality, and standardized volumetric imaging, allowing for more consistent biological measurements and diagnoses in the future.

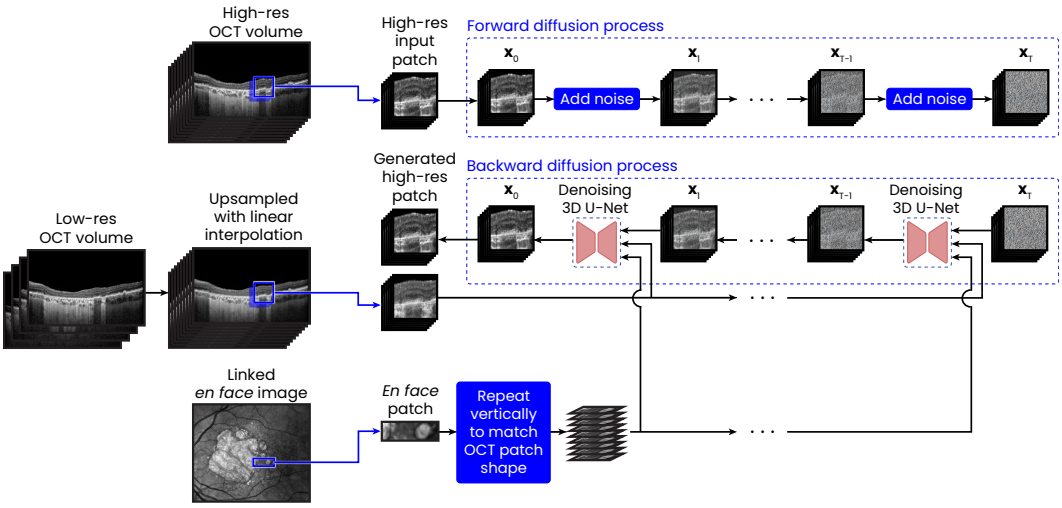


Figure 2: *En face*-conditioned diffusion model overview.

2 Methods

Our model is a 3D diffusion model trained to generate high-resolution volumes by upsampling the number of slices (i.e., increasing the B-scan density in OCT). It is conditioned on both *en face* imaging data and the low-resolution counterpart of the high-resolution target (see Fig. 2). We provide a brief introduction to diffusion models and their associated symbols in Section 2.1. In Section 2.2, we describe how we adapt diffusion models for *en face* conditioned super-resolution. In short, the low-resolution image is concatenated with a reshaped *en face* image along the channel dimension, which we subsequently input as conditional information to the denoising model. The sampling process, including the use of Denoising Diffusion Implicit Model (DDIM) [28] sampling, overlapping patches enabled by RePaint [16], and classifier-free guidance (CFG) [11], is detailed in Section 2.3. Finally, we present the network architecture and implementation details in Section 2.4.

2.1 Diffusion models

Diffusion models are generative models consisting of a forward diffusion process and a backward diffusion process [10]. In the forward diffusion process, over many timesteps T, more and more noise is gradually added to an input image $\mathbf{x}_0$, resulting in noisy images $\mathbf{x}_1, \ldots, \mathbf{x}_T$. This process q can be formulated with a variance schedule $\beta_1, \ldots, \beta_T$ as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

As shown by Ho *et al.* [10], we can directly obtain $\mathbf{x}_t$ given $\mathbf{x}_0$ using the following equation:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

In the backward diffusion process p_θ, $\mathbf{x}_{t-1}$ is predicted for any $t \in \{1, \ldots, T\}$ by denoising $\mathbf{x}_t$ using a trained denoising model, optimized by model parameters θ. This model generally uses some variant of the U-Net [23] architecture (see Section 2.4 for the implementation we use). Following Salimans *et al.* [25], our trained denoising model does not predict the noise $\boldsymbol{\epsilon}$ or image $\mathbf{x}_0$ directly, but uses $\mathbf{v}$-prediction parameterization as this prevents intensity shifting artifacts in super-resolution models [9], where $\mathbf{v}_t := \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\mathbf{x}_0$. We use the mean squared error (MSE) loss to train our denoising model v_θ:

$$\mathcal{L} := ||v_\theta(\mathbf{x}_t, t) - \mathbf{v}_t||_2^2. \tag{3}$$

2.2 *En face* conditioned super-resolution

To enable the use of diffusion models for *en face* conditioned super-resolution, we combine the approach for generating high-resolution from low-resolution images from SR3 [24] with our proposed method of conditioning the denoising model with *en face* information. In this paper, we use SLO images but our approach could be extended to other *en face* data.

Specifically, we adapt the input to the conditional denoising model as follows:

$$\hat{\mathbf{v}}_t = v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}), \tag{4}$$

where $\hat{\mathbf{v}}_t$ is the output of the denoising model, $\mathbf{x}_t$ is the noisy image at timestep t, $\mathbf{x}_{LR}$ is the corresponding low-resolution image upsampled to match the dimensions of $\mathbf{x}_t$ using linear interpolation, and $\mathbf{x}_{enface}$ is the *en face* image.

$\mathbf{x}_{LR}$ and $\mathbf{x}_{enface}$ are fed similarly into the denoising model. Following SR3 [24], we concatenate the noisy input $\mathbf{x}_t$ with the low-resolution image $\mathbf{x}_{LR}$ along the channel dimension. Since *en face* images generally do not align with their corresponding OCT scans in the *en face* plane, registration of these two images is required (see Section A.1). This registration step results in spatial correspondence and ensures the *en face* image has the same shape in the *en face* plane as the OCT volume. To reach an image with the same 3D shape as $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, we repeat the registered 2D $\mathbf{x}_{enface}$ image in the y-direction H times, where H is the height of $\mathbf{x}_t$. This 3D tensor is concatenated together with $\mathbf{x}_t$ and $\mathbf{x}_{LR}$ along the channel dimension. We subsequently input the resulting tensor into the model.

Due to computational limitations, we work with image patches for $\mathbf{x}_t$, $\mathbf{x}_{LR}$, and $\mathbf{x}_{enface}$. These patches all correspond in terms of their location and size. No noise is added to $\mathbf{x}_{LR}$ and $\mathbf{x}_{enface}$ for any timestep t.

2.3 Sampling process

During sampling, there are a few key distinctions in the processing pipeline compared to the backward diffusion process during training. Firstly, we use DDIM [28] sampling, which allows for accelerated sampling by reducing the number of timesteps while sampling.

Secondly, we train our denoising model with patches, but we are interested in generating full high-resolution volumes during sampling. To prevent artifacts near the borders of patches, we use overlapping patches. We use RePaint [16] to facilitate this overlapping strategy, an image inpainting approach for diffusion models (see Section A.2 for more details).

Thirdly, to minimize image artifacts showing an overall intensity in the generated slices that is different from the overall intensity in the slices already existing in the low-resolution volume, we implemented a post-processing normalization step. In this step, we normalized the mean and standard deviation of the intensities in the generated slices to match those in the slices that already existed in the low-resolution volume. This normalization step was performed separately for each OCT volume.

Lastly, to influence how much the denoising model uses the *en face* information for its generative decisions, we employ CFG [11]. In CFG, during training, the conditional information is dropped for a random number of samples in each batch with some probability p_{uncond}. In practice, when the conditional information is dropped, we feed an image with all pixels set to zero. This results in a jointly trained conditional denoising model $v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface})$ and unconditional denoising model $v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR})$. During sampling, we can then linearly combine the conditional and unconditional model predictions using a guidance scale hyperparameter w:

$$\tilde{v}_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}) = (1-w)v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}) + wv_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}). \tag{5}$$

2.4 Network architecture and implementation details

For our denoising network, we use the U-Net architecture described by Pinaya *et al.* [20], in which the timestep embedding is integrated into the model via residual connections. The U-Net uses 3D convolutions and has a depth of four U-Net levels with 32, 64, 128, and 256 channels, respectively, in each level with two residual blocks per level. We use self-attention at the deepest U-Net level with a

single attention head. We train using the Adam optimizer with a learning rate of 5×10^{-5}, a batch size of 16, for 20 000 epochs. All images were normalized between -1 and 1 before cropping.

In our experiments, the number of timesteps for the diffusion process was $T = 1000$ during training with a SCALED LINEAR [20] schedule and a β_t range of 0.0005 to 0.0195. We employed DDIM sampling with 100 timesteps, resulting in a $10\times$ time efficiency improvement. We do not perform any resampling steps when inpainting using RePaint [16] to allow for faster sampling. We use the MONAI generative AI framework [20] for implementing our diffusion model.

