

Article

Speckle Noise Reduction in OCT Retinal Images Based on a Multi-Scale Self-Attention Generative Adversarial Network

Haiyi Bian ^{1,*} , Yingzhou Zhu ¹, Yeman Liu ¹, Lei Liu ¹, Xiaoteng Yan ², Zhongjie Cheng ¹  and Yanrong Wang ¹¹ Faculty of Electronic Information Engineering, Huai'an University, Huai'an 223003, China² The Second People's Hospital of Huai'an, Huai'an 223003, China

* Correspondence: bianhaiyi@163.com

Abstract

Optical coherence tomography (OCT) is widely used in biomedical imaging and ophthalmology. However, OCT images are frequently corrupted by speckle noise from coherent light interference. This degradation hampers clinical diagnosis of retinal lesions and identification of tissue layers. We propose a speckle noise reduction algorithm based on a Multi-Scale Self-Attention Generative Adversarial Network combined with a Siamese network (SA-Siamese-GAN). To address the limited receptive fields of traditional convolutional neural networks (CNNs), which can cause broken or blurred retinal layers, a self-attention mechanism is integrated into the bottleneck layer of the generator to capture global pixel dependencies. A Siamese network enforces structural consistency, and a joint loss function combining Wasserstein distance, perceptual loss, and structural similarity (SSIM) is used. Experiments on 11,103 pairs of clinical OCT retinal images show that SA-Siamese-GAN outperforms BM3D, NLM, Wavelet, and GAN-ResNet. It removes speckle noise while preserving retinal layer structure and fine textures, and achieves the highest Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

Keywords: optical coherence tomography (OCT); speckle noise reduction; generative adversarial network (GAN); self-attention mechanism; Siamese network; multi-scale fusion

1. Introduction

Optical Coherence Tomography (OCT) [1] is a widely used high-resolution imaging modality in clinical ophthalmology. It supports tasks such as neovascularization prediction [2], photoreceptor disruption detection [3], and CNN-based biomedical segmentation [4,5], as well as optic nerve head analysis [6], retinal layer extraction [7,8], and 3D image registration [9]. However, OCT images are inherently degraded by multiplicative speckle noise caused by the low-coherence interference of scattered light. This noise obscures tissue boundaries and reduces the reliability of automated diagnostic algorithms.

Hardware-based strategies, including angular compounding [10], strain compounding [11], and spatial compounding [12,13], can suppress speckle noise but require specialized optical configurations that are often incompatible with commercial scanners. Software-based denoising is therefore a practical requirement for tasks such as deformable registration [14]. Early algorithmic approaches relied on spatial filters such as the SUSAN filter [15,16]. Researchers subsequently developed mathematical frameworks based on sparsity [17,18] and spectral fitting [19]. However, these methods rely on hand-crafted priors that often fail to distinguish speckle noise from fine tissue textures, resulting in over-smoothing of clinically significant details.



Received: 16 June 2026

Revised: 9 July 2026

Accepted: 11 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Deep learning has introduced more flexible alternatives for image restoration [20–23]. Convolutional neural networks (CNNs) employing residual learning [24] and perceptual loss functions [25] have established high performance benchmarks for OCT denoising. Generative Adversarial Networks (GANs) [26–28] and unsupervised paradigms such as Noise2Noise [29,30] and Noise2Void [31,32] have also been applied to model complex noise distributions. To maintain structural integrity during denoising, Ma et al. [33] proposed an edge-sensitive cGAN that uses externally computed edge maps as additional supervision. Wang et al. [34] introduced Caps-cGAN, which replaces standard convolutions with capsule networks and dynamic routing to capture spatial feature relationships with fewer parameters, and further employed a semi-supervised learning strategy to leverage unlabeled data. Yu et al. [26] designed a GAN combining multi-scale convolution and dilated convolution Res-blocks for OCT despeckling. Although these GAN-based methods have advanced the field, several limitations remain. The edge-sensitive cGAN relies on external edge detection that may fail under heavy noise. Caps-cGAN uses capsule routing to model spatial relationships but does not explicitly constrain structural consistency between the denoised output and the ground truth. GAN-ResNet employs dilated convolutions for multi-scale representation yet still operates within local receptive fields. More fundamentally, standard convolutional layers [24,25] are constrained by local receptive fields, which can fragment the continuous structural trajectories of retinal layers. Furthermore, unregularized adversarial training [26,27] can introduce periodic pseudo-textures that do not correspond to any underlying anatomy.

To address these issues, we propose a Multi-Scale Self-Attention Generative Adversarial Network with a Siamese consistency constraint (SA-Siamese-GAN) for speckle noise reduction in retinal OCT images. It is the first time that a self-attention mechanism, a Siamese consistency branch, and multi-scale feature fusion are simultaneously integrated within a single GAN framework for this task. Unlike previous GAN-based OCT denoising methods, SA-Siamese-GAN uniquely combines three mechanisms: (1) a self-attention module in the generator bottleneck that captures global inter-layer pixel dependencies to preserve retinal layer continuity during filtering, (2) a Siamese consistency branch that computes L_1 distance between the denoised output and the ground truth in a high-dimensional latent space to penalize structural discrepancies and suppress adversarial artifacts, and (3) multi-scale feature fusion in both the generator (via skip connections) and the discriminator (via cross-scale concatenation) to preserve macro-structural boundaries and micro-textural details simultaneously. The ablation study (Section 3.4) verifies that these three components contribute synergistically. Clinical validation on 11,103 paired OCT images demonstrates that SA-Siamese-GAN achieves superior PSNR and SSIM compared to existing methods.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

Training deep neural networks requires large datasets. Since paired OCT datasets are scarce, this study utilizes two retinal OCT datasets: the Duke University dataset [35] and the Kermany dataset [36]. The combined dataset includes normal fundus, CNV, DR, and ERM cases. All B-scan images were cropped and resized to 512×512 pixels. Data augmentation, including horizontal flipping and random cropping, was applied to expand the dataset to 11,103 image pairs. Target clean images were generated by multi-frame averaging. Specifically, 5 repeated frames for Dataset 1 and 10 to 15 frames for Dataset 2 were registered and averaged per location. A rigid registration algorithm was used to align frames and reduce motion artifacts. Since OCT speckle noise follows a random multiplicative process with a zero-mean characteristic in the logarithmic scale, multi-frame spatial compounding suppresses speckle fluctuations to produce clean reference images.

Experiments were run on a workstation with an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM) using TensorFlow 2.6. The network was trained with the Adam optimizer, a batch size of 4, and 100 epochs. Initial learning rates were 2×10^{-4} for the generator and 2×10^{-5} for the discriminator. After training, image quality metrics were computed on an independent test set to evaluate denoising performance.

2.2. Proposed Denoising Network

We propose a multi-scale generative adversarial network with a self-attention mechanism and a Siamese consistency constraint (SA-Siamese-GAN) for speckle noise reduction in OCT retinal images. The overall architecture is shown in Figure 1.

