OCTDiff: Bridged Diffusion Model for Portable OCT Super-Resolution and Enhancement

Ye Tian[1], **Angela McCarthy**[3], **Gabriel Gomide**[3], **Nancy Liddle**[2], **Jedrzej Golebka**[3],
Royce W.S. Chen[3], Jeffrey M. Liebmann[3], Kaveri A. Thakoor[1,2,3]

[1]Department of Biomedical Engineering, Columbia University, New York, NY, USA
[2]Department of Computer Science, Columbia University, New York, NY, USA
[3]Department of Ophthalmology, Columbia University Irving Medical Center, New York, NY, USA
{yt2793, k.thakoor}@columbia.edu

Abstract

Medical imaging super-resolution is critical for improving diagnostic utility and reducing costs, particularly for low-cost modalities such as portable Optical Coherence Tomography (OCT). We propose OCTDiff, a bridged diffusion model designed to enhance image resolution and quality from portable OCT devices. Our image-to-image diffusion framework addresses key challenges in the conditional generation process of denoising diffusion probabilistic models (DDPMs). We introduce Adaptive Noise Aggregation (ANA), a novel module to improve denoising dynamics within the reverse diffusion process. Additionally, we integrate Multi-Scale Cross-Attention (MSCA) into the U-Net backbone to capture local dependencies across spatial resolutions. To address overfitting on small clinical datasets and to preserve fine structural details essential for retinal diagnostics, we design a customized loss function guided by clinical quality scores. OCTDiff outperforms convolutional baselines and standard DDPMs, achieving state-of-the-art performance on clinical portable OCT datasets. Our model and its downstream applications have the potential to generalize to other medical imaging modalities and revolutionize the current workflow of ophthalmic diagnostics. The code is available at https://github.com/AI4VSLab/OCTDiff.

1 Introduction

Medical image analysis has been transformed by recent deep learning advances, including disease classification [1, 2], tissue and cellular segmentation [3], and contrast enhancement [4]. However, most existing models are trained on data from high-end imaging systems, making them less effective and even unusable in low-resource or point-of-care settings. One critical domain with this limitation is ophthalmology, where high-quality Optical Coherence Tomography (OCT) systems can cost over 50,000 dollars and weigh more than 50 pounds, significantly reducing their accessibility in underdeveloped regions.

Convolutional neural networks (CNNs) and generative adversarial networks (GANs) have been widely applied for OCT image enhancement tasks including super-resolution (SR) and denoising. For instance, an early approach [5] built on ESRGAN [6] and MedGAN [7] demonstrated promising visual improvements on simulated low-resolution OCT images. However, it suffered from mode collapse and generated unwanted artifacts; hence, its generalizability to real-world clinical scans remains untested. CNN-based super-resolution methods have also been applied on other medical imaging modalities such as computed tomographpy (CT) and magnetic resonance imaging (MRI) [8, 9]. Though quantitative performance improves, these models tend to hallucinate fine details and

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

fail to preserve delicate anatomical structures, which are essential for accurate diagnosis. Medical visual tasks for diagnostic assistance such as anatomical segmentation and motion correction have achieved clinically-usable performance [10, 11], but SR remains challenging due to the inherent low contrast of low-resolution images and the scarcity of large-scale patient datasets with low-resolution and high-resolution pairs, despite well-established SR benchmarks on natural images [12].

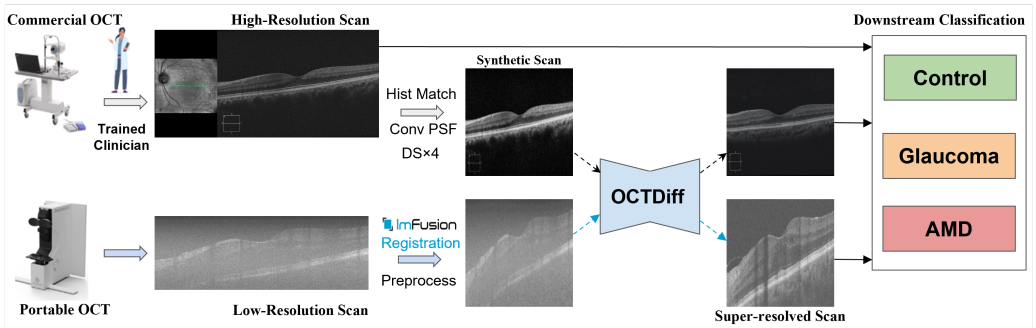

Figure 1: Pipeline of our study. The clinical low-resolution dataset was captured with a portable OCT device and then underwent necessary filtering, preprocessing, and registration. The synthetic dataset [5] was obtained by matching the histogram of the commercial OCT data to that of the portable OCT data, convolving the resulting data with the point spread function (PSF) of the portable OCT, and then downsampling the data by a factor of 4 [6]. OCTDiff super-resolves both datasets and generates high-resolution outputs for downstream classification of ophthalmic diseases such as glaucoma and age related macular degeneration (AMD).

Diffusion-based generative modeling, i.e. denoising diffusion probabilistic models (DDPMs) [13], has demonstrated superior performance over convolutional models and GANs in terms of both generative quality and training stability [14, 15]. DDPM variants such as latent diffusion models (LDMs) [16] and conditional DDPMs (CDMs) [17] extend the power of DDPMs by incorporating guided conditioning to generate more controlled outputs. However, they are still not guaranteed to reliably translate images between different domains when conditioning on a target image. A more recent approach, the bridged Brownian diffusion model (BBDM) [18], emphasizes conditioning by directly using the reference image as the initial point in the reverse diffusion process. This architecture is particularly well-suited for image-to-image tasks such as style transfer and semantic synthesis. Nevertheless, its potential for super-resolution and its effectiveness when training on medical images, which are small in dataset size and large in spatial resolution [19], remains relatively unexplored.

To address these limitations, we propose OCTDiff, an image-to-image super-resolution conditional diffusion model for portable OCT enhancement. OCTDiff builds upon the bridged framework originated from BBDM. We propose two novel components: (1) an Adaptive Noise Aggregation (ANA) algorithm that stabilizes the reverse denoising trajectory by aggregating noise predictions from previous time steps in the reverse process; and (2) a Multi-scale Cross-Attention (MSCA) UNet backbone that enables cross-resolution feature interaction between encoder and decoder to better capture both global anatomy and fine retinal details. (3) During model training, we also introduce a custom loss function with modulation from a clinical quality score to guide the model toward perceptually and diagnostically meaningful outputs.