During training, we use 3D patches of size $128 \times 128 \times 16$ for $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, and a 2D patch of size $128 \times 1 \times 16$ for $\mathbf{x}_{enface}$. During sampling, we use a patch size of $496 \times 496 \times 16$ for $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, and a 2D patch of size $496 \times 1 \times 16$ for $\mathbf{x}_{enface}$. When sampling full high-resolution volumes, we used an overlap of 25%, 25%, and 50% for the x-, y-, and z-direction, respectively. During training, we prepared our patches such that each 5$^{\text{th}}$ and each 13$^{\text{th}}$ slice in $\mathbf{x}_{LR}$ were identical to the 5$^{\text{th}}$ and 13$^{\text{th}}$ slice in $\mathbf{x}_0$, respectively. The other slices were interpreted using linear interpolation.

2.5 Data

For training and evaluating our diffusion models, we used OCT volumes and corresponding SLO images from the MACUSTAR study [7]. MACUSTAR is a clinical study on age-related macular degeneration (AMD) that is carried out across 20 sites in 7 European countries. The dataset from this study contains data from patients with varying disease severities (no, early, intermediate, and advanced AMD).

We used the patient data from the cross-sectional part of the MACUSTAR study and only included Heidelberg Spectralis OCTs with at least 237 B-scans. This resulted in a total set of 302 patients. We randomly split this set of patients in 181 (60%), 60 (20%), and 61 (20%) patients for training, validation, and testing, respectively. As multiple OCT volumes were available for each patient (from both eyes and either one, two, or three visits in the cross-sectional study) this resulted in 721 and 236 OCT volumes for the training and validation set, respectively. For the test set, we only used the OCT volume from the study eye, defined for the MACUSTAR study, from the first visit, resulting in 61 OCT volumes. More details regarding the dataset and pre-processing can be found in Appendix A.1.

3 Experiments

We evaluate our approach for the task of upsampling the number of slices in the image volume with a factor of 8. For the sake of simplicity, when generating full volumes as described in Section 2.3, we drop the last slice in the OCT volume in the test set, resulting in OCT volumes with 240 instead of 241 slices. Hence, we upsample low-resolution volumes with 30 slices to high-resolution volumes with 240 slices. The resolution of individual slices was not changed.

We refer to our proposed diffusion model with *en face* conditioning and CFG with guidance scale $w = 2$ as DDIM$_{\text{ef}}$. To measure the effect of *en face* conditioning and CFG, we compare DDIM$_{\text{ef}}$ with the two models: DDIM$_{\text{ef}}$ (no CFG), and DDIM. DDIM is the proposed approach with *en face* conditioning turned off during sampling. *En face* conditioning is turned off by feeding an image with all pixels set to zero as the conditional image, which is also done during training with a probability of p_{uncond} to enable CFG (see Section 2.3). The proposed model with *en face* conditioning, but without CFG, will be referred to as DDIM$_{\text{ef}}$ (no CFG). Furthermore, we compare these methods with the more traditional upsampling method of tricubic interpolation.

3.1 Evaluation

We report the classical image similarity metrics MSE, SSIM [31], and PSNR, computed between the $8\times$ upsampled low-resolution images using DDIMs and tricubic interpolation, and the high-resolution reference images. As noted by Saharia *et al.* [24], these classical metrics may not be optimal for evaluating super-resolution methods, as they were shown to correlate poorly with human perception and heavily penalize synthetic high-frequency details that deviate from the reference, favoring blurry images instead.

Therefore, we also evaluate with Learned Perceptual Image Patch Similarity (LPIPS) [34]. We used the LPIPS metric implementation provided by the authors of the original LPIPS paper [34]

Table 1: Classical image similarity metrics (MSE, SSIM, and PSNR) and perceptual metrics (all LPIPS variants) calculated on the test set between the original high-resolution OCT volumes and low-resolution images that were $8\times$ upsampled in the slice-direction using tricubic interpolation and our proposed DDIM methods. Results are presented as the mean $\pm$ standard deviation over all OCT volumes. Bolded values indicate the best values in terms of the mean performance.

	Tricubic	DDIM	DDIM_{ef} (no CFG)	DDIM_{ef}
MSE $\downarrow$	**0.006** $\pm$ **0.002**	**0.006** $\pm$ **0.003**	**0.006** $\pm$ **0.003**	**0.006** $\pm$ **0.003**
SSIM $\uparrow$	**0.451** $\pm$ **0.116**	0.444 $\pm$ 0.107	0.447 $\pm$ 0.107	0.447 $\pm$ 0.107
PSNR (dB) $\uparrow$	22.472 $\pm$ 1.418	22.401 $\pm$ 1.644	**22.495** $\pm$ **1.673**	22.450 $\pm$ 1.683
$\text{LPIPS}_{\text{axi}}$ $\downarrow$	**0.120** $\pm$ **0.027**	0.138 $\pm$ 0.030	0.138 $\pm$ 0.030	0.141 $\pm$ 0.031
$\text{LPIPS}_{\text{cor}}$ $\downarrow$	0.548 $\pm$ 0.103	**0.158** $\pm$ **0.047**	**0.158** $\pm$ **0.048**	0.162 $\pm$ 0.050
$\text{LPIPS}_{\text{sag}}$ $\downarrow$	0.540 $\pm$ 0.088	**0.144** $\pm$ **0.049**	**0.144** $\pm$ **0.049**	0.147 $\pm$ 0.050
$\text{LPIPS}_{\text{2.5D}}$ $\downarrow$	0.403 $\pm$ 0.072	**0.147** $\pm$ **0.041**	**0.147** $\pm$ **0.042**	0.150 $\pm$ 0.043
$\text{LPIPS}_{\text{efproj}}$ $\downarrow$	0.231 $\pm$ 0.055	0.063 $\pm$ 0.039	**0.060** $\pm$ **0.039**	0.064 $\pm$ 0.039

for their metric based on an ImageNet pre-trained AlexNet [14]. Since this evaluation method was developed for 2D images, we modified it for 3D data by calculating the LPIPS metric on all 2D slices in the axial, coronal and sagittal planes in the volume, resulting in the metrics $\text{LPIPS}_{\text{axi}}$, $\text{LPIPS}_{\text{cor}}$, and $\text{LPIPS}_{\text{sag}}$, respectively. We also report $\text{LPIPS}_{\text{2.5D}}$, which is the average of these three metrics. This approach for 3D data is available in the publicly available implementation from the MONAI generative AI framework [20]. Additionally, we calculated $\text{LPIPS}_{\text{efproj}}$, which compares two OCT *en face* projections generated by averaging all columns in each slice of the volume, resulting in a 2D *en face* image.

Even though LPIPS may be considered a more suitable evaluation approach than the classical evaluation metrics, it is not guaranteed to consider the structures of interest enough, since the underlying model was only trained on natural images. Therefore, we also present qualitative examples.

4 Results

The first three rows of Table 1 shows the classical image similarity metrics MSE, SSIM [31], and PSNR, computed between the $8\times$ upsampled low-resolution images using DDIMs and tricubic interpolation, and the high-resolution reference images. The perceptual metrics based on LPIPS are shown in the last five rows of Table 1.

For the proposed method, ablated methods and the tricubic interpolation method, we present several figures to illustrate the difference in structural similarity to the high-resolution reference images, sharpness, and coherence within the generated volumes. In Fig. 3 and A.2, we aim to point out these aspects using the *en face* projections, allowing one to observe the overall structure of the full generated volumes. Fig. 4 and Fig. A.3 show additional examples of these *en face* projections while highlighting relevant regions, alongside the corresponding image patches from the underlying OCT volume. Further examples are presented in Fig. A.4, which shows 3D renders of generated and reference volumes, and in Fig. A.5, which shows patches of consecutive B-scans with difference maps between generated and reference images. Fig. A.6 shows generation examples and difference maps for several randomly picked image patches. The effect of increasing the guidance scale w from CFG is illustrated in Fig. 5. Specifically, a case of an example OCT patch is shown with several drusen that are also visible in the SLO image.

Model training took approximately 9 days on an NVIDIA A100 GPU. Sampling a full volume with 240 B-scans of 768×496 pixels, the most common size in the test set, resulted in 58 patches. On the aforementioned GPU type, for DDIM_{ef}, sampling a whole volume took approximately 46 minutes. For DDIM and DDIM_{ef} (no CFG), this sampling time was about half (approximately 23 minutes), as CFG doubles the number of required forward passes of the denoising model.