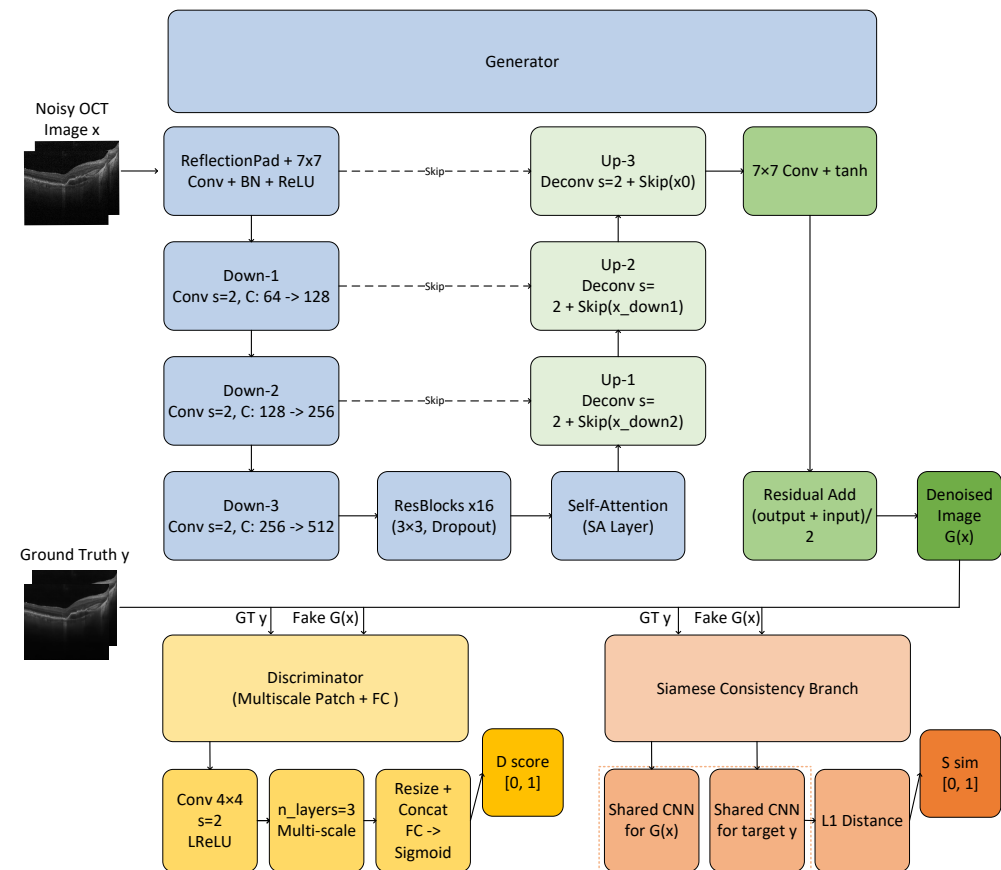


Figure 1. The overall architecture of the proposed SA-Siamese-GAN. It consists of a multi-scale generator, a multi-scale discriminator, and a Siamese consistency branch. x denotes the noisy OCT input, y represents the clean Ground Truth, and $G(x)$ is the denoised output. The dashed boxes denote two shared CNN layers with identical weights.

The SA-Siamese-GAN consists of three components: a multi-scale generator, a multi-scale discriminator, and a Siamese consistency branch.

The noisy OCT image x is fed into the generator, which uses an Encoder-Decoder structure with skip connections. The input passes through a 7×7 convolution with Reflection Padding, Batch Normalization (BN), and ReLU activation. Three down-sampling stages (Down-1, Down-2, Down-3) use convolutions with a stride of 2, increasing feature channels from 64 to 512. The features then enter a bottleneck layer with 16 Residual Blocks (3×3 convolutions with Dropout) and a Self-Attention (SA) layer.

In the decoding phase, three up-sampling stages (Up-1, Up-2, Up-3) use deconvolutions with a stride of 2 to restore spatial resolution. The up-sampled features are fused with the corresponding encoder features through skip connections (Skip(x_{down2}),

Skip(x_{down1}), Skip(x_0)). A final 7×7 convolution with tanh activation and a residual addition produces the denoised image $G(x)$.

The reconstructed image $G(x)$ is evaluated by a dual-branch mechanism. The multi-scale discriminator receives $G(x)$ and the clean ground truth y alternately, processing them through a 4×4 convolution with LeakyReLU and three multi-scale layers. The extracted features are resized, concatenated, and passed through fully connected (FC) layers with Sigmoid activation to output a real/fake probability (D score) in $[0, 1]$.

In the Siamese branch, $G(x)$ and y are fed in parallel into two shared-weight CNNs. The L_1 distance between the encoded feature representations produces a similarity score (S_{sim}) in $[0, 1]$. This constraint penalizes structural discrepancies, helping the generator preserve retinal layer continuity without introducing pseudo-artifacts.

The networks are trained alternately. For each batch, the discriminator and the Siamese network receive 8 weight updates, followed by a single weight update for the generator (critic_updates = 8). The Wasserstein loss is used without gradient penalty. No spectral normalization is applied. All three networks use the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with a learning rate of 10^{-4} .

2.3. Self-Attention Mechanism

Standard convolutions have localized receptive fields and cannot capture long-range dependencies across retinal layers in OCT images. When filtering speckle noise, local processing can fracture continuous tissue boundaries. To address this, a Self-Attention mechanism is integrated into the generator's bottleneck. The module is shown in Figure 2.

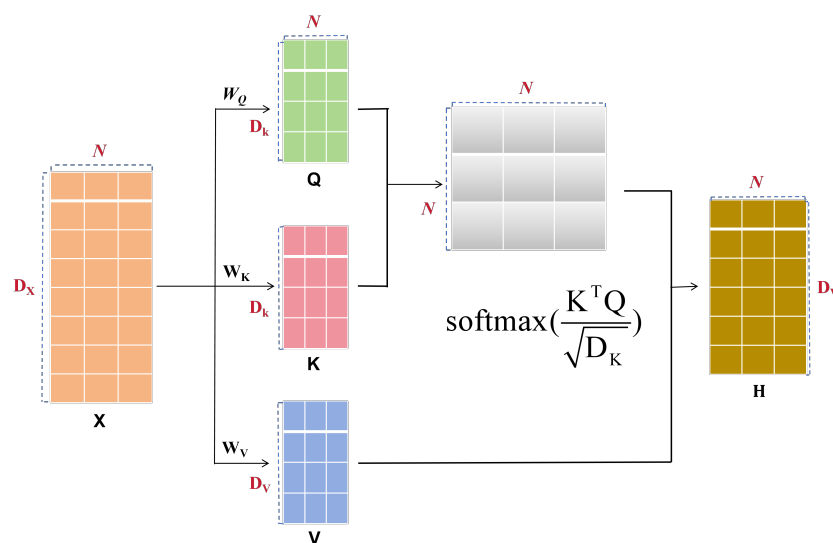


Figure 2. The structure and operations of the Self-Attention mechanism.

Given an input feature map from the preceding convolutional layers, it is flattened along the spatial dimensions to a matrix $\mathbf{X} \in \mathbb{R}^{D_X \times N}$, where $N = H \times W$ and D_X is the number of input channels. To compute dependencies between any two spatial positions, $\mathbf{X}$ undergoes three linear transformations parameterized by $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ (implemented as 1×1 convolutions), producing the Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) matrices.

To reduce computation, the channel dimensions of $\mathbf{Q}$ and $\mathbf{K}$ are compressed to D_K (set to $1/8$ of the original channels), while the dimension D_V of $\mathbf{V}$ is unchanged. The inner product $\mathbf{K}^T \mathbf{Q}$ captures global spatial correlations. The result is scaled by $\sqrt{D_K}$ and passed through a softmax function to produce an $N \times N$ attention weight matrix (Attention_Score):

$$\text{Attention_Score} = \text{softmax}\left(\frac{\mathbf{K}^T \mathbf{Q}}{\sqrt{D_K}}\right) \quad (1)$$

The Value matrix $\mathbf{V}$ is multiplied by the attention weight matrix, producing the globally aggregated output $\mathbf{H} \in \mathbb{R}^{D_V \times N}$.