In summary, our main contributions are:

- We propose OCTDiff, which significantly outperforms baseline methods in both quantitative metrics and qualitative structural fidelity on real-world OCT datasets. The computational efficiency and training speed are also preserved, as demonstrated in Section 4.1.
- We are the first to address the semantic misalignment issue [20] in conditional DDPMs through temporal fusion across denoising steps, enabling stronger conditioning guidance throughout the generation process. This strategy is uniquely feasible within the bridged diffusion framework, where the conditioning image directly anchors the reverse process.
- Clinically, OCTDiff pioneers a new direction at the intersection of AI and healthcare, tailored for ultra-low-resolution images from portable devices. Its utility in downstream classification tasks (Section 4.3) shows OCTDiff's potential to deliver affordable and reliable vision care to

under-served and remote populations, and to generalize to other resource-constrained medical imaging settings.

2 Related Work

Heuristic Optimization in Diffusion Models Early improvements in diffusion models relied heavily on heuristic strategies for loss weighting, noise scheduling, and sampling. Hybrid loss designs [21, 22, 23, 24] combine pixel-wise noise prediction with latent-space objectives to balance reconstruction accuracy and perceptual realism. For noise scheduling, linear [25] or cosine β_t schedules are commonly used, while more recent works [26, 27] propose handcrafted tuning based on empirical settings that better align with DDPM's denoising capacity across timesteps. Additionally, timestep reweighting prioritizes intermediate steps where gradients are more stable [21, 28, 29, 30]. These heuristic strategies improve sample diversity and convergence without modifying model architecture, but they are inherently static and task-agnostic, often relying on fixed schedules and an assumption that certain timesteps are more important than others. In contrast, our Adaptive Noise Aggregation (ANA) module in OCTDiff dynamically aggregates multiple denoising predictions across timesteps, assigning different noise schedules to different images. This temporal fusion captures complementary information from various noise levels that represent multi-scale features, which is conceptually similar to ensemble learning. ANA is only realizable within a bridged diffusion framework, being particularly effective for conditional tasks such as super-resolution and reconstruction for degraded OCT images.

Learnable noise modulation MuLAN [31] learns per-sample noise schedules by predicting the optimal noise scale σ for each input, while other methods [32, 33] use auxiliary networks or learned noise embeddings to refine denoising behavior. ANA also aims to improve signal-noise alignment but achieves this through temporal fusion rather than direct modulation. Unlike MuLAN's learned noise schedules, ANA adapts noise dynamically per image without introducing additional supervision, making it well-suited for resource-constrained applications. While ANA may sacrifice flexibility in handling highly variable noise patterns compared to MuLAN, it offers a simpler and more efficient solution, particularly for OCT scans that require consistent, low-complexity adaptations.

Attention and Diffusion Recent works increasingly integrate attention mechanisms into diffusion models. Transformer-based architectures such as latent diffusion [16], ImageCraft [34], and CDM [35] incorporate self- and cross-attention to enhance spatial coherence. Other designs introduce hierarchical [36], spatially-aware [37], or multi-scale attention [38, 39] to improve structure preservation while maintaining efficiency. Some models further combine attention with guidance signals (e.g., text, edge maps) for stronger control [40, 41]. While prior multi-scale works stack features within encoder or decoder branches or process them in parallel at a single resolution, our MSCA introduces explicit cross-attention between encoder and decoder features at different scales, enabling context-aware guidance across resolutions.

Temporal Ensembling and Recurrent Denoising We introduce ANA as a regularizing prior across timesteps, inspired by previous aggregation algorithms. For example, temporal ensembling reduces prediction variance and enhances consistency in semi-supervised learning [42], and recurrent denoising benefits from multi-step integration to avoid error explosion [26]. The idea of combining global and local features in medical images also resembles multi-scale fusion of frequency components via weighted aggregation [43]. Our work is also distinctive from differential-equation–based methods [44] because ANA is discrete-time, controllable (with attenuation α) and operates during inference (not training). Also rather than modifying the noise sampler like DPM-Solver [45], ANA assumes a fixed noise schedule (e.g., linear or cosine) and aggregates the previously-sampled noise vectors.

3 Method

3.1 Adaptive Noise Aggregation

Our OCTDiff builds upon bridged diffusion [18] that learns the translation between two image domains directly through a bidirectional diffusion process. We propose Adaptive Noise Aggregation (ANA) strategy to improve the stability of the reverse denoising process. Noise predictions at different time steps t encode complementary information particularly for high-frequency retinal details. Aggregating different noise levels basically leverages different image scales' information. ANA with adaptive weights enhances fine structure reconstruction, especially when the input is severely degraded as in portable OCT scans.

The reverse process starts from the high-resolution condition $\hat{x}_T = y$, and iteratively estimates intermediate latent states $\hat{x}_t$ to recover a clean image $\hat{x}_0$ over T denoising steps. At each step t, the model defines the reverse transition as:

$$p_\theta(x_t \mid x_{t+1}, y) \sim \mathcal{N}\left(\mu_\theta(x_{t+1}, t, y), \Sigma_\theta(x_{t+1}, t, y)\right) \tag{1}$$

where p_θ is a learned Gaussian distribution with parameters predicted by the U-Net backbone. The noise component is estimated as $\hat{\epsilon}_t = \varepsilon_\theta(x_{t+1}, t, y)$, where ε_θ is the noise prediction network. This predicted noise $\hat{\epsilon}_t$ is then used to reconstruct an intermediate clean estimate $\hat{x}_0^t$, which subsequently informs the mean function $\mu_\theta(x_{t+1}, t, y)$ in the reverse transition.

To improve the robustness of the reverse process, our ANA algorithm does not rely solely on the current noise prediction. Instead, it adaptively aggregates noise predictions from all previous time steps in the reverse process $\{\hat{\epsilon}_\tau\}_{\tau=t}^{T-1}$ using exponential decay. This temporal fusion yields a more stable and informative estimate $\bar{\epsilon}_t$. The aggregation is formally defined as:

$$\bar{\epsilon}_t = \frac{1}{Z_t} \sum_{\tau=t}^{T-1} \exp(-\alpha(\tau - t)) \cdot \hat{\epsilon}_\tau, \tag{2}$$

where the weight function $w(\tau, t) = \exp(-\alpha(\tau - t))$ introduces a time-decay prior, emphasizing predictions closer to step t while still leveraging long-range information. The scalar $\alpha > 0$ controls how fast the weight decays over time. We discuss the impact of different α values in an ablation study (Section 4.2). The normalization term Z_t ensures the aggregated weights sum to 1: $\sum_{\tau=t}^{T-1} \frac{w(\tau,t)}{Z_t} = 1$. This updated ϵ_t forms soft temporal fusion of noise estimates.