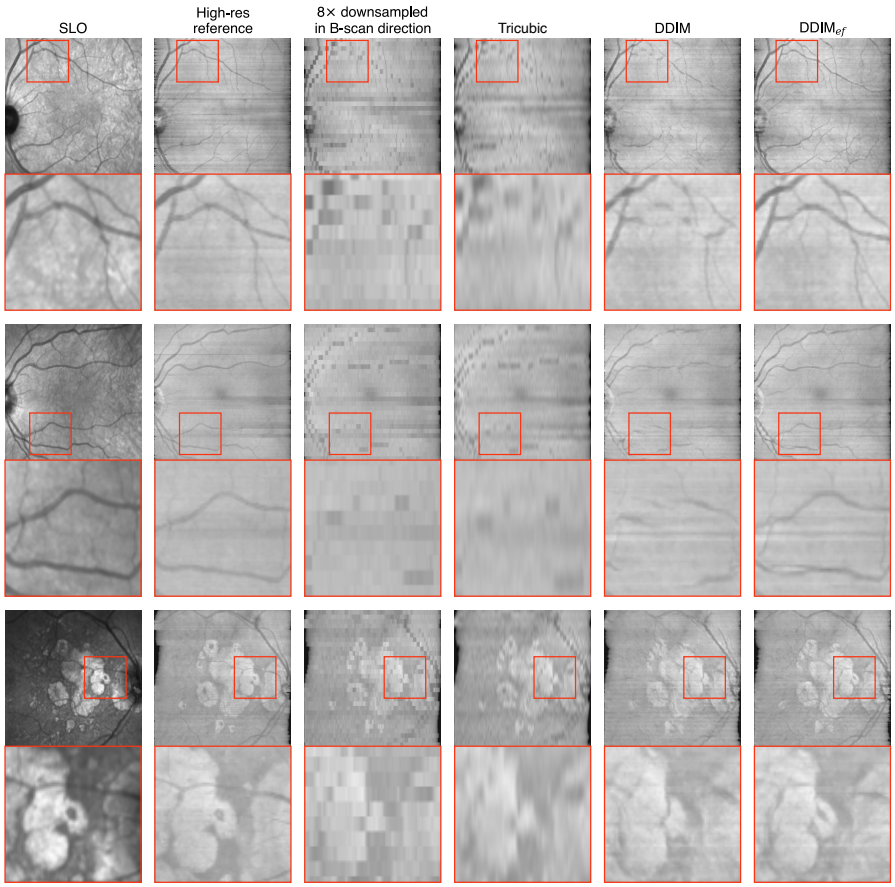


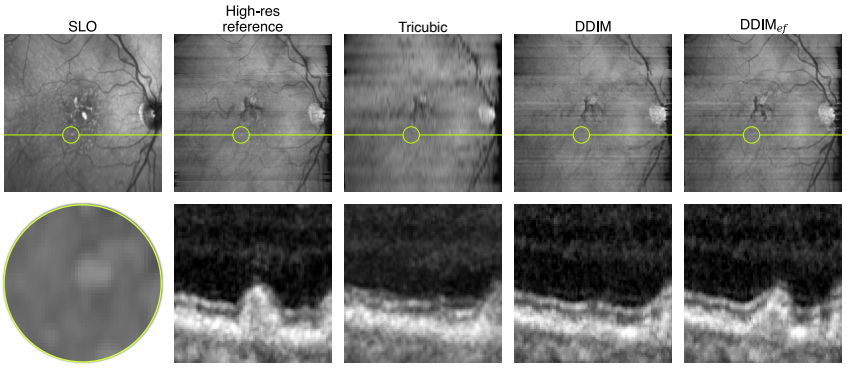
Figure 3: Examples of *en face* projections of the OCT volumes generated using tricubic interpolation, the unconditional diffusion model DDIM, and our proposed *en face* conditioned diffusion model DDIM$_{\text{ef}}$. These three projections are shown in the last three columns and were all generated from the 8 $\times$ downsampled (in the B-scan direction) volume as input, which is shown in the third column. The first and second rows show the corresponding high-resolution (high-res) reference and scanning laser ophthalmoscopy (SLO) image, respectively. A separate example from a different test set patient is shown in each row. The top images in each row show the full image. Zoomed-in versions of the image patches (red boxes) are shown at the bottom of each row.

5 Discussion

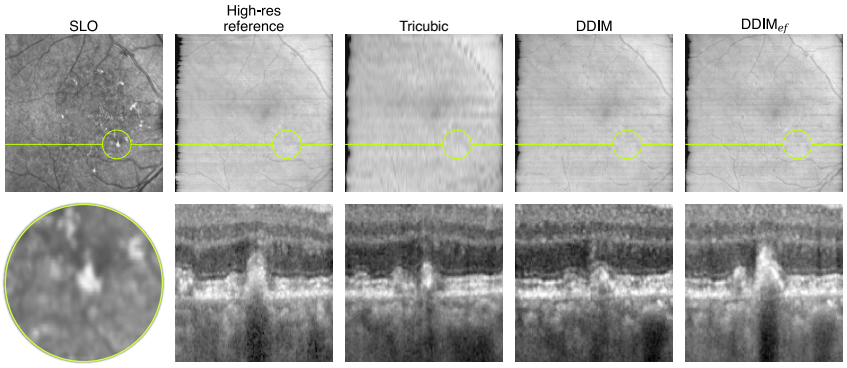
We addressed the large variability in anisotropy across OCT scans, which can lead to inconsistent quantifications. We propose a super-resolution approach that uses SLO images to condition 3D super-resolution diffusion models, aiming for better informed image generations that are closer to the biological truth. SLO is commonly already acquired alongside OCT scans, ensuring our method often will not require any additional data beyond what is already available in clinical practice. Furthermore, SLO acquisition is relatively fast, while OCT acquisition time increases with every additional line to be acquired, potentially speeding up the overall acquisition process.

Our qualitative results indicate that our approach can upsample the number of B-scans in OCT volumes by a factor of 8 while improving similarity to high-resolution reference images and overall coherence, compared to a diffusion-based approach without *en face* information as conditional input. Our method specifically improves the reconstruction of superficial blood vessels and geographic atrophy (see Fig. 3 and A.2). Furthermore, our diffusion models demonstrate visually sharper images than tricubic interpolation.

In terms of the classical image similarity metrics MSE, SSIM, and PSNR, which are suboptimal for evaluating super-resolution methods [24], tricubic interpolation performed roughly the same as our



(a)



(b)

Figure 4: Examples showing the effect of using DDIM_{ef}, compared to leaving out the *en face* information (DDIM), and tricubic interpolation. In the top rows, the SLO image is shown on the left, followed by the *en face* projections of each OCT volume. The bottom rows show zoomed-in patches. The patch locations are indicated in the top row with a lime circle. The lime horizontal line corresponds to the B-scan location from the shown patches. (a) DDIM_{ef} reconstructs a druse that is present in the high-resolution reference, while it is missing in the OCT patches from tricubic interpolation and DDIM. The druse also seems to be subtly visible in the SLO image. (b) The large lesion in the center of the OCT patches are more similar to the reference for DDIM_{ef} than for tricubic interpolation and DDIM. It is also visible in the SLO images as a hyper-intense lesion.

diffusion models. In terms of $\text{LPIPS}_{\text{cor}}$, $\text{LPIPS}_{\text{sag}}$, and $\text{LPIPS}_{\text{efproj}}$, our diffusion models outperformed tricubic interpolation but slightly underperformed in terms of $\text{LPIPS}_{\text{axi}}$. This finding is in line with our visual observations in Appendix Section A.4.

In terms of all reported quantitative metrics, using CFG to guide the diffusion model more towards the information in the *en face* image either slightly decreased performance, or showed no effect (see Table 1). Paradoxically, we found using CFG could lead to structural features that more closely resembled those in the high-resolution reference than when CFG was not employed. An example of this effect is shown in Fig. 5. Besides, this example shows that setting the guidance scale w too high can lead to exaggerated structural features (e.g., too large drusen) and artifacts.