Finally, $\mathbf{H}$ is reshaped to $(H \times W \times D_V)$. A learnable scalar γ (initialized to zero) scales the attention output, and a residual connection adds it to the original input $\mathbf{X}$. This lets the network learn to balance local convolution and global attention during training, preserving retinal layer continuity while suppressing noise.

2.4. Discriminator and Siamese Consistency Module

A multi-scale feature fusion discriminator evaluates the realism of generated images. As shown in the lower-left of Figure 1, the discriminator consists of five convolutional blocks. The first block uses a 4×4 kernel with stride 2, same-padding, and LeakyReLU ($\alpha = 0.2$) without Batch Normalization. The subsequent three blocks each use 4×4 kernels with stride 2, same-padding, LeakyReLU, and Batch Normalization. The fifth block uses a 4×4 kernel with stride 1. The channel dimensions are 64, 128, 256, 512, and 512, respectively. For multi-scale fusion, feature maps from all five blocks are bilinearly resized to the spatial resolution of the first block and concatenated. A 3×3 convolution is then applied for feature integration, followed by Batch Normalization and LeakyReLU. The output is flattened and passed through two fully connected layers (1024 with tanh, then 1 with sigmoid) to produce the final probability (D score) in $[0, 1]$.

Binary discriminator feedback alone is often insufficient to maintain anatomical fidelity, and can lead to texture distortion during adversarial training. To address this, a Siamese network is introduced as a consistency module, as shown in the lower-right of Figure 1. This branch consists of two parallel CNNs with shared weights. During training, $G(x)$ and y are fed into the two branches. The shared CNN maps each image to a high-dimensional feature space, and the L_1 distance between the two feature representations is mapped to a similarity score (S_{sim}) in $[0, 1]$. This Siamese constraint penalizes structural differences, guiding the generator to preserve retinal layer continuity without introducing pseudo-artifacts.

2.5. Joint Loss Functions

Optimizing a generative model with a single adversarial objective often involves a trade-off between speckle suppression and the preservation of anatomical structures. The optimization objective of SA-Siamese-GAN combines three loss functions with the similarity feedback from the Siamese branch. The total loss $\mathcal{L}_{\text{total}}$ is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}} \mathcal{L}_W + \lambda_{\text{per}} \mathcal{L}_p + \lambda_{\text{ms}} \mathcal{L}_M + \lambda_{\text{siam}} \mathcal{L}_{\text{Siamese}}, \quad (2)$$

where λ_{adv} , λ_{per} , λ_{ms} , and λ_{siam} are hyperparameter weights. In our experiments, these were set to $\lambda_{\text{adv}} = 5$, $\lambda_{\text{per}} = 100$, $\lambda_{\text{ms}} = 100$, and $\lambda_{\text{siam}} = 1$.

$\mathcal{L}_W$ is the Wasserstein adversarial loss. Unlike the Jensen-Shannon divergence used in standard GANs, the Earth-Mover distance optimized by $\mathcal{L}_W$ reduces mode collapse and stabilizes training:

$$\mathcal{L}_W = \mathbb{E}_{x \sim \mathbb{P}_{\text{data}}} [D(x)] - \mathbb{E}_{x \sim \mathbb{P}_{\text{noisy}}} [D(G(x))], \quad (3)$$

where $\mathbb{P}_{\text{data}}$ and $\mathbb{P}_{\text{noisy}}$ denote the real ground truth distribution and the generated data distribution, respectively.

$\mathcal{L}_p$ is the perceptual loss. Instead of pixel-wise error, it compares high-level semantic features extracted by a pre-trained VGG-16 network (the *block3_conv3* layer) from $G(x)$ and y :

$$\mathcal{L}_p = \frac{1}{N} \|\phi(y) - \phi(G(x))\|_2^2, \quad (4)$$

where $\phi(\cdot)$ denotes the semantic feature maps extracted by the specified layer of the VGG network, and N represents the total number of elements in the feature maps.

$\mathcal{L}_M$ is the Structural Similarity loss, minimized as the negative SSIM index to constrain luminance, contrast, and structure:

$$\mathcal{L}_M = -\text{SSIM}(y, G(x)). \quad (5)$$

$\mathcal{L}_{\text{Siamese}}$ is the structural discrepancy penalty from the Siamese branch. The shared CNN encodes y and $G(x)$ into latent vectors. Their L_1 distance is passed through a fully connected layer with Sigmoid activation, producing a similarity score $S \in [0, 1]$. The loss then penalizes deviations from the expected target ($y_{\text{true}} = 1$ during generator optimization):

$$\mathcal{L}_{\text{Siamese}} = \mathbb{E}[y_{\text{true}} \cdot S(G(x), y)]. \quad (6)$$

3. Results

3.1. Visual Evaluation of the Proposed Model

Subjective visual evaluation is important for assessing the clinical utility of denoising methods. Figure 3 shows the denoising results of our model on representative OCT retinal slices.

As shown in Figure 3, the original images contain speckle noise that obscures boundaries between retinal layers. The denoised outputs (b, d, f, h) remove most background speckle noise. Even under heavy noise (e.g., Figure 3a,e,g), the model produces smooth backgrounds while retaining the continuous layered structure of the retina. The network avoids the over-smoothing common in traditional filters.

To further illustrate the micro-structural preservation capabilities of the proposed architecture, Figure 4 provides a magnified view of the Region of Interest (ROI) marked by the red bounding box in Figure 3e.

The magnified ROI shows a morphologically complex region. While speckle masks edge gradients in the original noisy crop, the contours and textural variations are re-constructed in the denoised image. This preservation of fine structure results from the self-attention mechanism and Siamese constraints, which penalize structural distortions.

To assess generalization across different pathological conditions, we evaluated SA-Siamese-GAN on three representative retinal cases: choroidal neovascularization (CNV), diabetic macular edema (DME), and a normal retina. As shown in Figure 5, the model effectively suppresses speckle noise in all three conditions. In the CNV case, the sub-retinal fluid boundary and neovascular membrane remain clearly visible. In the DME case, intra-retinal cystoid spaces and the distorted layer architecture are well preserved. In the normal case, the intact layered structure of the retina is maintained without over-smoothing. These results demonstrate that SA-Siamese-GAN is not specialized to a single disease morphology and generalizes across diverse tissue structures.