An intuitive explanation for ANA is the exponential moving average (EMA) [46]. We adapt the EMA principle to the spatially-structured noise tensor space, making ANA a temporally-aware ensemble over latent signals within diffusion. In OCTDiff, the forward process generates a sequence of interpolated images ("bridges") between the low-resolution input and high-resolution target. This bridging mechanism ensures all intermediate steps are structurally meaningful and semantically anchored. Therefore, aggregating noise predictions across time is more stable and informative here. On the contrary, in standard DDPMs, reverse steps are inherently noisy in early steps and often sensitive to prediction errors, which makes temporal aggregation less stable.

The proposed ANA strategy enables the model to leverage not only the current noise prediction but also aggregated predictions across future steps. This stabilizes the reverse trajectory and enhances final image quality. The ANA algorithm is summarized in Algorithm 1.

Algorithm 1 Adaptive Noise Aggregation (ANA)

Require: High-resolution input $\hat{x}_T = y$, total steps T
Ensure: Super-resolved output $\hat{x}_0$

```
for t = T−1 to 0 do
    ε̂_t ← ε_θ(x̂_{t+1}, t)
    if t > T/2 then
        ε̂_t ← Refine(ε̂_t, ∇L_denoise)
    end if
    ε̄_t ← (1/Z_t) Σ_{τ=t}^{T−1} exp(−α(τ − t)) · ε̂_τ
    x̂_0^t ← Reconstruct(x̂_{t+1}, ε̄_t, t)
    x̂_t ← μ_θ(x̂_{t+1}, x̂_0^t, t)
end for
return x̂_0
```

Refinement Module: To improve robustness at early stages ($t > T/2$), we apply a gradient-based refinement:

$$\hat{\epsilon}_t \leftarrow \hat{\epsilon}_t - \eta \cdot \nabla_{\hat{\epsilon}_t} \mathcal{L}_{\text{denoise}},$$

where η is a small step size and $\mathcal{L}_{\text{denoise}}$ denotes a pixel-level loss against a pseudo-ground truth.

Reconstruction Module: Given $\hat{x}_{t+1}$ and the aggregated noise $\bar{\epsilon}_t$, we use the DDIM [47] inversion rule:

$$\hat{x}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\hat{x}_{t+1} - \sqrt{1 - \bar{\alpha}_t} \cdot \bar{\epsilon}_t\right),$$

to approximate the clean image at step 0. Then $\hat{x}_0^t$ is used to compute the posterior mean $\hat{x}_t$ as in the DDPM/DDIM formulation.

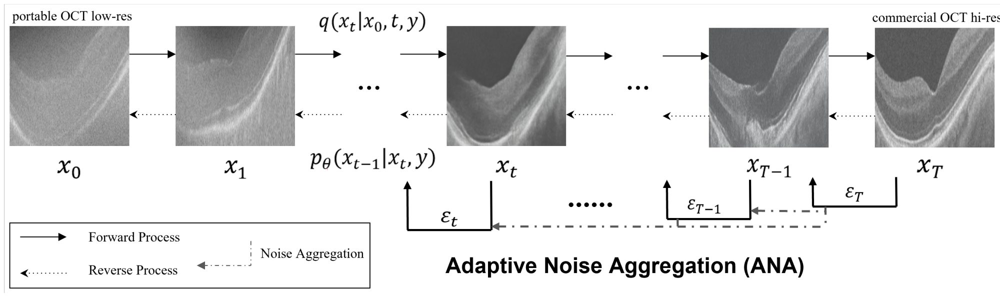

Figure 2: The Adaptive Noise Aggregation (ANA) process in reverse diffusion.

3.2 Multi-Scale Cross Attention

We implement multi-scale cross-attention (MSCA) in the UNet backbone of our OCTDiff model. The MSCA enables encoder features at each scale to attend to decoder features at different scales. This cross-scale interaction is particularly important for our super-resolution task on OCT images. Medical image scans such as OCT often contain crucial diagnostic patterns in small, localized patches observed at different scales. For example, at a scale of 32×32, only the vessels may be visible, while at 128×128, the scan may show larger structures like the retina and tissue. The MSCA helps preserve fine retinal structures while capturing global context in OCT.

For each encoder query Q_{enc}^s at scale $s \in \mathcal{S}_{\text{enc}}$, the attention is computed with key-value pairs $(K_{\text{dec}}^{s'}, V_{\text{dec}}^{s'})$ from different scales $s' \in \mathcal{S}_{\text{dec}}$ where $s' \neq s$. The attention is computed by the softmax of the scaled dot product between the query and the key. The output of the attention is then added to the original query Q_{enc}^s to preserve the residual connection. (Eq. (3)) depicts the MSCA mechanism:

$$\tilde{Q}_{\text{enc}} = \sum_{s \in \mathcal{S}_{\text{enc}}} \sum_{\substack{s' \in \mathcal{S}_{\text{dec}} \\ s' \neq s}} \left(Q_{\text{enc}}^s + \text{Softmax}\left(\frac{Q_{\text{enc}}^s (K_{\text{dec}}^{s'})^\top}{\sqrt{C}} \right) V_{\text{dec}}^{s'} \right) \tag{3}$$

3.3 Loss Function with Clinical Quality Score

Due to the physical constraints of portable OCT devices, some acquired OCT scans are highly degraded. We thus incorporate clinical expert knowledge into model training by introducing a quality-aware loss function. Each high-resolution training image is assigned a perceptual quality score $S_{\text{quality}}^{(i)}$ derived from subjective ratings provided by ophthalmologists. These ratings (e.g., from 1 to 10) are aggregated via voting and normalized to form a continuous score, reflecting the perceived clinical value of each target OCT image. We design a focal-style loss that enables high-quality scans to have a larger impact during model training. This is formulated as a perceptual quality modulated mean squared error:

$$\text{MSE}_{\text{focal}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - S_{\text{quality}}^{(i)} \right)^\gamma \cdot (x_i - \hat{x}_i)^2 \tag{4}$$

Here, $\gamma < 0$ is a focusing parameter that controls the degree to which high-quality samples are prioritized, usually being set in [-2, -1]. The modulation term $(1 - S_{\text{quality}}^{(i)})^\gamma$ prevents the model from being confused by relatively suboptimal OCT data.