This study has limitations. (1) Our DDIM models can sometimes introduce imaging artifacts (see Appendix Section A.5). (2) When visually inspecting the dataset, we found that the registration information between OCT B-scans and the SLO image was not always perfect, possibly not allowing the model to utilize the SLO/OCT mapping well sometimes. An improved registration strategy will likely improve generation results and learning speed. (3) Although we present various qualitative results, we only evaluated our approach quantitatively using image similarity metrics. Therefore, we cannot draw conclusions about whether our super-resolution approach improves biomarker quantification. This requires a vast amount of manual annotation labour or a reliable segmentation model for the type of OCT data used in this work. As we did not have access to these resources, we leave this evaluation to future work. (4) Hallucination is a large risk of most generative models,

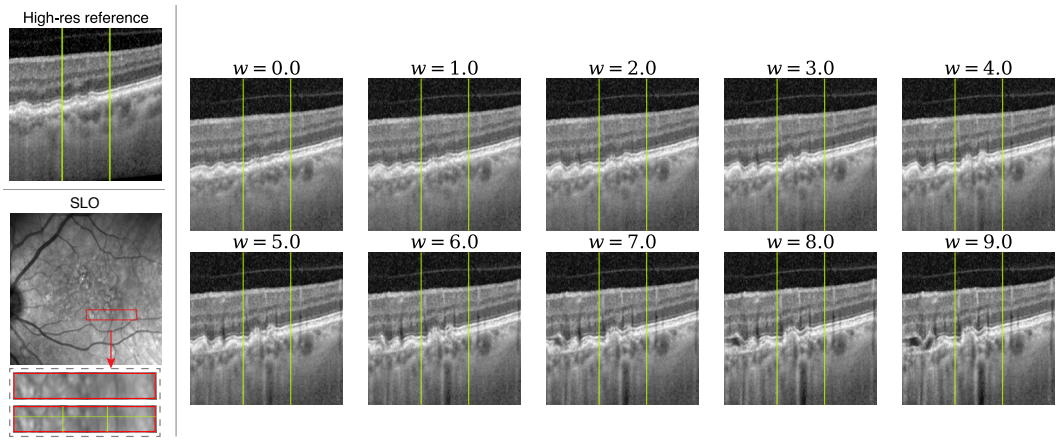


Figure 5: The effect of CFG and increasing the guidance scale w. In the top left, the high-resolution reference is shown. In the bottom left, the *en face* SLO image is shown, in which the patch location is indicated with a red box. The vertical lime lines correspond to the same physical locations throughout the figure. The horizontal lime line in the bottom left indicates the B-scan location of the shown OCT patches. The two rows on the right show the results from our *en face* conditioned diffusion model with CFG for various guidance scales. $w = 0$ is equivalent to the unconditional DDIM, $w = 1$ to DDIM_{ef} (no CFG), $w = 2$ to DDIM_{ef}, and $w > 2$ is similar to DDIM_{ef}, but with a larger guidance scale. The two drusen between the lime vertical lines seem to grow with a larger guidance scale.

including diffusion models [13]. In the context of medical imaging, generative models like ours risk the generation of non-existent lesions (leading to false positives), inflating them (leading to over-quantification), removing them (leading to false negatives), or shrinking them (leading to under-quantification). Although our *en face* conditioning mechanism may reduce hallucinations by providing more context to make well-informed generative decisions, sufficient empirical evidence of our method completely preventing this is lacking. (5) The sampling time for our diffusion model is relatively long. However, approaches exist to reduce this sampling time [25, 22].

We only explored the effect of conditioning OCT super-resolution diffusion models with near-infrared SLO images. Future work could include more *en face* modalities as conditional information, such as CFP and FAF. As images from those modalities would likely provide additional information than SLO images, we expect this could lead to more accurate super-resolution models. Other metadata, such as functional vision exam data and OCT scans from other devices or protocols, may contain even more useful information.

In our current implementation, we resize and vertically repeat the SLO image patch, enabling concatenation in the channel dimension with the OCT volume. This turns the 2D SLO image into a volume that is processed by 3D convolutions. This is computationally inefficient and the initial resizing can lead to information loss. Future work could focus on designing an architecture that more effectively leverages this multimodal data, possibly using a separate encoder for *en face* images and a cross-attention mechanism to combine the features from the different encoders.

The approach of using relatively high-resolution 2D images from a certain modality to condition diffusion models for super-resolving 3D data from another modality could potentially be applied in other medical domains. For example, using high-resolution 2D X-ray imaging as conditional information for super-resolution CT or MRI scans may be an interesting future direction.

In conclusion, we have shown the feasibility of conditioning super-resolution diffusion models to reduce anisotropy in volumetric images with additional and readily available image data, enabling well-informed generative decisions. Specifically, we showed this in the context of OCT super-resolution conditioning on *en face* images. We think this can be an important next step towards standardized high-quality OCT and other volumetric imaging, leading to more consistent measurements – obtained from either downstream manual quantifications or machine learning models – within and across datasets, studies, and clinical practices. Furthermore, our approach could facilitate the trustworthiness of generative models and their regulatory approval by mitigating the risk of hallucinations compared to uninformed super-resolution models.

Acknowledgments and Disclosure of Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116076. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. This communication reflects the author's view and neither IMI nor the European Union and EFPIA are responsible for any use that may be made of the information contained therein.

References

[1] Silke Aumann, Sabine Donner, Jörg Fischer, and Frank Müller. Optical coherence tomography (OCT): principle and technical realization. *High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics*, pages 59–85, 2019.

[2] Anne E Barañano, Pearse A Keane, Humberto Ruiz-Garcia, Alexander C Walsh, and Srinivas R Sadda. Impact of scanning density on spectral domain optical coherence tomography assessments in neovascular age-related macular degeneration. *Acta Ophthalmologica*, 90(4):e274–e280, 2012.

[3] Hrvoje Bogunovic, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, Karthik Gopinath, Amirali K Gostar, Kiwan Jeon, Zexuan Ji, Sung Ho Kang, Dara D Koozekanani, Donghuan Lu, Dustin Morley, Keshab K Parhi, Hyoung Suk Park, Abdolreza Rashno, Marinko Sarunic, Saad Shaikh, Jayanthi Sivaswamy, Ruwan Tennakoon, Shivin Yadav, Sandro De Zanet, Sebastian M Waldstein, Bianca S Gerendas, Caroline Klaver, Clara I Sánchez, and Ursula Schmidt-Erfurth. Retouch: The retinal oct fluid detection and segmentation benchmark and challenge. *IEEE Transactions on Medical Imaging*, 38:1858–1874, 8 2019.

[4] Vineeta Das, Samarendra Dandapat, and Prabin Kumar Bora. Unsupervised super-resolution of oct images using generative adversarial network for improved age-related macular degeneration diagnosis. *IEEE Sensors Journal*, 20(15):8746–8756, 2020.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[6] Sina Farsiu, Stephanie J. Chiu, Rachelle V. O'Connell, Francisco A. Folgar, Eric Yuan, Joseph A. Izatt, and Cynthia A. Toth. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.

[7] Robert P. Finger, Steffen Schmitz-Valckenberg, Matthias Schmid, Gary S. Rubin, Hannah Dunbar, Adnan Tufail, David P. Crabb, Alison Binns, Clara I. Sánchez, Philippe Margaron, Guillaume Normand, Mary K. Durbin, Ulrich F. O. Luhmann, Parisa Zamiri, Jose Cunha-Vaz, Friedrich Asmus, Frank G. Holz, and on behalf of the MACUSTAR consortium. Macustar: Development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. *Ophthalmologica*, 241:61–72, 8 2019.

[8] Davide Garzone, Jan Henrik Terheyden, Olivier Morelle, Maximilian WM Wintergerst, Marlene Saßmannshausen, Steffen Schmitz-Valckenberg, Maximilian Pfau, Sarah Thiele, Stephen Poor, Sergio Leal, et al. Comparability of automated drusen volume measurements in age-related macular degeneration: a macustar study report. *Scientific reports*, 12(1):21911, 2022.

[9] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[12] Yongqiang Huang, Zexin Lu, Zhimin Shao, Maosong Ran, Jiliu Zhou, Leyuan Fang, and Yi Zhang. Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network. *Optics express*, 27(9):12289–12307, 2019.

[13] Seunghoi Kim, Chen Jin, Tom Diethe, Matteo Figini, Henry FJ Tregidgo, Asher Mullokandov, Philip Teare, and Daniel C Alexander. Tackling structural hallucination in image translation with local diffusion. *arXiv preprint arXiv:2404.05980*, 2024.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[15] Jiawei Li, Jae Chul Koh, and Won-Sook Lee. Hrinet: alternative supervision network for high-resolution ct image interpolation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1916–1920. IEEE, 2020.

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[17] Martijn J. Mulder, Max C. Keuken, Pierre-Louis Bazin, Anneke Alkemade, and Birte U. Forstmann. Size and shape matter: The impact of voxel geometry on the identification of small nuclei. *PLOS ONE*, 14(4):1–19, 04 2019.

[18] Muneeswar Gupta Nittala, R Konduru, H Ruiz-Garcia, and SR Sadda. Effect of oct volume scan density on thickness measurements in diabetic macular edema. *Eye*, 25(10):1347–1355, 2011.