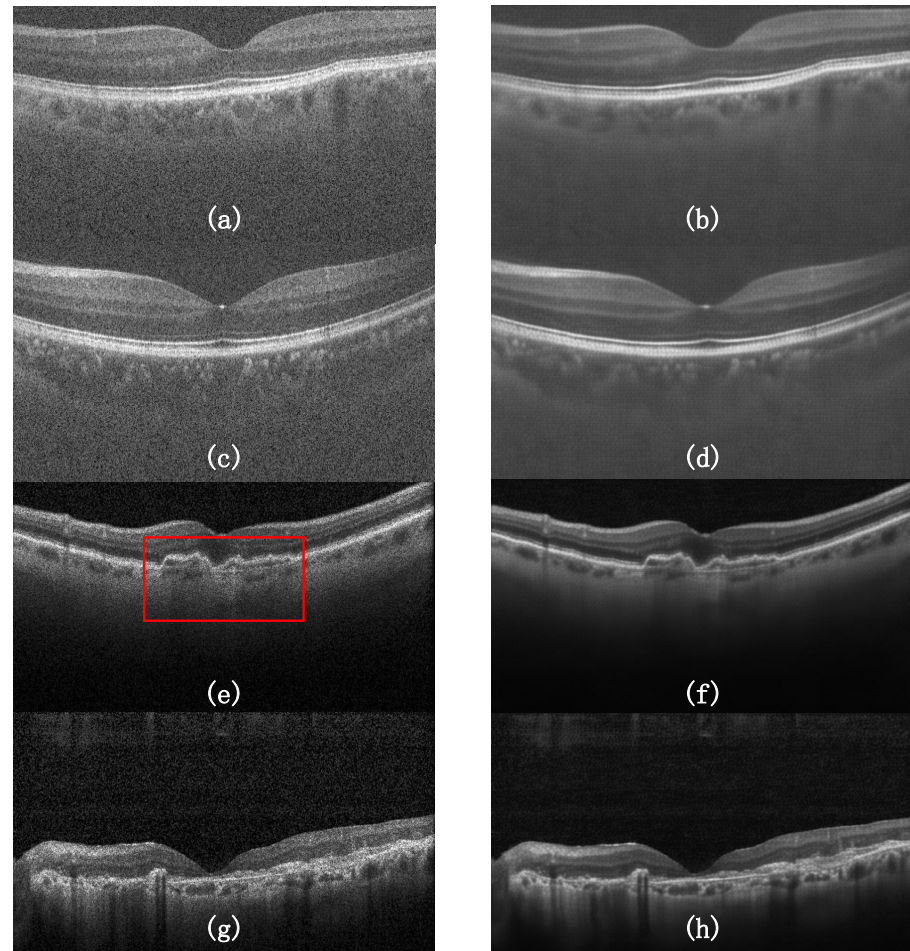


Figure 3. Qualitative denoising results. The left column (a,c,e,g) displays the original noisy OCT images. The right column (b,d,f,h) exhibits the corresponding outputs generated by our model. The red bounding box in (e) indicates the region of interest (ROI) that is magnified and visualized in Figure 4.

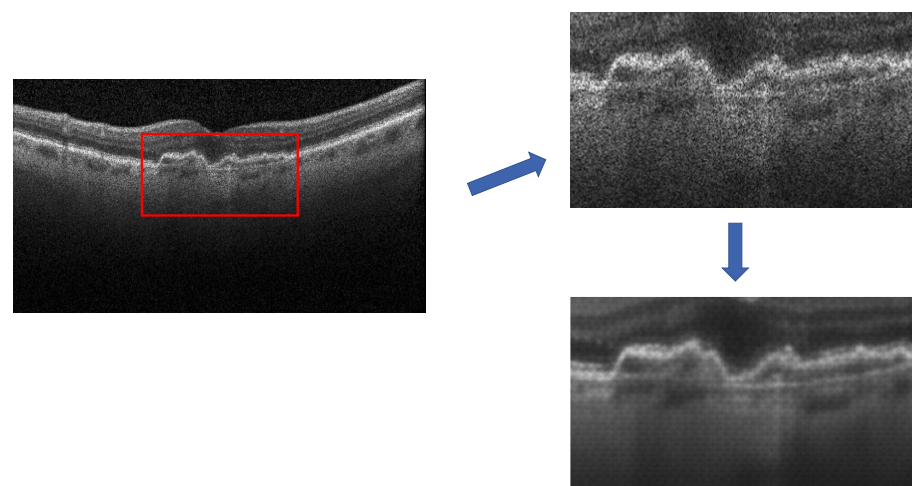


Figure 4. Magnified visualization of local micro-structures. (Top): The original noisy ROI from the red bounding box in Figure 3e. (Bottom): The denoised ROI generated by our model.

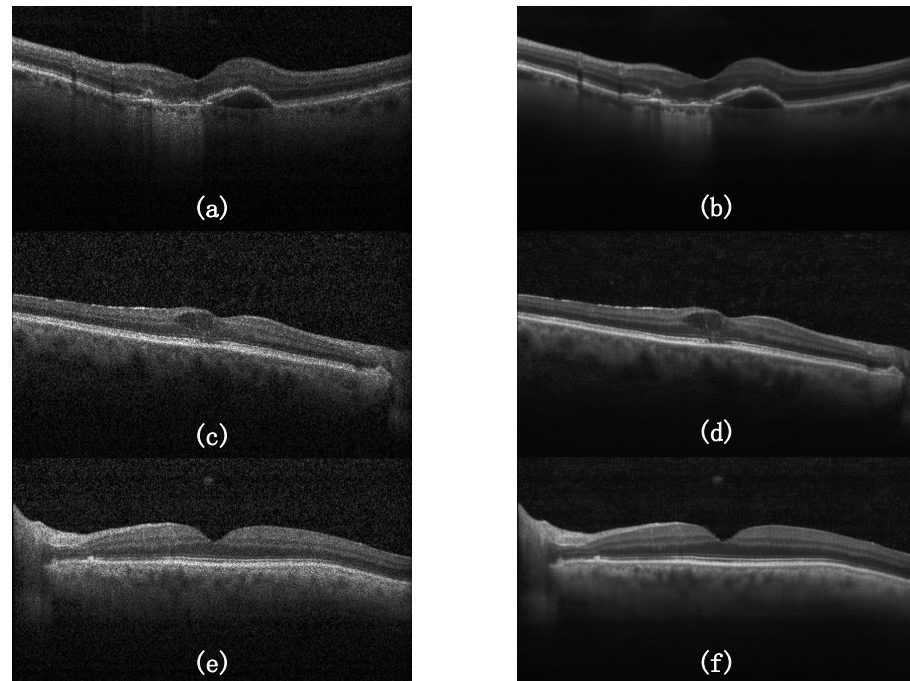


Figure 5. Denoising results across three pathological conditions. (a) CNV noisy input, (b) CNV denoised, (c) DME noisy input, (d) DME denoised, (e) Normal noisy input, (f) Normal denoised.

3.2. Qualitative Comparison with State-of-the-Art Methods

We compared SA-Siamese-GAN against six denoising methods: Wavelet, BM3D, SBSDI, NLM-DIP, MNLM-DIP, and GAN-ResNet. Figure 6 shows the denoising results on a heavily corrupted OCT slice, and Figure 7 provides a magnified view of the central foveal region.

As shown in Figure 6a, the original OCT image is contaminated by dense speckle noise, particularly in the apical and background regions, which impairs diagnostic visibility. The Wavelet method (Figure 6b) removes high-frequency noise but introduces block-like artifacts that distort the tissue appearance. BM3D (Figure 6c) reduces noise but leaves residual speckle and causes slight structural distortions. SBSDI (Figure 6d) over-smoothes the image, blurring retinal layer boundaries.

The deep learning models NLM-DIP (Figure 6e) and MNLM-DIP (Figure 6f) balance noise suppression and structure retention better, but still do not fully recover the hierarchical textures of retinal layers. GAN-ResNet (Figure 6g) produces relatively sharp boundaries. However, constrained by its localized convolutions and unregularized adversarial training, it produces periodic pseudo-textures, including false anatomical lines in the background.

The magnified crops in Figure 7 confirm these observations. SA-Siamese-GAN (Figure 7h) produces the best visual quality. It suppresses background speckle while reconstructing retinal layer boundaries with continuous clarity, and preserves hierarchical textural details without synthetic artifacts.

3.3. Quantitative Evaluation

To supplement visual assessment, we evaluated PSNR and SSIM across two test datasets: OCT-1 (Duke dataset [35]) and OCT-2 (Kermany dataset [36]). Results are shown in Table 1.