3.4 Datasets and Experiments

Synthetic 500 Dataset We use two medical OCT datasets in our experiments: one synthetic and one clinical. The synthetic dataset is referred to as the Synthetic 500 dataset and contains 524 pairs of high-resolution / simulated low-resolution OCT B-scans; this dataset was designed to simulate the imaging characteristics of portable OCT devices. The high-resolution images were acquired from a Carl Zeiss© Cirrus HD-OCT 5000 machine and serve as ground truth. The low-resolution scans were simulated by first matching the high-resolution scans with the intensity histogram of low-resolution scans, then convolving the histogram-matched scans with the point spread function (PSF) of a

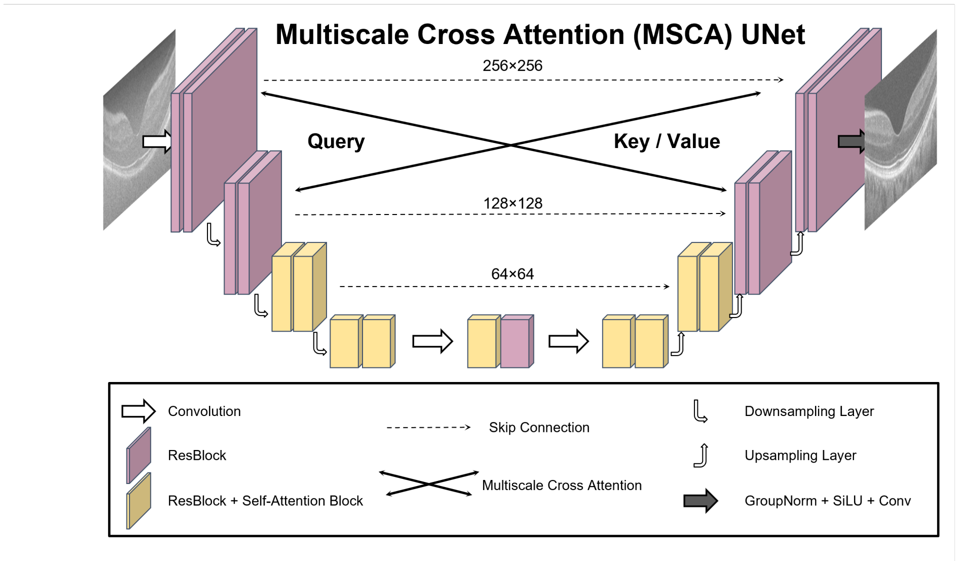

Figure 3: Illustration of the Multi-scale Cross-Attention (MSCA) mechanism. Encoder features at each scale attend to decoder features at different scales, enabling cross-scale information fusion.

portable OCT device, and finally downsampling the scans by a factor of 4 as done in ESRGAN [6] experiments. The Synthetic 500 dataset serves as a controlled setting for proof-of-concept evaluation and comparison with baselines to demonstrate the superiority of OCTDiff.

Philophos 84 Dataset Due to the lack of publicly available low-resolution OCT datasets, we collected our own real-world dataset with the Philophos© KUOS-O100 portable OCT device. OCT B-scans were captured from patients who visited Columbia Ophthalmology from May through July 2024. To obtain paired high-resolution scans under consistent physical and clinical conditions, each participant also underwent an additional OCT scan using a commercial device (Zeiss Cirrus 5000) during the same visit. The portable device has significantly limited imaging capabilities: shallower imaging depth of 2.0 mm (commercial: 2.9 mm), smaller field of view 6.5 $\times$ 6.5 mm (commercial: >8 mm), and reduced spatial resolution of 562 $\times$ 1286 (commercial: >3k).

These hardware limitations result in low-resolution, high-noise, low-contrast images, often with off-centered or partially missing retinal structures. We first filtered out technically corrupted or clinically irrelevant scans, then applied an unsupervised denoising approach [48] to all images. Next, affine registration was performed using ImFusion© software [49] to align remaining scans to a standard OCT template [50] and resizing to 256 for model training. Then we conducted augmentation including flipping, scaling, rotation, elastic deformation, and contrast enhancement. After preprocessing, we obtained 504 paired B-scans from 84 patients, forming the Philophos 84 Dataset that presents realistic challenges for low-quality portable OCT enhancement. OCTDiff is proposed to faithfully super-resolve these degraded scans from "unusable" to "usable", thereby making portable OCT devices clinically valuable.

4 Results

We present quantitative and qualitative results of our OCTDiff against baseline models. We conduct ablation studies to analyze the effects of the ANA and MSCA modules and the quality-score informed loss on model performance, complexity, and training efficiency, including different cross-attention types and the exponential decay rate α in ANA. To demonstrate real-world applicability, we perform downstream disease classification using images generated by OCTDiff, comparing results with those from original low-resolution and target high-resolution scans. Finally, we discuss the limitations of our approach and directions for future work.

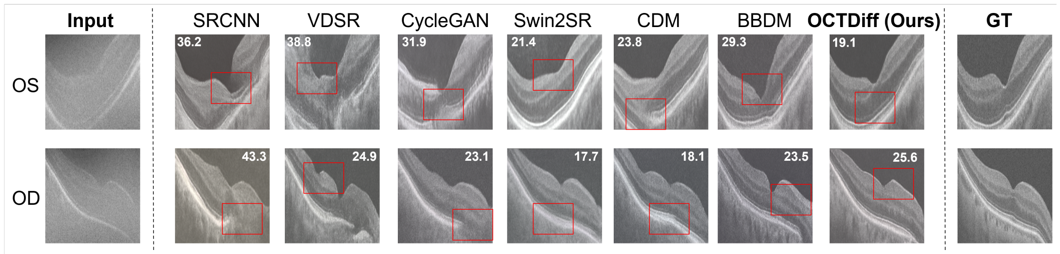

Figure 4: Two examples of reconstructed images from baseline models and our proposed OCTDiff. The first row shows a left-eye (OS) scan, and the second row shows a right-eye (OD) scan. The first and last columns correspond to the input low-resolution image and the ground truth (GT) high-resolution image, respectively. Baseline methods include SRCNN [51], VDSR [52], CycleGAN [53], Swin2SR [54], CDM [35], and BBDM [18]. Red boxes highlight regions with notable degradation or artifacts compared to GT. Each image is annotated with its BRISQUE score [55] to quantify perceptual quality.

Table 1: Quantitative comparison of models on Philophos 84 and Synthetic 500 datasets. Arrows indicate the desirable direction for each metric.