[19] Deniz Oncel, Navid Manafi, Muneeswar Gupta Nittala, Swetha Bindu Velaga, Dwight Stambolian, Margaret A Pericak-Vance, Jonathan L Haines, and Srinivas R Sadda. Effect of oct b-scan density on sensitivity for detection of intraretinal hyperreflective foci in eyes with age-related macular degeneration. *Current eye research*, 47(9):1294–1299, 2022.

[20] Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*, 2023.

[21] Bin Qiu, Yunfei You, Zhiyu Huang, Xiangxi Meng, Zhe Jiang, Chuanqing Zhou, Gangjun Liu, Kun Yang, Qiushi Ren, and Yanye Lu. N2nsr-oct: Simultaneous denoising and super-resolution in optical coherence tomography images using semisupervised deep learning. *Journal of biophotonics*, 14(1):e202000282, 2021.

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

[23] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015.

[24] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[25] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[26] Jörg Sander, Bob D de Vos, and Ivana Išgum. Autoencoding low-resolution mri for semantically smooth interpolation of anisotropic mri. *Medical image analysis*, 78:102393, 2022.

[27] Ursula Schmidt-Erfurth, Gregor S Reiter, Sophie Riedl, Philipp Seeböck, Wolf-Dieter Vogl, Barbara A Blodi, Amitha Domalpally, Amani Fawzi, Yali Jia, David Sarraf, et al. AI-based monitoring of retinal fluid in disease activity and under therapy. *Progress in retinal and eye research*, 86:100972, 2022.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[29] Jan H Terheyden, Frank G Holz, Steffen Schmitz-Valckenberg, Anna Lüning, Matthias Schmid, Gary S Rubin, Hannah Dunbar, Adnan Tufail, David P Crabb, Alison Binns, et al. Clinical study protocol for a low-interventional study in intermediate age-related macular degeneration developing novel clinical endpoints for interventional clinical trials with a regulatory and patient access intention—macustar. *Trials*, 21:1–11, 2020.

[30] SB Velaga, MG Nittala, RK Konduru, F Heussen, PA Keane, and SR Sadda. Impact of optical coherence tomography scanning density on quantitative analyses in neovascular age-related macular degeneration. *Eye*, 31(1):53–61, 2017.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[32] Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Yue Huang, Qinji Yu, Sifan Song, Xinxing Xu, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023.

[33] Ying Xu, Bryan M Williams, Baidaa Al-Bander, Zheping Yan, Yao-chun Shen, and Yalin Zheng. Improving the resolution of retinal oct with deep learning. In *Medical Image Understanding and Analysis: 22nd Conference, MIUA 2018, Southampton, UK, July 9-11, 2018, Proceedings 22*, pages 325–332. Springer, 2018.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

A Appendix

A.1 Additional dataset and pre-processing details

For the training and validation set combined, the number of B-scans per OCT volume varied, with 1 volume containing 237 B-scans, 927 volumes containing 241 B-scans, and 29 volumes containing 512 B-scans. The B-scan sizes ranged from 512 $\times$ 496 pixels to 1536 $\times$ 496 pixels, and the physical OCT volume size ranged from 2.9 $\times$ 1.9 $\times$ 2.9 mm^3 to 9.2 $\times$ 1.9 $\times$ 7.7 mm^3. For the test set, all OCT volumes contained 241 B-scans. The B-scan sizes ranged from 512 $\times$ 496 pixels to 1536 $\times$ 496 pixels, and all OCT volumes had a physical size of approximately 8.8 $\times$ 1.9 $\times$ 7.3 mm^3.

Next to the previously described set of OCT volumes, we used the near-infrared confocal SLO images, which Heidelberg Spectralis devices acquire alongside the OCT, as *en face* modality for our conditional diffusion models. The SLO image was registered to the OCT volume according to the physical linkage information between these two images that was provided by the camera software. We subsequently cropped and resized these SLO images to the same width and height as, respectively, the width and depth from their corresponding OCT volumes.

For some OCT volumes in the dataset, we observed that incidentally adjacent B-scans were not correctly aligned vertically. Therefore, during sampling, B-scans were registered vertically using a grid search for the vertical translation amount and MSE as a cost function, based on a flattened representation of the B-scans (collapsed into columns by averaging over the x-axis).

A.2 Overlapping patches and inpainting

During sampling of full volumes, we use overlapping patches, facilitated through inpainting with RePaint [16]. Inpainting is the task of filling in new content in a specific part of an image, which can be defined by a binary mask. We refer to the image part that needs to be filled in as “unknown” and the other part as “known”. During each timestep t in the sampling process, RePaint combines the “known” image part from the input image, which has been noised to the appropriate noise level of timestep $t-1$, with the “unknown” part from the denoised image $\mathbf{x}_{t-1}$ (see the bottom part of Fig. A.1). When we generate a complete volume using this patch overlapping strategy, the “known" region is defined as the area previously generated from an adjacent patch, combined with slices from the original low-resolution image volume (see the top part of Fig. A.1). Besides, the patch size during sampling is larger than the one used during training, as we empirically found this to improve the coherence and fidelity of the generated volumes.

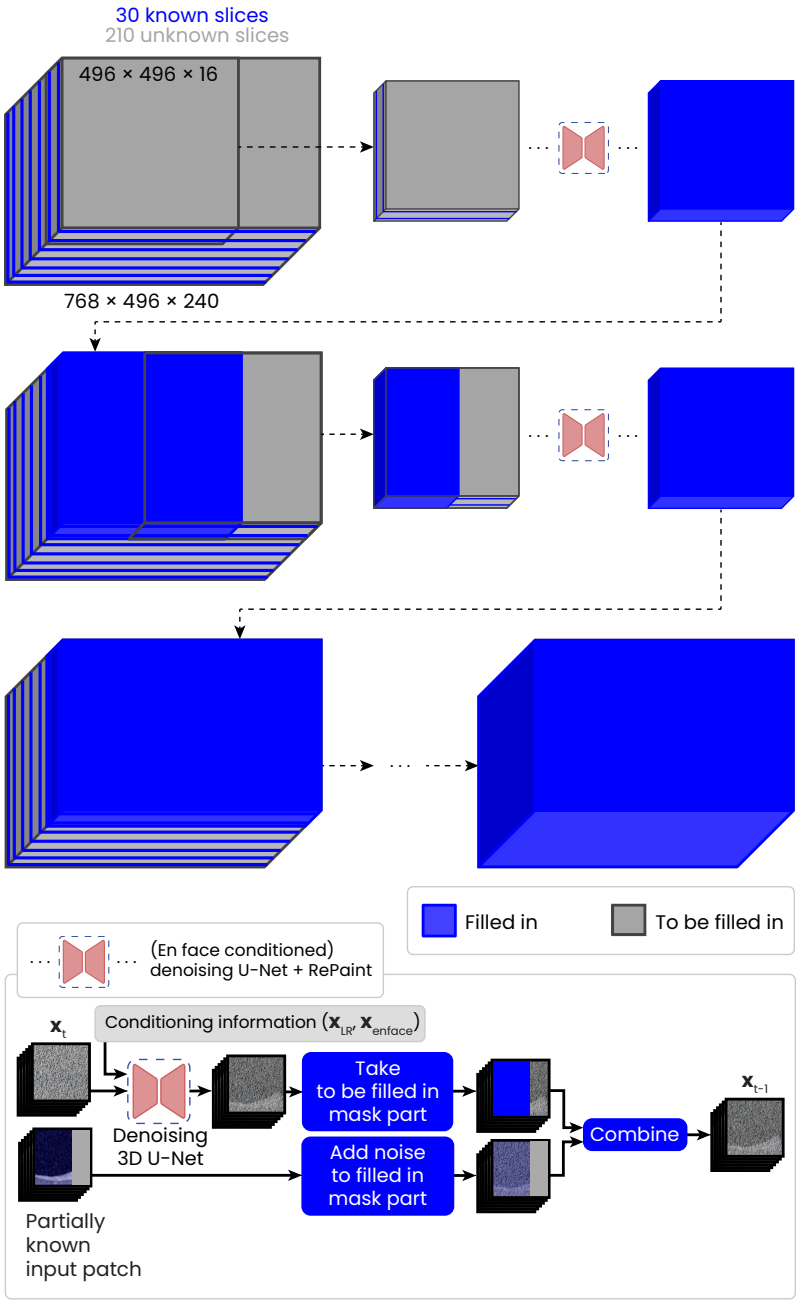


Figure A.1: Overview of our overlapping strategy facilitated through RePaint [16].