Table 1. Quantitative comparison of PSNR (dB) and SSIM metrics for different denoising methods across the OCT-1 and OCT-2 datasets. The best results are highlighted in bold.

Method	OCT-1		OCT-2	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
Wavelet	26.26	0.72	27.66	0.75
BM3D	25.84	0.56	27.28	0.61
SBSDI	26.02	0.66	27.38	0.70
NLM-DIP	26.62	0.75	29.15	0.87
MNLM-DIP	27.51	0.80	29.06	0.81
GAN-ResNet	29.85	0.88	30.24	0.88
Ours (SA-Siamese-GAN)	34.52	0.91	33.13	0.91

As shown in Table 1, SA-Siamese-GAN outperforms all other methods on both datasets.

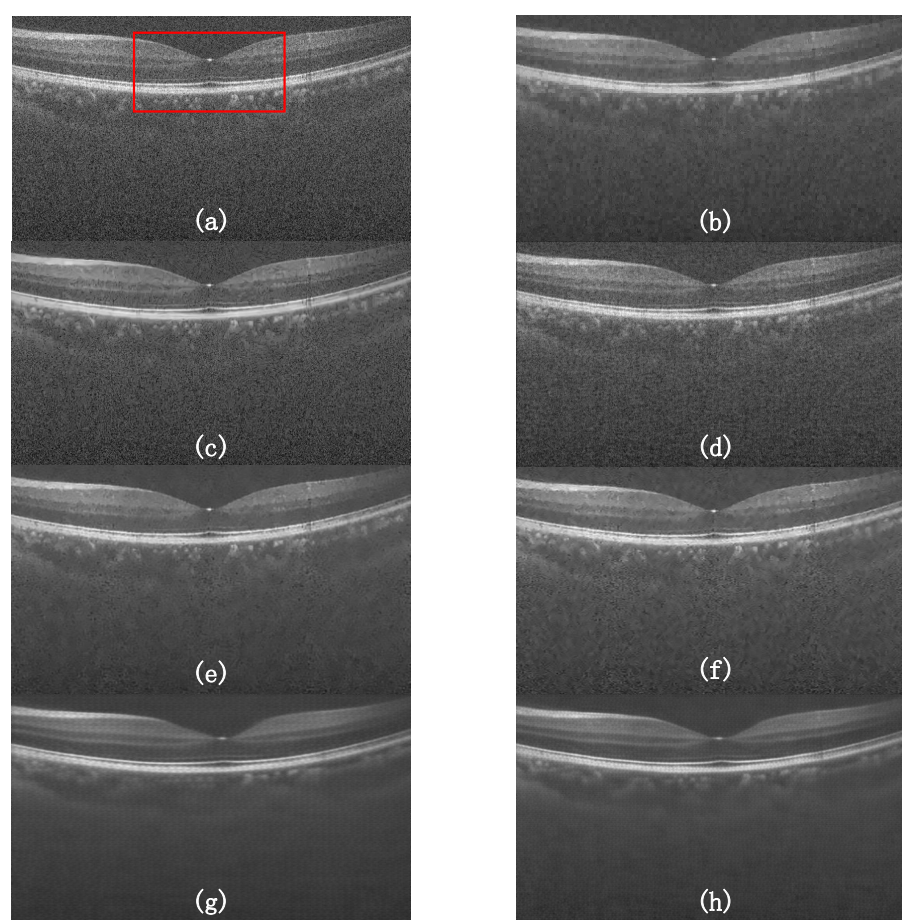


Figure 6. Qualitative comparison of different denoising methods on a severely degraded OCT image. (a) Original noisy image, (b) Wavelet, (c) BM3D, (d) SBSDI, (e) NLM-DIP, (f) MNLM-DIP, (g) GAN-ResNet, and (h) Ours (SA-Siamese-GAN). The red bounding box in (a) indicates the region of interest (ROI) that is magnified and visualized in Figure 7.

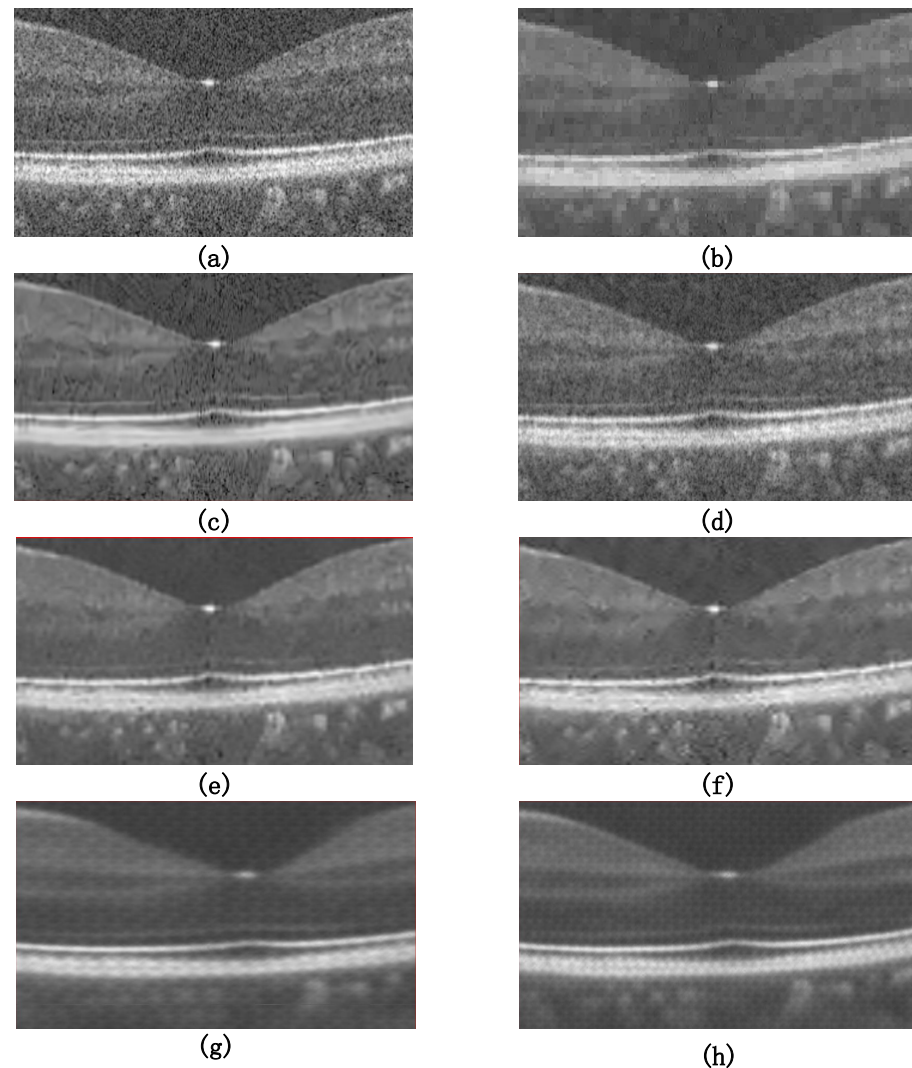


Figure 7. Magnified visualization of the central foveal region (corresponding to the red box in Figure 6a) to compare micro-structural preservation across different methods: (a) original noisy image, (b) Wavelet, (c) BM3D, (d) SBSDL, (e) NLM-DIP, (f) MNLM-DIP, (g) GAN-ResNet, and (h) ours (SA-Siamese-GAN).