	Philophos 84 Dataset			Synthetic 500 Dataset[5]		
Model	SSIM% ↑	PSNR ↑	LPIPS% ↓	SSIM% ↑	PSNR ↑	LPIPS% ↓
SRCNN	39.7	18.4	49.3	91.7	28.2	9.2
VDSR	26.1	17.9	47.7	85.2	33.0	12.7
CycleGAN	58.4	29.2	28.3	95.3	30.3	17.9
Swin2SR	78.8	35.9	18.3	96.9	38.2	4.7
CDM	71.9	33.2	31.5	98.6	34.2	14.3
BBDM	87.2	35.3	27.9	98.1	**42.7**	6.2
OCTDiff(ours)	**93.6**	**38.8**	**16.1**	**98.9**	41.0	**1.7**

4.1 Performance

OCTDiff is trained and tested separately on the Philophos 84 and Synthetic 500 dataset. Training is conducted on a Lambda Labs Vector server equipped with two NVIDIA©A6000 GPUs, requiring approximately 48 hours to complete from scratch with input images resized to 256×256 pixels and a total diffusion time step $T = 1000$. For quantitative evaluation, we employ structural similarity index measure (SSIM) [56], peak signal-to-noise ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [57] to comprehensively assess reconstruction fidelity, pixel-level accuracy, and perceptual quality, respectively.

The baseline convolutional models include SRCNN [51] and VDSR [52]. Since previous work [5] on the Synthetic 500 dataset implemented ESRGAN and MedGAN, we include CycleGAN [53] for comparison. Swin2SR is chosen as a state-of-the-art transformer-based super-resolution model leveraging hierarchical self-attention and shifted windows to compare with our MSCA strategy. Regarding diffusion models, we implement CDM and the bridged model BBDM [18] as a foundation for our work. All models are trained from scratch under the same training strategy without pretraining to ensure a fair comparison.

Quantitative Results Our OCTDiff achieves the best performance in terms of SSIM (0.936 and 0.989 on the Philophos 84 and Synthetic 500 datasets, respectively) and attains the highest PSNR of 38.8 on the Philophos 84 dataset, as shown in Table 1. Most baseline models perform well on the Synthetic 500 dataset and achieve over 0.95 SSIM, which is comparable to OCTDiff, with BBDM reaching the highest PSNR of 42.7. This is likely because the Synthetic 500 dataset preserves scaling and local structural information during the synthesis process from high-resolution to low-resolution images, making the super-resolution task relatively easier for models employing local upsampling techniques such as convolutional kernels. Model performances drop on the Philophos 84 dataset, while our OCTDiff outperforms all baselines.

Table 2: Ablation study on ANA and MSCA in OCTDiff on Philophos 84 dataset.

ANA	MSCA	SSIM% ↑	Params (M)	FLOPs (G)
		86.1	22.8	406.2
✓		90.7	22.8	451.3
	✓	88.0	23.7	613.9
✓	✓	**93.6**	23.7	762.6

Table 3: Ablation study on ANA exponential decay rate α on Philophos 84 dataset.

α	SSIM% ↑	PSNR ↑
0.1	91.2	37.1
0.3	**93.6**	37.8
0.6	84.6	**38.8**
1.0	82.8	38.0

Qualitative Outcome Figure 4 visualizes two representative examples of outputs generated by all models on the Philophos 84 dataset. Each image is annotated with its BRISQUE score [55], a reference-less perceptual quality metric for which lower values indicate better visual quality. Red boxes highlight regions of structural deviation from the ground truth. Convolution-based models SRCNN and VDSR fail to produce structurally stable outputs, often introducing artifacts and distorted anatomical layers. CycleGAN generates visually-coherent results but tends to average out fine-grained variations, missing subtle retinal layers especially at the bottom of the scan. Swin2SR produces smooth and consistent textures but over-flattens important curvature details that compromise anatomical realism. Diffusion-based models like CDM and BBDM better reconstruct the global retinal structure but struggle with precise local detail, leading to extra peaks, dips, or distortions around the fovea region, which are critical in clinical interpretation. Our OCTDiff not only preserves global coherence but also recovers sharp structural boundaries close to the ground truth, although it cannot replicate small subject-specific variations.

4.2 Ablation Study

Impact of ANA and MSCA Modules To evaluate the contribution of the ANA and MSCA modules, we tested on different combinations of them and measured their impact on model performance, size (number of trainable parameters), and training cost (in FLOPs) [58], as summarized in Table 2. Another example of output image and corresponding residual maps compared to the ground truth are shown in Figure 5. While ANA does not visibly alter the residual ratio, it effectively suppresses fundamental errors (i.e., the discontinuities in retinal layers) that occur when only MSCA is present.

ANA introduces no additional trainable parameters but increases FLOPs, offering a trade-off that results in SSIM increase. In contrast, MSCA yields more modest performance improvements but is crucial for maintaining spatial consistency, particularly for medical images that require structural integrity for diagnosis. In summary, ANA serves as the primary driver of performance gain and MSCA provides structural regularization. Both components together give rise to the superior performance of OCTDiff.

Weight Decay Factor Another key hyperparameter in ANA is the exponential decay rate α, which controls how quickly the adaptive noise modulation weights diminish as introduced in Section 3.1. Table 3 reports the results with varying α values. When α is low (e.g., 0.1), the model keeps dependency on noise from earlier time steps over a longer duration and yields relatively high SSIM but slightly lower PSNR, reflecting good structural preservation but less sharpness. Increasing α further (e.g., 1.0) makes the model focus on the most recent two to three steps. This over-smoothing causes loss of fine structural details. This trade-off suggests that a moderate α achieves the best balance, and thus we selected $\alpha = 0.3$ for our experiments.

Choices of Cross Attention To justify our MSCA design of using encoder features as queries to attend to decoder features, we provide an empirical analysis to compare different cross attentions (CA), including unidirectional and bidirectional CA. The results are reported in Table 4.

Reverse CA (Decoder→Encoder) performed marginally worse though it converges faster, possibly due to the decoder lacking detailed structural localization at early stages that makes the query less efficient. Bidirectional CA is computationally expensive (twice as many attention layers per scale), with limited benefits. Unidirectional CA (Encoder→Decoder) yields the most significant improvement. The cost and scalability are also key impact factors for our decision when connecting two scales, as our ultimate goal is to deploy OCTDiff onto clinical OCT devices to achieve real-time image processing. We thus prioritized encoder-to-decoder attention in our task.

Table 4: Ablation study on cross attention types in OCTDiff on Philophos 84 dataset.

MSCA Type	SSIM% ↑	Params (M)	FLOPs (G)
No CA	86.1	22.8	406.2
Enc → Dec CA	**88.4**	23.7	613.9
Dec → Enc CA	87.2	23.7	607.7
Bidirectional CA	87.8	25.2	678.4

Table 5: Comparison of loss with and without quality score on Philophos 84 dataset.

Metric	No QS	With QS
SSIM% ↑	74.1	**93.6**
PSNR ↑	34.9	**38.8**
LPIPS% ↓	22.4	**16.1**
BRISQUE ↓	31.7	**24.5**

Quality Score in Loss Function To quantify how much the clinical input contributed to the overall performance, we compared all metrics with and without quality-aware loss as illustrated in 3.3. The results are in Table 5; the quality score significantly improved performance. This highlights how domain knowledge is required for the perceptual and structural understanding of OCT images.