A.3 Additional results

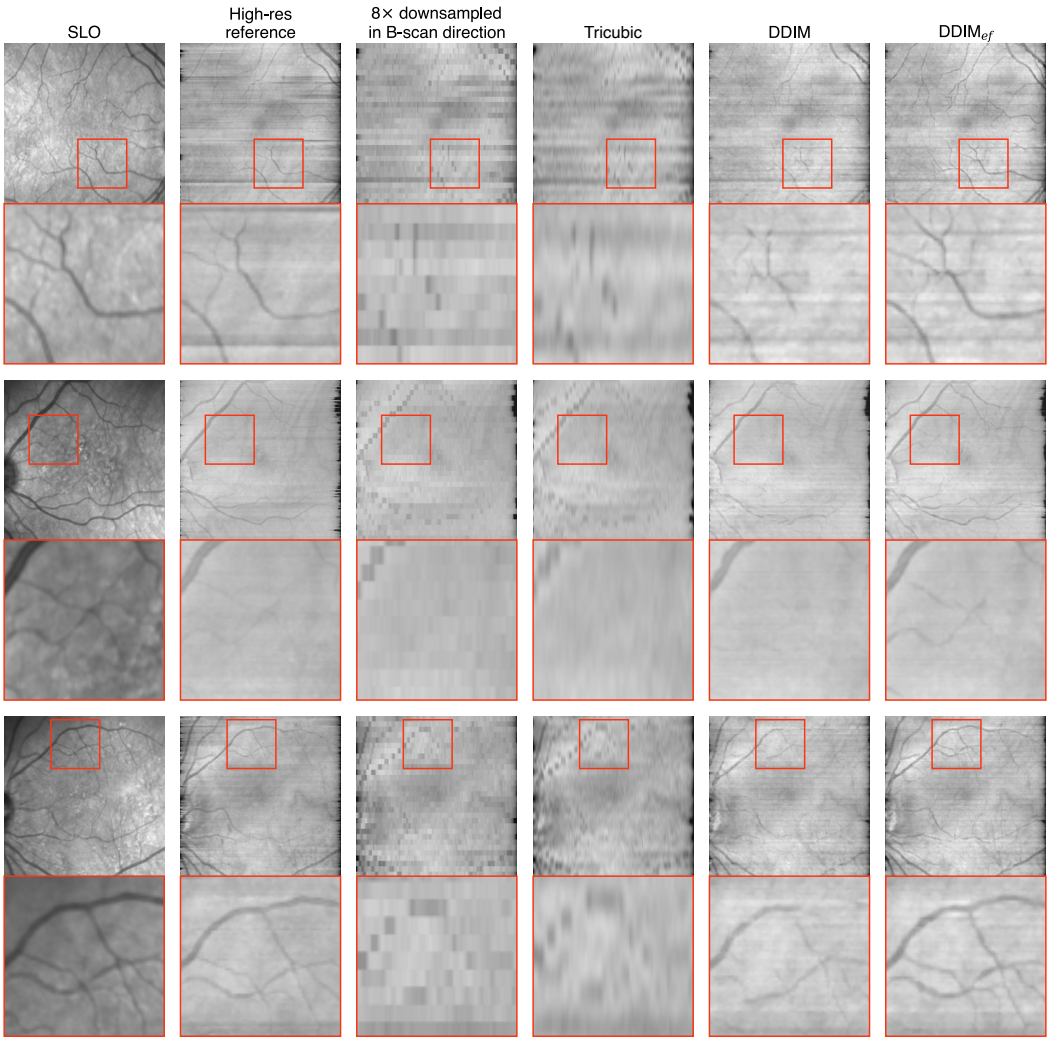
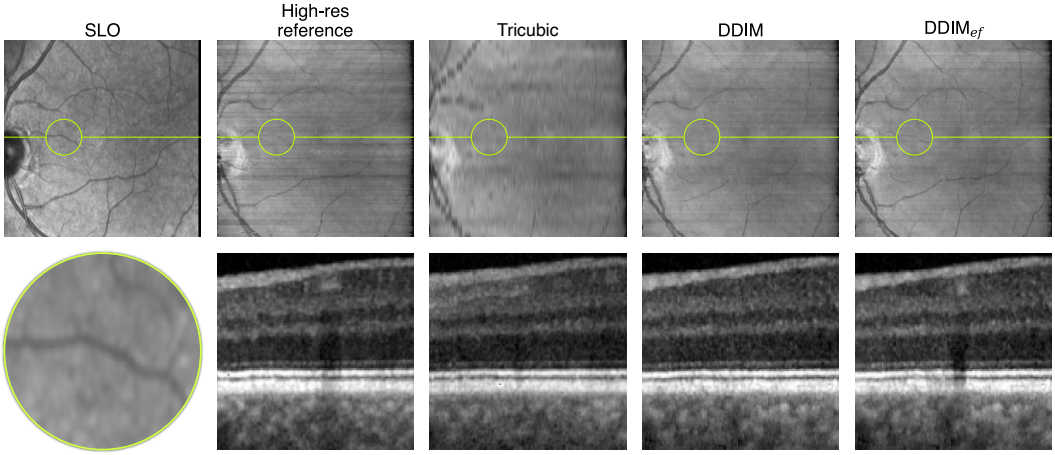
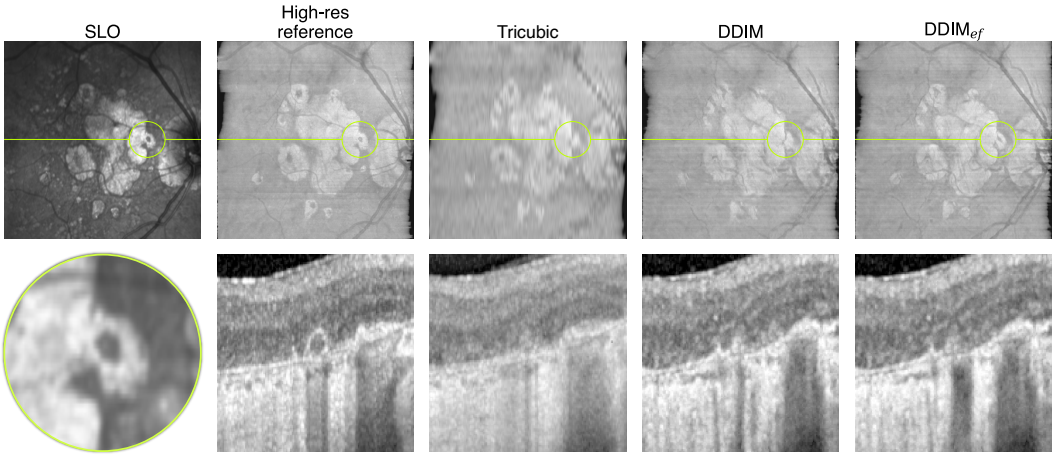


Figure A.2: Additional examples of *en face* projections of the OCT volumes generated using tricubic interpolation, the unconditional diffusion model DDIM, and our proposed *en face* conditioned diffusion model DDIM$_{\text{ef}}$ (presentation similar to Fig. 3). These three projections are shown in the last three columns and were all generated from the 8 $\times$ downsampled (in the B-scan direction) volume as input, which is shown in the third column. The first and second rows show the corresponding high-resolution (high-res) reference and scanning laser ophthalmoscopy (SLO) image, respectively. A separate example from a different test set patient is shown in each row. The top images in each row show the full image. Zoomed-in versions of the image patches (red boxes) are shown at the bottom of each row.



(a)



(b)

Figure A.3: Additional examples showing the effect of using DDIM_{ef}, compared to leaving out the *en face* information (DDIM), and tricubic interpolation (presentation similar to Fig. 4). In the top rows, the scanning laser ophthalmoscopy (SLO) image is shown on the left, followed by the *en face* projections of each OCT volume. In the bottom row, zoomed-in patches are shown. The patch locations are indicated in the top row with a lime circle. The lime horizontal line corresponds to the B-scan location from the shown patches. (a) A blood vessel is only reconstructed by DDIM_{ef}. This vessel is also visible in the SLO image. (b) The hypertransmission pattern seems to be best reconstructed by DDIM_{ef}. The *en face* location of the hypertransmission area is also visible in the SLO image.