Traditional filtering methods (Wavelet, BM3D) produce the lowest PSNR and SSIM scores (e.g., BM3D achieves an SSIM of 0.56 on OCT-1), consistent with the blurring and block artifacts observed in the visual comparison. The deep-learning DIP variants (NLM-DIP, MNLM-DIP) improve PSNR to 27–29 dB, but their SSIM scores suggest that high-frequency structural details remain compromised.

GAN-ResNet achieves a PSNR of 29.85 dB and an SSIM of 0.88 on OCT-1. SA-Siamese-GAN improves PSNR to 34.52 dB on OCT-1 (a gain of 4.67 dB) and 33.13 dB on OCT-2, with SSIM reaching 0.91 on both datasets. These results indicate that the self-attention mechanism and Siamese constraints together enable better speckle removal while preserving anatomical structure. Notably, the performance gap between the two datasets is small (1.39 dB in PSNR and identical SSIM of 0.91), despite their different acquisition protocols and device configurations. This consistency provides evidence of cross-domain robustness. The hyperparameters of all baseline methods are listed in Appendix A.

3.4. Ablation Study

To isolate the contribution of each architectural component, ablation experiments were conducted under a unified reduced configuration (256×256 resolution, 6 residual blocks, 6000 training pairs, and 20 training epochs) to manage computational cost. All variants shared the same training data, random seed, and optimizer settings, with PSNR and SSIM evaluated on a held-out test set of 1000 images. Four variants were compared: (1) the full SA-Siamese-GAN, (2) without the self-attention module in the generator bottleneck (w/o SA), (3) without the Siamese consistency branch (w/o Siamese), and (4) without multi-scale feature fusion in the discriminator (w/o MS-D).

As shown in Table 2, the full model achieves the highest PSNR (37.21 dB) and SSIM (0.9643). Removing any single module consistently degrades both metrics, confirming that each component contributes positively to denoising performance. The multi-scale discriminator contributes the largest individual gain (+0.45 dB). Without multi-scale fusion, the discriminator relies solely on the deepest feature map, which weakens its capacity to capture fine-grained texture discrepancies and reduces the quality of adversarial gradients provided to the generator.

Table 2. Ablation study results. All variants were trained under identical settings (256×256 , 6 residual blocks, 6000 training pairs) and evaluated on a held-out test set of 1000 images.

Model Variant	PSNR (dB)	SSIM	Δ PSNR
SA-Siamese-GAN (Full)	37.21	0.9643	—
w/o Multi-Scale Discriminator	36.76	0.9639	−0.45
w/o Self-Attention	36.95	0.9635	−0.26
w/o Siamese Branch	37.00	0.9629	−0.21

The self-attention module contributes +0.26 dB. While this numerical gain appears modest, pixel-wise metrics alone may understate its value: the self-attention mechanism captures long-range dependencies across the entire feature map, which is essential for preserving the continuity of retinal layer boundaries. As shown in the visual comparison (Figure 8), the w/o SA variant exhibits subtle layer fragmentation that is not fully reflected in the aggregate PSNR score. Similarly, the Siamese branch contributes +0.21 dB. Its primary role is to regularize the generator against pseudo-artifacts during adversarial training—a qualitative benefit visible in the background regions of the denoised outputs.

The combined gain of all three modules (+0.45 dB) is achieved through their synergistic interaction: the multi-scale discriminator enforces global texture realism, the self-attention mechanism preserves structural continuity, and the Siamese branch suppresses spurious anatomical patterns. Removing any one component weakens this synergy, demonstrating that the integrated design of SA-Siamese-GAN is necessary for optimal speckle noise reduction in retinal OCT images.

At the standard configuration (512×512 , 16 residual blocks), average inference time per B-scan is 26.8 ms on an RTX 3090 GPU (~ 37 FPS). Reducing the number of residual blocks to 6 lowers this to 17.7 ms, offering a lightweight variant with a minor trade-off in denoising quality. The complexity of the self-attention module is $O(H^2W^2)$, quadratic in spatial resolution, resulting in 6.8 ms at 256×256 versus 26.8 ms at 512×512 .

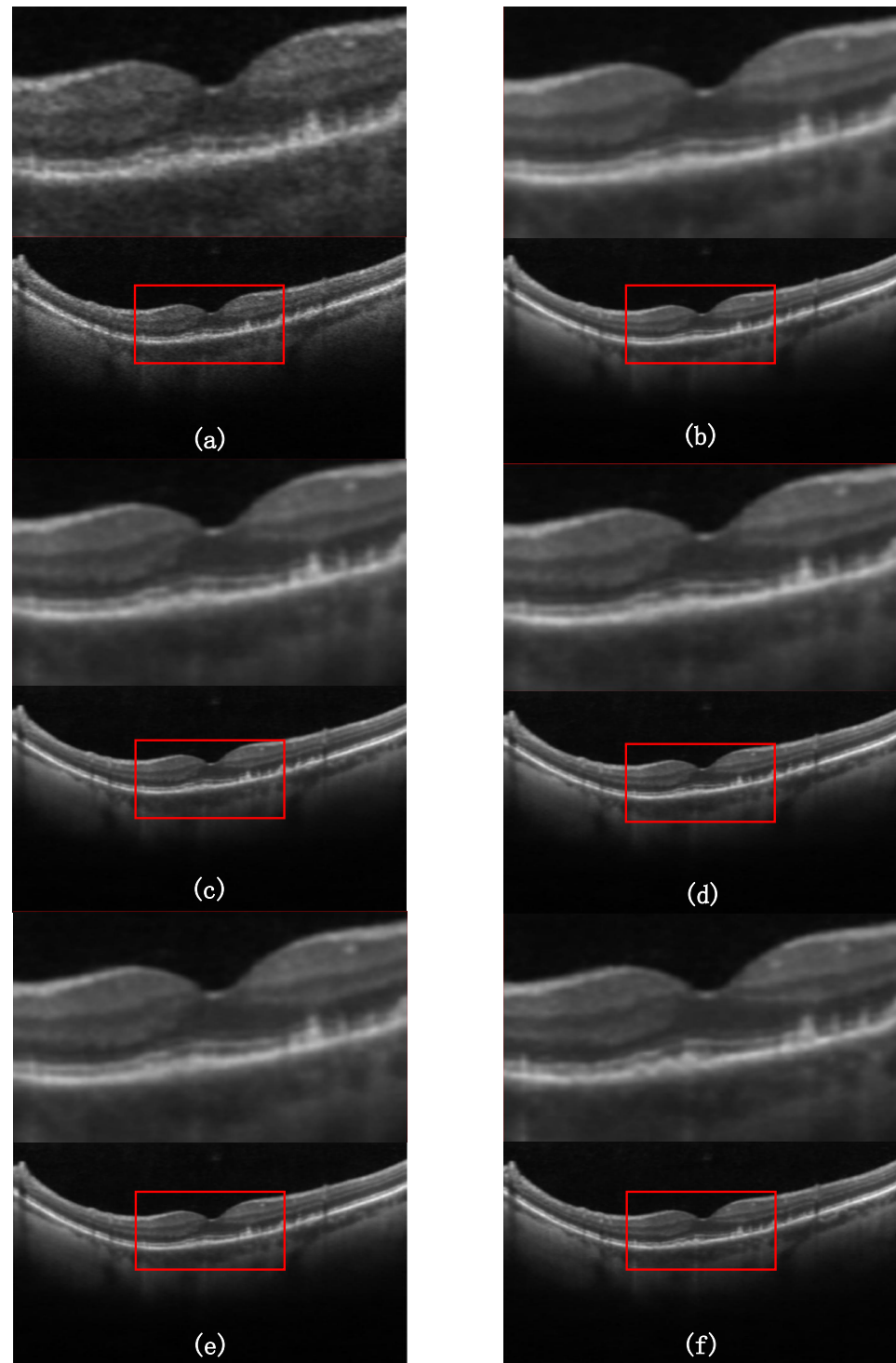


Figure 8. Visual comparison of ablation variants on a representative OCT B-scan. (a) Noisy input; (b) Full SA-Siamese-GAN; (c) w/o Self-Attention; (d) w/o Siamese; (e) w/o Multi-Scale D; (f) Ground truth. In each subfigure, the top panel provides a magnified view of the region of interest (ROI) indicated by the red bounding box. The magnified views highlight localized retinal layer fragmentation in the ablated variants, which is most prominent in the w/o SA output.