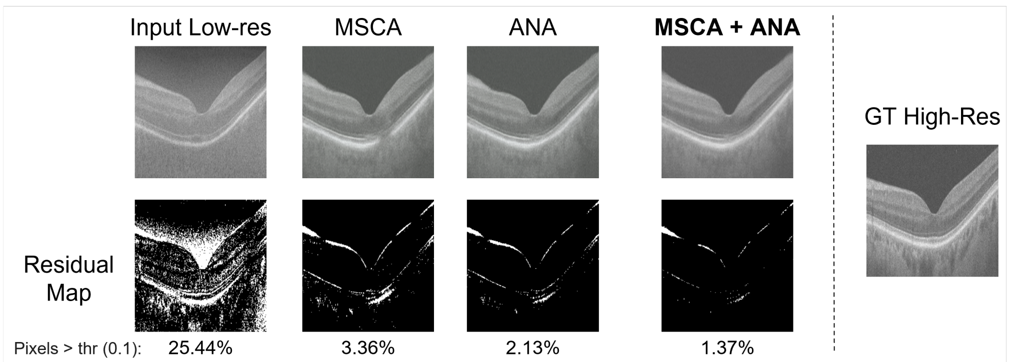

Figure 5: Example output images when toggling MSCA and ANA. The second row shows the corresponding residual maps of the images in the first row compared with GT on the right. At bottom, the ratio of pixels that have a difference over 10% are shown.

4.3 Downstream Disease Classification

Table 6: Accuracy (%) on downstream disease classification using high-resolution images from commercial OCT, low-resolution images from portable OCT, images generated from our OCTDiff and BBDM models, in columns 1, 2, 3 and 4, respectively. The * indicates p-value $\leq$ 0.05 compared to the OCTDiff-generated input. The numbers are in the format of mean $\pm$ standard deviation.

Disease Class	**Model**	**High Res.**	**Low Res.**	**OCTDiff**	**BBDM**
Glaucoma	ViT	74.4 $\pm$1.9	50.1 $\pm$1.8*	74.5 $\pm$3.1	62.0 $\pm$2.5*
	CNN2D	93.4 $\pm$5.0	83.3 $\pm$2.2*	93.5 $\pm$0.5	81.0 $\pm$1.0*
	SwinT	75.5 $\pm$2.6*	49.5 $\pm$3.1*	55.6 $\pm$4.3	57.1 $\pm$3.2
AMD	ViT	86.5 $\pm$1.9	48.9 $\pm$2.7*	85.8 $\pm$1.6	79.5 $\pm$2.1
	CNN2D	94.6 $\pm$0.9*	83.0 $\pm$1.4*	96.4 $\pm$0.4	91.8 $\pm$0.7*
	SwinT	82.3 $\pm$1.4*	49.4 $\pm$2.4*	66.3 $\pm$1.8	64.9 $\pm$2.6

We perform downstream disease classification using images generated by OCTDiff to directly demonstrate its real-world utility in clinical applications. To reduce model bias, we trained three classification architectures: ViT [59], vanilla CNN, and SwinT [60] on three types of inputs: (1) high-resolution images from commercial OCT devices, (2) low-resolution images from portable OCT devices, (3) OCTDiff super-resolved images, and (4) BBDM super-resolved images. The pretrained weights on ImageNet1K [61] are imported. We performed 5-fold cross-validation for each model-dataset pair, with the results shown in Table 6, and we conducted Mann–Whitney U tests [62] to assess statistical significance. This evaluation was conducted independently for two representative ophthalmic tasks: glaucoma diagnosis and age-related macular degeneration (AMD) classification, both of which are widely studied in AI ophthalmology [63, 64, 65, 66, 67], as they are the top two causes of blindness worldwide.

Among the six models trained on high-resolution images, three showed no statistically significant difference, with the simplest vanilla CNN performing equally well in accuracy compared to counterparts trained on OCTDiff-generated images. In contrast, models trained on low-resolution portable OCT images exhibited a statistically significant drop. BBDM is selected as a representative of SR baselines, which is outperformed by OCTDiff in 5 out of 6 rows. These comparisons highlight OCTDiff's ability to transform previously suboptimal portable OCT scans into diagnostically reliable images, indicating that our method can closely match gold-standard OCT-image quality, or even exceed the performance achieved by commercial high-resolution scans.

4.4 Limitations and Future Work

While OCTDiff demonstrates SOTA performance within each clinical dataset, the cross-dataset generalization, such as training on Synthetic 500 and testing on Philophos 84 and vice versa, is still under investigation due to the limited amount of data in each set. While OCTDiff is not positioned to be a general-purpose SR benchmark, further experiments on widely used natural image SR datasets are still needed to more comprehensively prove the advantages of the OCTDiff algorithm. Additionally, although OCTDiff is effective, the model remains relatively large and computationally demanding, leading to longer training times. The cross-scale attention mechanism shows promising benefits, but its performance is sensitive to the choice of scale combinations. Currently, only scales of 128, 64, and 32 have been explored.

Future work will focus on overcoming current limitations by exploring latent-space methods [16] such as LDM, pre-training the model using natural images, and cross-dataset evaluations to further validate OCTDiff's generalizability. Toward system-level advancement, future work will include integrating OCTDiff into physical portable OCT devices using edge computing platforms like NVIDIA©Jetson Orin Nano [68] to enable real-time image enhancement for point-of-care clinical applications. Safeguarding against hallucination in generated OCT images will be addressed in future work via posthoc clinical quality assessment.

We plan to extend OCTDiff to other medical imaging and broader fields. OCT shares core signal degradation characteristics, particularly speckle noise, with radiological imaging modalities such as ultrasound and low-dose CT. Also our algorithm's novel modules (ANA and MSCA) are not restricted to a specific image type. For these reasons, OCTDiff has potential for cross-modal generalization. There are no publicly available portable OCT datasets to our knowledge, and portable medical datasets are very rare in general, indicating the paradigm-shifting role of AI's entry into this field. Our work paves the way for the growing trend of developing tandem AI + imaging technology: making AI algorithms more scalable, interoperable, and easier to deploy within a portable form factor, enabling accessibility for the broadest populations at point-of-care.