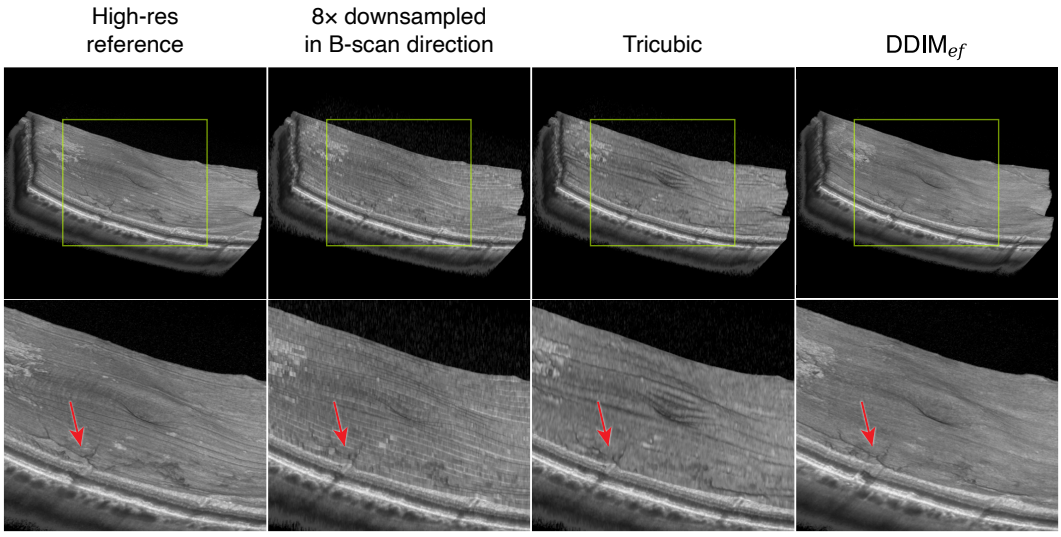


Figure A.4: 3D renders of full OCT volumes, depicted for the high-resolution (high-res) reference, downsampled volume, tricubic interpolated volume, and our proposed method DDIM$_{\text{ef}}$. The bottom row shows zoomed-in versions of the renders in the top row. The lime squares in the top row indicate the zoomed-in area. The red arrows point to a vessel on the inner retina.

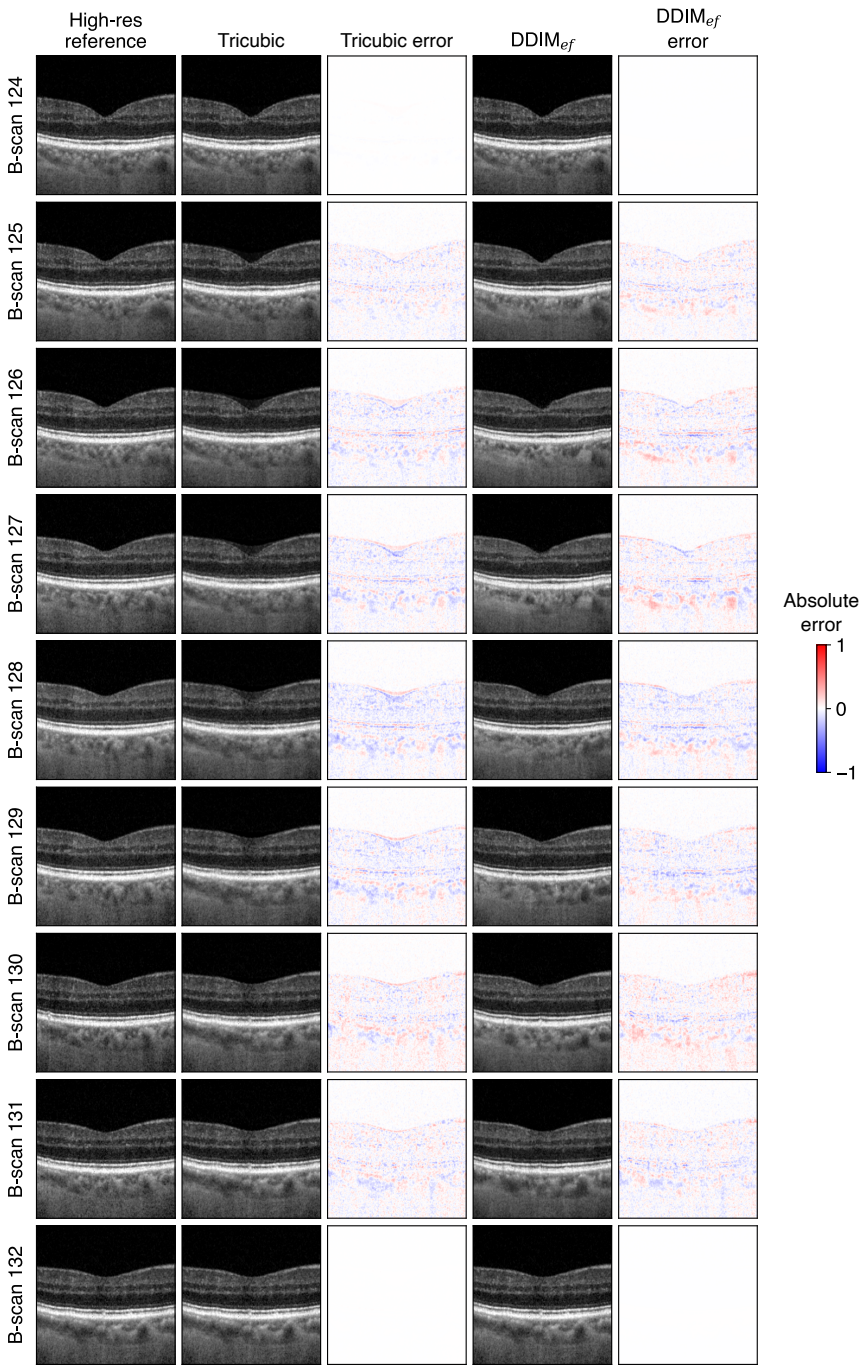


Figure A.5: An example showing a sequence of B-scans, illustrating that the volume generated with tricubic interpolation is more smoothed than the volume generated by our diffusion model DDIM$_{\text{ef}}$, which is much sharper. This is most evident around slice 128. The first and last shown slices were already present in the low-resolution OCT volume. The third and fifth columns indicate the absolute error.