4. Discussion

In this article, we proposed SA-Siamese-GAN for speckle noise reduction in retinal OCT images. It is the first time that a self-attention mechanism is combined with a Siamese consistency branch within a multi-scale GAN framework for this task. Unlike previous

GAN-based OCT denoising methods that rely solely on local convolutional operations, SA-Siamese-GAN captures global inter-layer dependencies through self-attention while the Siamese branch penalizes structural discrepancies in latent space. The ablation study (Section 3.4) confirmed that each component contributes independently. The multi-scale discriminator provided the largest gain (+0.45 dB PSNR). The self-attention module and the Siamese branch contributed +0.26 dB and +0.21 dB, respectively. It can be observed from the visual comparison (Figure 8) that the w/o SA variant exhibits subtle layer fragmentation, which indicates that pixel-wise metrics alone understate the structural role of self-attention.

As shown in Table 1, SA-Siamese-GAN achieved a PSNR of 34.52 dB and an SSIM of 0.91 on OCT-1, outperforming all compared methods. Traditional methods such as Wavelet and BM3D produced block-like artifacts and excessive blurring. BM3D achieved an SSIM of only 0.56 on OCT-1. These results confirm that local mathematical priors are insufficient for OCT speckle noise. Deep-learning variants such as NLM-DIP improved PSNR to the 27–29 dB range, but high-frequency textures remained compromised, which may be caused by their restricted receptive fields. GAN-ResNet generated periodic pseudo-textures in background regions. This points to a known limitation of standard GAN architectures: without structural supervision, the generator can produce textures that fool the discriminator without being anatomically correct.

The self-attention module introduces computational overhead that scales quadratically with spatial resolution. As reported in Section 3.4, inference at 512×512 takes 26.8 ms per B-scan on an RTX 3090 GPU. Reducing residual blocks from 16 to 6 lowers this to 17.7 ms. This latency remains well within clinical requirements. Future work could explore linear-complexity attention variants to further reduce computational cost.

There are still several limitations in this study. The comparison methods do not include recent Transformer- or diffusion-based OCT denoising models (e.g., THFN-OCT [37], GARD [38]), because their code and pretrained weights are not publicly available as of this writing. The method has only been validated on 2D B-scans. Although cross-pathology generalization has been demonstrated qualitatively (Figure 5) and the two datasets originate from different devices, a systematic evaluation under varying scan densities and across additional commercial OCT platforms requires dedicated multi-manufacturer data collection and is planned for future work. Extension to 3D volumetric OCT data and real-time 4D-OCT denoising requires further investigation. The benefit of improved denoising on downstream tasks such as retinal layer segmentation and lesion detection has not been quantified. Therefore, evaluating the impact of SA-Siamese-GAN on these clinical tasks is one of our future research directions. Besides, how to further reduce the parameter count while maintaining denoising performance is another focus that will be continuously explored in future work.

5. Conclusions

We presented SA-Siamese-GAN, a multi-scale generative adversarial network with a self-attention mechanism and Siamese consistency constraints for OCT retinal image denoising. The self-attention module in the generator's bottleneck captures global spatial dependencies, enabling the network to reconstruct continuous retinal layers. The Siamese network provides structural supervision that reduces pseudo-artifacts during adversarial training. On clinical OCT datasets, SA-Siamese-GAN achieves a PSNR of 34.52 dB and an SSIM of 0.91, outperforming existing methods. This approach offers a practical solution for improving OCT image quality and supporting automated clinical analysis.

Author Contributions: Conceptualization, H.B.; methodology, H.B.; software, Y.Z. and Y.L.; validation, Y.Z.; formal analysis, Y.Z. and Y.L.; investigation, H.B., Y.Z. and Y.L.; resources, H.B. and L.L.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.,

H.B., X.Y., Z.C. and Y.W.; visualization, Y.Z.; supervision, H.B., L.L. and X.Y.; project administration, H.B. and L.L.; funding acquisition, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China (62205120); the Project of State Key Laboratory of Radiation Medicine and Protection, Soochow University (GZK1202217).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The OCT datasets used in this study are publicly available. Dataset 1 (Duke SD-OCT) can be found at https://people.duke.edu/~sf59/Chiu_BOE_2014_dataset.htm (accessed on 12 January 2025). Dataset 2 (Kermany OCT) is available at <https://doi.org/10.17632/rscbjbr9sj.3> (accessed on 12 January 2025).

Acknowledgments: The authors would like to thank the reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OCT	Optical Coherence Tomography
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index
SA	Self-Attention
BN	Batch Normalization
FC	Fully Connected

Appendix A

Table A1. Hyperparameters of baseline denoising methods.

Method	Hyperparameters
Wavelet	db4 basis, 5 decomposition levels, BayesShrink soft thresholding
BM3D	sigma_psd = 0.20, log-transform VST, default BM3D profile
SBSDI	Default parameters from the original implementation
NLM-DIP	NLM: $h = 7$, search window = 5, block size = 5; DIP: 8000 iterations, $lr = 10^{-4}$
MNLM-DIP	NLM: $h = 10$, search window = 21, block size = 7; DIP: 8000 iterations, $lr = 10^{-4}$
GAN-ResNet	ngf = 64, ndf = 64, 16 ResBlocks, Adam ($lr = 10^{-4}$), batch size = 4, 30 epochs

References

1. Huang, D.; Swanson, E.A.; Lin, C.P.; Schuman, J.S.; Stinson, W.G.; Chang, W.; Hee, M.R.; Flotte, T.; Gregory, K.; Fujimoto, G. Optical coherence tomography. *Science* **1991**, *254*, 1178–1181. [CrossRef] [PubMed]

2. Zhu, S.; Shi, F.; Xiang, D.; Zhu, W.; Chen, H.; Chen, X. Choroid neovascularization growth prediction with treatment based on reaction-diffusion model in 3-D OCT images. *IEEE J. Biomed. Health Informat.* **2017**, *21*, 1667–1674. [CrossRef]

3. Zhu, W.; Huang, Y.; Xie, Y.; Li, J.; Jia, S.; Hao, Q. Automatic three-dimensional detection of photoreceptor ellipsoid zone disruption caused by trauma in the OCT. *Sci. Rep.* **2016**, *6*, 25433. [CrossRef] [PubMed]

4. Rong, Y.; Xiang, D.; Zhu, W.; Yu, K.; Shi, F.; Chen, X. Deriving external forces via convolutional neural networks for biomedical image segmentation. *Biomed. Opt. Express* **2019**, *10*, 3800–3814. [CrossRef] [PubMed]