5 Conclusion

We propose OCTDiff, a bridged diffusion framework specifically designed for enhancing portable OCT images. We further propose Adaptive Noise Aggregation (ANA) to improve noise scheduling within the diffusion process, allowing the model to adaptively leverage information from multiple time steps.We incorporate Multi-Scale Cross-Attention (MSCA) to effectively capture spatial dependencies at multiple resolutions. To mitigate overfitting on limited clinical datasets and preserve diagnostically critical structures, we introduce a customized loss function guided by clinical quality scores. Downstream disease classification demonstrates that OCTDiff significantly improves image quality and makes low-cost OCT scans clinically usable. Our work lays the groundwork to revolutionize the current workflow of ophthalmic diagnostics with AI's power, making it more accessible and cost-effective, ultimately improving healthcare outcomes.

6 Acknowledgments and Disclosure

The authors would like to thank Tharun Kumar Jayaprakash, Dr. Vlad DiaConita, Dr. Ives (Tony) Valenzuela, Dr. Jack Cioffi, and our funding sources: Research to Prevent Blindness, Inc., the Peacock Trust, and Columbia University's Office of the Provost.

References

[1] Samir S Yadav and Shivajirao M Jadhav. "Deep convolutional neural network based medical image classification for disease diagnosis". In: *Journal of Big data* 6.1 (2019), pp. 1–18.

[2] Qing Li et al. "Medical image classification with convolutional neural network". In: *2014 13th international conference on control automation robotics & vision (ICARCV)*. IEEE. 2014, pp. 844–848.

[3] Risheng Wang et al. "Medical image segmentation using deep learning: A survey". In: *IET image processing* 16.5 (2022), pp. 1243–1267.

[4] Monika Agarwal and Rashima Mahajan. "Medical image contrast enhancement using range limited weighted histogram equalization". In: *Procedia Computer Science* 125 (2018), pp. 149–156.

[5] Kaveri A Thakoor et al. "Enhancing portable OCT image quality via GANs for AI-based eye disease detection". In: *International Workshop on Distributed, Collaborative, and Federated Learning*. Springer. 2022, pp. 155–167.

[6] Xintao Wang et al. "Esrgan: Enhanced super-resolution generative adversarial networks". In: *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018.

[7] Karim Armanious et al. "MedGAN: Medical image translation using GANs". In: *Computerized medical imaging and graphics* 79 (2020), p. 101684.

[8] Li Kang et al. "3D-MRI super-resolution reconstruction using multi-modality based on multi-resolution CNN." In: *Comput. Methods Programs Biomed.* 248 (2024), p. 108110.

[9] Junyoung Park et al. "Computed tomography super-resolution using deep convolutional neural network". In: *Physics in Medicine & Biology* 63.14 (2018), p. 145011.

[10] Hu Chen et al. "Low-dose CT denoising with convolutional neural network". In: *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE. 2017, pp. 143–146.

[11] Tobit Klug et al. "Motionttt: 2d test-time-training motion estimation for 3d motion corrected mri". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 40615–40650.

[12] Zheng Chen et al. "Binarized diffusion model for image super-resolution". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 30651–30669.

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

[14] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[15] Xingyi Yang and Xinchao Wang. "Diffusion model as representation learner". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 18938–18949.

[16] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[17] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[18] Bo Li et al. "Bbdm: Image-to-image translation with brownian bridge diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*. 2023, pp. 1952–1961.

[19] Hayit Greenspan. "Super-resolution in medical imaging". In: *The computer journal* 52.1 (2009), pp. 43–63.

[20] Chenlin Meng et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations". In: *arXiv preprint arXiv:2108.01073* (2021).

[21] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models". In: *International conference on machine learning*. PMLR. 2021, pp. 8162–8171.

[22] Jacob Austin et al. “Structured denoising diffusion models in discrete state-spaces”. In: *Advances in neural information processing systems* 34 (2021), pp. 17981–17993.

[23] Sonia Rehman et al. “Improved Image Denoising: A Combination Method Using Multiscale Contextual Fusion and Recursive Learning”. In: *IET Image Processing* 19.1 (2025), e70143.

[24] Shanchuan Lin and Xiao Yang. “Diffusion model with perceptual loss”. In: *arXiv preprint arXiv:2401.00110* (2023).

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. URL: `https://arxiv.org/abs/2006.11239`.

[26] Tero Karras et al. “Elucidating the Design Space of Diffusion-Based Generative Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. URL: `https://arxiv.org/abs/2206.00364`.

[27] Hui Shen et al. “Efficient diffusion models: A survey”. In: *arXiv preprint arXiv:2502.06805* (2025).

[28] Tong Xie et al. “Data Attribution for Diffusion Models: Timestep-induced Bias in Influence Estimation”. In: *arXiv preprint arXiv:2401.09031* (2024).

[29] Yize Li et al. “Pruning then reweighting: Towards data-efficient training of diffusion models”. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.

[30] Mengfei Xia et al. “Towards more accurate diffusion model acceleration with a timestep tuner”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 5736–5745.

[31] Subham S Sahoo et al. “Diffusion models with learned adaptive noise”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 105730–105779.

[32] Yiming Lei et al. “Denoising diffusion path: Attribution noise reduction with an auxiliary diffusion model”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 54003–54025.

[33] Trevor Norton and Debswapna Bhattacharya. “Sifting through the noise: A survey of diffusion probabilistic models and their applications to biomolecules”. In: *Journal of Molecular Biology* 437.6 (2025), p. 168818.

[34] Daivik Rao, P Yatish Sriram, Aniket Gupta, et al. “ImageCraft-Text To Image Synthesis”. In: *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*. IEEE. 2024, pp. 1–5.

[35] Chitwan Saharia et al. “Palette: Image-to-image diffusion models”. In: *ACM SIGGRAPH 2022 conference proceedings*. 2022, pp. 1–10.

[36] Jonathan Ho et al. “Cascaded diffusion models for high fidelity image generation”. In: *Journal of Machine Learning Research* 23.47 (2022), pp. 1–33.

[37] Wenqiang Sun et al. “Spatial-aware latent initialization for controllable image generation”. In: *arXiv preprint arXiv:2401.16157* (2024).

[38] Zhangxuan Gu, Haoxing Chen, and Zhuoer Xu. “Diffusioninst: Diffusion model for instance segmentation”. In: *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2024, pp. 2730–2734.

[39] Weifeng Chen et al. “Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning”. In: *arXiv preprint arXiv:2305.13840* (2023).

[40] Yash Balaji et al. “eDiffi: Text-to-Image Diffusion Models with Editable Outputs”. In: *arXiv preprint arXiv:2211.01324* (2022).

[41] Yuki Endo. “Masked-attention diffusion guidance for spatially controlling text-to-image generation”. In: *The visual computer* 40.9 (2024), pp. 6033–6045.