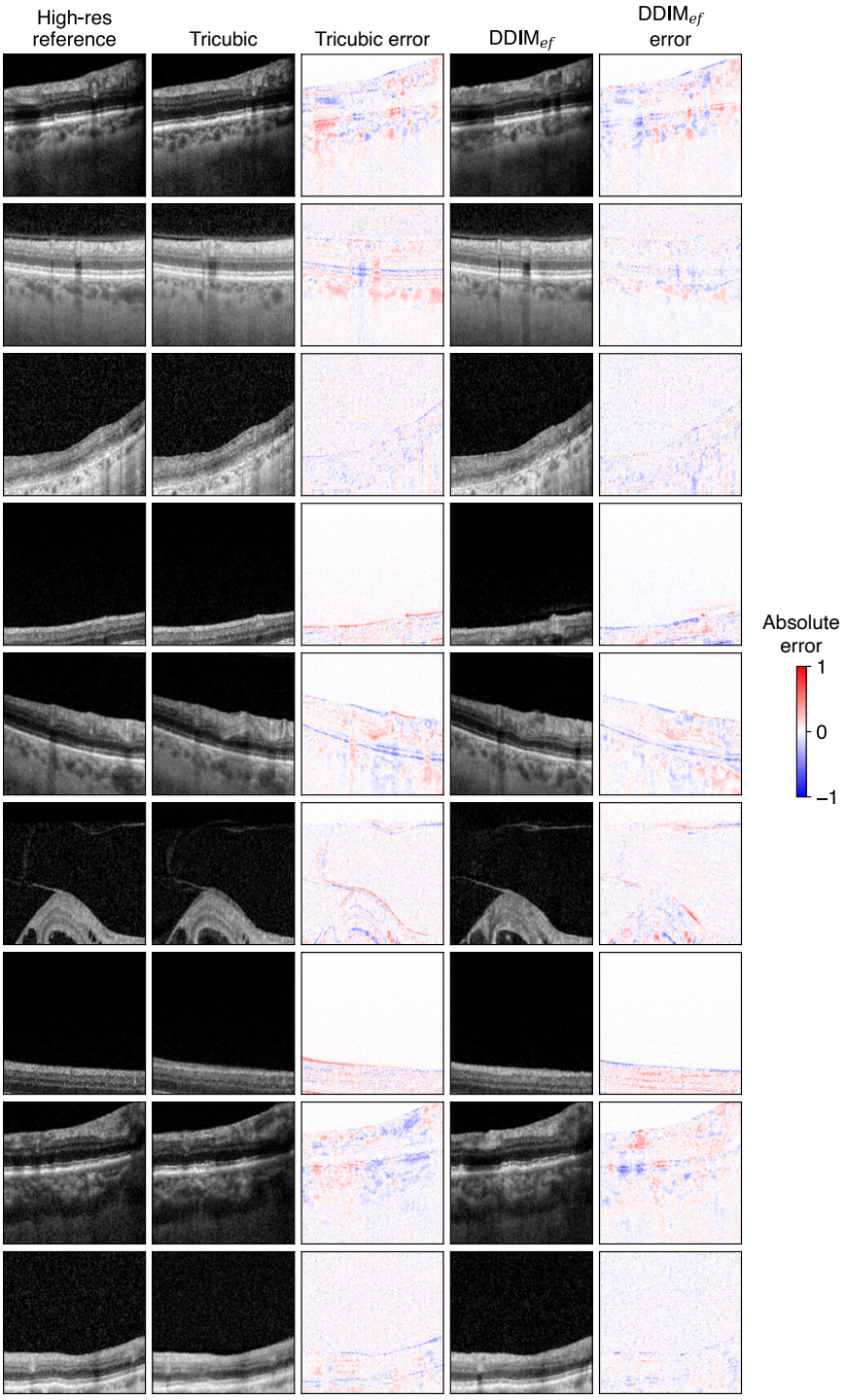


Figure A.6: Randomly picked 256×256 B-scan patches from "unknown" slices in the test set, depicting the origins of absolute errors for tricubic interpolation and DDIM$_{\text{ef}}$. The third and fifth columns indicate the absolute error.

A.4 LPIPS for different anatomical planes

We found that our diffusion models outperformed interpolation in terms of $\text{LPIPS}_{\text{cor}}$, $\text{LPIPS}_{\text{sag}}$, and $\text{LPIPS}_{\text{efproj}}$, but not in terms of $\text{LPIPS}_{\text{axi}}$. This finding is in line with the visual observations we make when manually inspecting slices from these anatomical planes. This is illustrated in Fig. A.7, in which DDIM_{ef} shows higher visual similarity with the high-resolution reference than the image upsampled with tricubic interpolation for the coronal plane, sagittal plane, and the *en face* projection. For the axial plane, however, we think this is less evident. The LPIPS values, which are also shown in Fig. A.7 for the shown slices, correlate well with these visual observations.

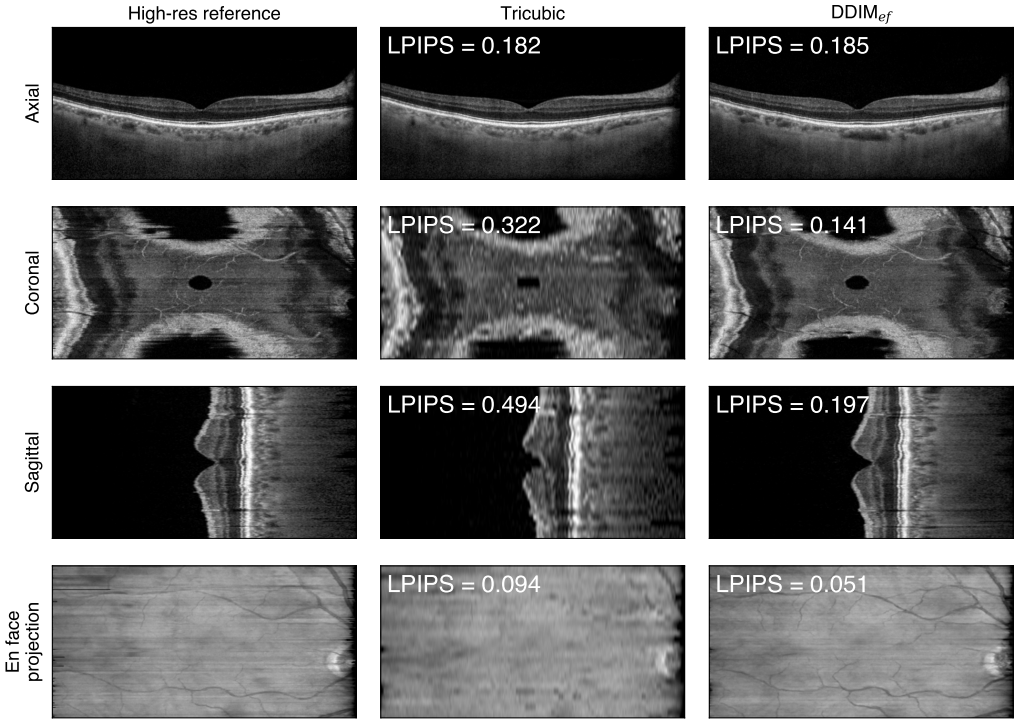
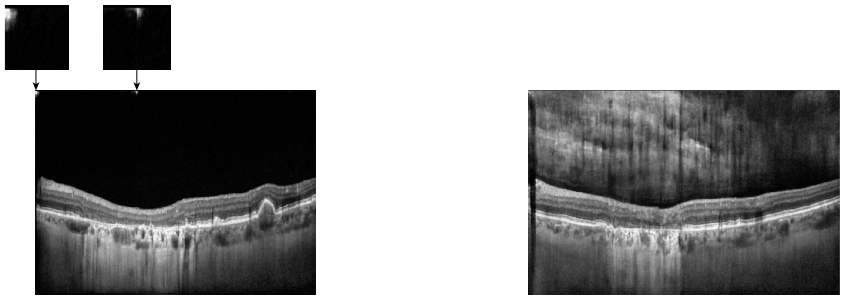


Figure A.7: Examples of slices from the three orthogonal anatomical planes and the *en face* projection image. All shown images originate from the same OCT volume. For tricubic interpolation and DDIM_{ef}, Learned Perceptual Image Patch Similarity (LPIPS) values, computed only using the depicted slices, are shown in the top left corner of each image. For the anatomical planes, the middle slices are shown. For the axial (B-scan) slice, this corresponds to a slice exactly in the middle of two slices that were also present in the low-resolution volume (i.e., , the slice was as far away from a "known" slice as possible, which is generally a slice that is more difficult to generate accurately than a slice that is closer to a "known" slice).

A.5 Generated imaging artifacts

We found that a number of imaging artifacts can occur when generating volumes using our DDIM approach (see A.8). The occurrence frequency depends on the artifact type. For example, small groups of white pixels are sometimes generated near the top of B-scans (see an example in Fig. A.8a). If they are present, they always seem to be located in the top left of the sampling patches. Furthermore, sometimes very large, bright areas are generated in the vitreous body (see an example in Fig. A.8b). This mainly seems to occur when two adjacent B-scans in the low-resolution volume were not registered well.



(a) (b)

Figure A.8: Artifact types that sometimes occur in volumes generated by our diffusion models. These artifacts were neither present in the low-resolution input nor the high-resolution reference. (a) Small groups of white pixels in the top left of sampling patches. (b) White areas in the vitreous body. This artifact type mainly seems to occur when two adjacent B-scans in the low-resolution volume were not registered well.

A.6 Ethical approval

Ethical approval of the clinical study data has been obtained, as described in previous works describing the MACUSTAR study [29].

A.7 MACUSTAR consortium members

H. Agostini, I. D. Aires, L. Altay, R. Atia, F. Bandello, P. G. Basile, J. Batuca, C. Behning, M. Belmouhand, M. Berger, A. Binns, C. J. F. Boon, M. Böttger, J. E. Brazier, C. Carapezzi, J. Carlton, A. Carneiro, A. Charil, R. Coimbra, D. Cosette, M. Cozzi, D. P. Crabb, J. Cunha-Vaz, C. Dahlke, H. Dunbar, R. P. Finger, E. Fletcher, M. Gutfleisch, F. Hartgers, B. Higgins, J. Hildebrandt, E. Höck, R. Hogg, F. G. Holz, C. B. Hoyng, A. Kilani, J. Krätzschmar, L. Kühlewein, M. Larsen, S. Leal, Y. T. E. Lechanteur, D. Lu, U. F. O. Luhmann, A. Lüning, N. Manivannan, I. Marques, C. Martinho, A. Miliu, K. P. Moll, Z. Mulyukov, M. Paques, B. Parodi, M. Parravano, S. Penas, T. Peters, T. Peto, S. Priglinger, R. Ramamirtham, R. Ribeiro, D. Rowen, G. S. Rubin, J. Sahel, C. Sánchez, O. Sander, M. Saßmannshausen, M. Schmid, S. Schmitz-Valckenberg, J. Siedlecki, R. Silva, E. Souied, G. Staurenghi, J. Tavares, D. J. Taylor, J. H. Terheyden, A. Tufail, P. Valmaggia, M. Varano, A. Wolf, N. Zakaria