5. Zhu, W.; Xiang, D.; Rong, Y.; Yu, K.; Shi, F.; Chen, X. An automated framework for intra-retinal cystoid macular edema segmentation in 3D-OCT images with macular hole. *J. Biomed. Opt.* **2017**, *22*, 076014. [\[CrossRef\]](#)
6. Yu, K.; Shi, F.; Gao, E.; Zhu, W.; Chen, H.; Chen, X. Shared-hole graph search with adaptive constraints for 3D optic nerve head optical coherence tomography image segmentation. *Biomed. Opt. Express* **2018**, *9*, 962–983. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Xiang, D.; Zhu, W.; Rong, Y.; Yu, K.; Shi, F.; Chen, X. Automatic segmentation of retinal layer in OCT images with choroidal neovascularization. *IEEE Trans. Image Process.* **2018**, *27*, 5880–5891. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Xiang, D.; Zhu, W.; Rong, Y.; Yu, K.; Shi, F.; Chen, X. Automatic retinal layer segmentation of OCT images with central serous retinopathy. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 283–295. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Pan, L.; Guan, L.; Chen, X. Segmentation guided registration for 3D spectral-domain optical coherence tomography images. *IEEE Access* **2019**, *7*, 138833–138845. [\[CrossRef\]](#)
10. Iftimia, N.; Bouma, B.E.; Tearney, G.J. Speckle reduction in optical coherence tomography by ‘path length encoded’ angular compounding. *J. Biomed. Opt.* **2003**, *8*, 260–263. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Kennedy, B.F.; Hillman, T.R.; McLaughlin, R.A.; Quirk, B.C.; Sampson, D.D. Speckle reduction in optical coherence tomography by strain compounding. *Opt. Lett.* **2010**, *35*, 2445–2447. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Desjardins, A.; Vakoc, B.; Tearney, G.; Bouma, B. Speckle reduction in OCT using massively-parallel detection and frequency-domain ranging. *Opt. Express* **2006**, *14*, 4736–4745. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Huang, B.; Bu, P.; Wang, X.; Chen, X. Speckle reduction in parallel optical coherence tomography by spatial compounding. *Opt. Laser Technol.* **2013**, *45*, 69–73. [\[CrossRef\]](#)
14. Chen, M.; Lang, A.; Ying, H.S.; Calabresi, P.A.; Prince, J.L.; Carass, A. Analysis of macular OCT images using deformable registration. *Biomed. Opt. Express* **2014**, *5*, 2196–2214. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Smith, S.M.; Brady, J.M. SUSAN—A new approach to low level image processing. *Int. J. Comput. Vis.* **1997**, *23*, 45–78. [\[CrossRef\]](#)
16. Wei, X.; Sui, R. A review of machine learning algorithms for retinal cyst segmentation on optical coherence tomography. *Sensors* **2023**, *23*, 3144. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Fang, L.; Li, S.; Nie, Q.; Izatt, J.A.; Toth, C.A.; Farsiu, S. Sparsity based denoising of spectral domain optical coherence tomography images. *Biomed. Opt. Express* **2012**, *3*, 927–942. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Fang, L.; Li, S.; McNabb, R.P.; Nie, Q.; Kuo, A.N.; Toth, C.A.; Izatt, J.A.; Farsiu, S. Fast acquisition and reconstruction of optical coherence tomography images via sparse representation. *IEEE Trans. Med. Imaging* **2013**, *32*, 2034–2049. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Bian, H.Y.; Duan, H.N.; Liu, L.; Yan, X.T.; Wu, X.P.; Xu, W.C.; He, N. Fast and high-resolution optical coherence tomography reconstruction via a resampling-free spectral fitting framework. *Opt. Lett.* **2026**, *51*, 2684–2687. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Jebur, R.S.; Zabil, M.H.B.M.; Hammood, D.A.; Al-Shammari, H.A. A comprehensive review of image denoising in deep learning. *Multimed. Tools Appl.* **2024**, *83*, 58181–58199.
21. Yang, Q.; Yan, P.; Zhang, Y.; Yu, H.; Shi, Y.; Mou, X.; Kalra, M.K.; Zhang, Y.; Sun, L.; Wang, G. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans. Med. Imaging* **2018**, *37*, 1348–1357. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Yuan, J.; Zhou, F.; Guo, Z.; Zhang, L.; Wang, Y.; Chen, X.; Liu, Y.; Zhang, J.; Li, Y.; Wang, Z. HCformer: Hybrid CNN-transformer for LDCT image denoising. *J. Digit. Imaging* **2023**, *36*, 2290–2305. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Ilesanmi, A.E.; Ilesanmi, T.O. Methods for image denoising using convolutional neural network: A review. *Complex Intell. Syst.* **2021**, *7*, 2179–2198. [\[CrossRef\]](#)
24. Gour, N.; Khanna, P. Speckle denoising in optical coherence tomography images using residual deep convolutional neural network. *Multimed. Tools Appl.* **2020**, *79*, 15679–15695.
25. Qiu, B.; Huang, Z.; Liu, X.; Zeng, S.; Meng, X.; Chen, X. Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function. *Biomed. Opt. Express* **2020**, *11*, 817–830. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Yu, X.; Li, M.; Ge, C.; Chen, X. A generative adversarial network with multi-scale convolution and dilated convolution Res-network for OCT retinal image despeckling. *Biomed. Signal Process. Control* **2023**, *80*, 104231. [\[CrossRef\]](#)
27. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
28. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013; p. 3.
29. Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; Aila, T. Noise2Noise: Learning image restoration without clean data. *arXiv* **2018**, arXiv:1803.04189.
30. Qiu, B.; Zeng, S.; Meng, X.; Huang, Z.; Liu, X.; Chen, X. Comparative study of deep neural networks with unsupervised Noise2Noise strategy for noise reduction of optical coherence tomography images. *J. Biophotonics* **2021**, *14*, e202100151. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Krull, A.; Buchholz, T.-O.; Jug, F. Noise2Void-learning denoising from single noisy images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2129–2137.

32. Nienhaus, J.; Matten, P.; Britten, A.; Bork, M.; Koch, P. Live 4D-OCT denoising with self-supervised deep learning. *Sci. Rep.* **2023**, *13*, 5760. [[CrossRef](#)] [[PubMed](#)]
33. Ma, Y.; Chen, X.; Zhu, W.; Xiang, D.; Yu, K.; Shi, F. Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN. *Biomed. Opt. Express* **2018**, *9*, 5129–5146. [[CrossRef](#)] [[PubMed](#)]
34. Wang, M.; Zhu, W.; Yu, K.; Chen, Z.; Shi, F.; Zhou, Y.; Ma, Y.; Peng, Y.; Bao, D.; Feng, S.; et al. Semi-supervised capsule cGAN for speckle noise reduction in retinal OCT images. *IEEE Trans. Med. Imaging* **2021**, *40*, 1168–1183. [[CrossRef](#)] [[PubMed](#)]
35. Chiu, S.J.; Li, X.T.; Nicholas, P.; Toth, C.A.; Izatt, J.A.; Farsiu, S. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. *Opt. Express* **2010**, *18*, 19413–19428. [[CrossRef](#)] [[PubMed](#)]
36. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [[CrossRef](#)] [[PubMed](#)]
37. Yang, B.; Huang, Y.; Xie, Y.; Li, J.; Jia, S.; Hao, Q. Time-step encoded high-frequency enhanced diffusion model for OCT retinal image denoising. *Biomed. Opt. Express* **2025**, *16*, 4571–4587. [[CrossRef](#)] [[PubMed](#)]
38. Fazekas, B.; Pinetz, T.; Aresta, G.; Emre, T.; Bogunović, H. GARD: Gamma-based anatomical restoration and denoising for retinal OCT. *arXiv* **2025**, arXiv:2509.10341.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.