[42] Samuli Laine and Timo Aila. “Temporal ensembling for semi-supervised learning”. In: *arXiv preprint arXiv:1610.02242* (2016).

[43] Xiangzuo Huo et al. “HiFuse: Hierarchical multi-scale feature fusion network for medical image classification”. In: *Biomedical Signal Processing and Control* 87 (2024), p. 105534.

[44] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *arXiv preprint arXiv:2011.13456* (2020).

[45] Cheng Lu et al. “Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps”. In: *Advances in neural information processing systems* 35 (2022), pp. 5775–5787.

[46] Diederik P Kingma. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising diffusion implicit models”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2010.02502.

[48] Dewei Hu, Yuankai K Tao, and Ipek Oguz. “Unsupervised denoising of retinal OCT with diffusion probabilistic model”. In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 25–34.

[49] Oliver Zettinig et al. “Recent advances in point-of-care ultrasound using the ImFusion Suite for real-time image analysis”. In: *International Workshop on Point-of-Care Ultrasound*. Springer. 2018, pp. 47–55.

[50] Lingjiao Pan and Xinjian Chen. “Retinal OCT image registration: methods and applications”. In: *IEEE reviews in biomedical engineering* 16 (2021), pp. 307–318.

[51] Chao Dong et al. “Learning a deep convolutional network for image super-resolution”. In: *European conference on computer vision (ECCV)*. Springer. 2014, pp. 184–199.

[52] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Accurate image super-resolution using very deep convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 1646–1654.

[53] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2017, pp. 2223–2232.

[54] Marcos V Conde et al. “Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration”. In: *European conference on computer vision*. Springer. 2022, pp. 669–687.

[55] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-reference image quality assessment in the spatial domain”. In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[56] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[57] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[58] Raphael Tang, Ashutosh Adhikari, and Jimmy Lin. “Flops as a direct optimization objective for learning sparse neural networks”. In: *arXiv preprint arXiv:1811.03060* (2018).

[59] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[60] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10012–10022.

[61] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 248–255.

[62] Henry B Mann and Donald R Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[63] Daniel SW Ting et al. “Deep learning in ophthalmology: The technical and clinical considerations”. In: *Progress in Retinal and Eye Research* 72 (2019), p. 100759.

[64] Felix Grassmann et al. “A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography”. In: *Ophthalmology* 125.9 (2018), pp. 1410–1420.

[65] Daniel S Kermany et al. “Identifying medical diagnoses and treatable diseases by image-based deep learning”. In: *Cell* 172.5 (2018), pp. 1122–1131.

[66] Daniel B Russakoff et al. “Deep learning for prediction of AMD progression: a pilot study”. In: *Investigative ophthalmology & visual science* 60.2 (2019), pp. 712–722.

[67] Liu Li et al. “Attention based glaucoma detection: A large-scale database and CNN model”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10571–10580.

[68] Agus Kurniawan. “Introduction to nvidia jetson nano”. In: *IoT Projects with NVIDIA Jetson Nano: AI-Enabled Internet of Things Projects for Beginners*. Springer, 2020, pp. 1–6.

A Technical Appendices and Supplementary Material

Supplementaty materials will be uploaded in a different PDF file.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our proposal of OCTDiff, a bridged diffusion model for super-resolution and enhancement of portable OCT images, along with the introduction of key components such as Adaptive Noise Aggregation (ANA) 3.1, Multi-Scale Cross-Attention (MSCA) 3.2 and a customized loss function 3.3 with clinical quality score. The results include improved image quality, quantitative metrics 4.1, downstream clinical applications 4.3 and ablation studies 4.2. The claims in the abstract are well supported by our results and align with the content presented throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly acknowledge several limitations in Section 4.4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our work is primarily empirical and methodological, focusing on developing and evaluating the OCTDiff model for OCT image super-resolution and enhancement. It does not present new theoretical results that require formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary details on the model architecture in Figures 2 and 3, training procedure including dataset splits, hyperparameters (including the weight decay factor α), and evaluation protocols. This will enable reproduction of the main experimental results supporting the paper's claims. Additionally, the code and sample data (through figures in this publication) are now being made publicly available in this camera-ready version.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code URL is being provided in this camera-ready version of the paper, with clear instructions and documentation. The complete clinical dataset is currently under IRB restrictions and ongoing collection, and will be made publicly available in a separate publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We clearly specify the training details including the data splits, α hyperparameter settings in Section 4.2, GPU usage and rationale behind their selection. These details provide sufficient transparency to understand and reproduce the results. Other details such as batch size and optimizer option will be upon each user's preference.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance in Section 4.3, and we have updated the results using the Mann–Whitney U test in this camera-ready version.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify computational cost of models (Section 4.2) and hardware specifications such as GPU (Section 4.2) and training time (Section 4.2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work complies fully with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We emphasized and evaluated the significant societal impacts of our work in AI healcare (Section 1, 4.3, 4.4 and 5). OCTDiff has the potential to revolutionize the current workflow of ophthalmic diagnostics with AI's power, making it more accessible and cost-effective, ultimately improving industrial healthcare outcomes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our work focuses on medical image enhancement and does not involve high-risk generative models such as large language models or general-purpose image generators. The model is intended solely for clinical research use. The data used are de-identified and subject to IRB protocols, and the release of code will be accompanied by clear usage disclaimers to prevent misuse outside intended medical and academic contexts.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets used in the paper, including OCT imaging devices (Philophos and Zeiss) pretrained models (e.g., ViT, SwinT) and datasets (e.g., ImageNet1K), are properly cited with corresponding references. Their usage complies with the original licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced (e.g. the OCTDiff model, the Adaptive Noise Aggregation (ANA) module, and a clinical-quality-guided loss function) are thoroughly documented in the main text. In this camera-ready paper, we are now releasing the full source code with comprehensive instructions. A sample dataset will also be provided, while the full clinical dataset will be released separately due to IRB constraints

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: Our work does not involve crowdsourcing or direct research with human subjects. All human-related data (retinal OCT scans) were retrospectively collected under IRB-approved clinical protocols, with appropriate anonymization and ethical safeguards. No compensation was provided, and no participants were recruited specifically for this study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: All clinical OCT data used in our study were collected under Institutional Review Board (IRB) approval number IRB-AAAV1311, in accordance with ethical standards for research involving human subjects. The data were fully deidentified before AI training. No additional risks were posed to participants beyond standard OCT scanning, and informed consent was obtained as part of the data collection protocol.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our work does not employ large language models (LLMs) as part of its core methodology or experimental process. Any use of LLMs was limited to writing and editing assistance that does not affect the scientific novelty or originality of our research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.