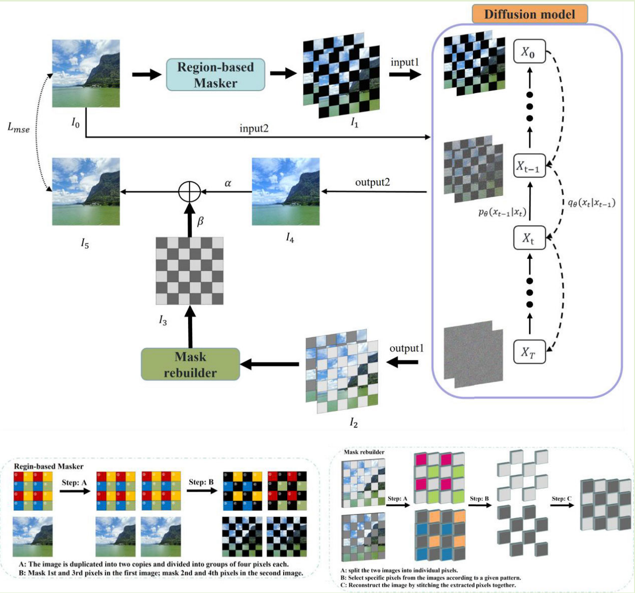
MR-DDPM: Unsupervised OCT Denoising Using Masked Region-Based DDPM

Zijian Li, Hua Wei, Yanshuo Li, Xiang He, Muhao Xu, Quanqing Xu, Kongzheng Yang,
Baochen Fu, Wei Yi, and Weiye Song

***Abstract*—Denoising is a commonly used technique in medical image preprocessing, particularly in the processing of optical coherence tomography (OCT) images. Due to the characteristics of OCT systems and their operational principles, the acquired images inevitably contain a superposition of various types of noise, including speckle noise, structural noise, and random noise, which consequently reduces the accuracy of subsequent segmentation tasks. Moreover, medical image processing places a high emphasis on image details, making it challenging to remove noise without causing excessive smoothing of the images. In addition, the difficulty in obtaining ground truth for medical images limits the application of many supervised learning methods. To address these issues, this article proposes an unsupervised masked region-based denoising diffusion probabilistic model algorithm. This algorithm employs pixel-level masking, which not only effectively enhances the removal of random noise but also enables the model to focus more on adjacent regions, thereby avoiding excessive alteration of the original image's details and edges. Furthermore, the method leverages information complementarity among data within the same group, effectively preventing information loss caused by masking. Experimental results demonstrate that compared to the baseline, the proposed algorithm increases the CNR to 3.77 with minimal reduction in the structural similarity index measure (SSIM). In addition, further segmentation experiments confirm the effectiveness of the denoising algorithm, achieving a 5.5% improvement in segmentation accuracy. This study presents the MR-DDPM algorithm, which improves the denoising quality of OCT images and consequently enhances segmentation accuracy, without the need for labeled data. This contribution is valuable to the advancement of medical image processing algorithms.**


Diffusion model
Region-based Masker
input1
X_0
L_{mse}
I_0
I_1
input2
α
output2
X_{t-1}
$p_\theta(x_{t-1}|x_t)$
$q_\theta(x_t|x_{t-1})$
β
I_5
I_4
X_t
I_3
Mask rebuilder
output1
X_T
I_2
Regin-based Masker
Step: A
Step: B
A: The image is duplicated into two copies and divided into groups of four pixels each.
B: Mask 1st and 3rd pixels in the first image; mask 2nd and 4th pixels in the second image.
Mask rebuilder
Step: A
Step: B
Step: C
A: split the two images into individual pixels.
B: Select specific pixels from the images according to a given pattern.
C: Reconstruct the image by stitching the extracted pixels together.

***Index Terms*—Denoising, denoising diffusion probabilistic model (DDPM), optical coherence tomography (OCT), unsupervised.**

I. Introduction

OPTICAL coherence tomography (OCT) [1], [2] is a technique that utilizes the principle of interferometry with low-coherence light to obtain high-resolution tissue cross-sectional images [3]. As a 3-D imaging modality, OCT has gained widespread application in medical and biological imaging [4], [5], [6], [7], [8], as well as in clinical diagnostic support, owing to its noninvasive nature, fast imaging speed, and high resolution. In particular, in the field of ophthalmology [9], the images obtained through OCT exhibit a distinct layered structure and contain rich pathological information, making it a valuable tool for assisting in the diagnosis of retinal diseases, glaucoma, and other conditions. However, due to the use of low-coherence light for image acquisition, combined with the inherent characteristics of its structure, OCT images inevitably suffer from various types of noise, including speckle noise, structural noise, and random noise [10]. Severe noise can significantly affect the accuracy of downstream tasks such as image segmentation and feature recognition [11]. Consequently, denoising is of paramount importance for the effective application of OCT technology. Currently, the

Received 8 May 2025; revised 25 June 2025; accepted 10 July 2025. Date of publication 23 July 2025; date of current version 15 September 2025. This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China, under Grant ZR2023QC262; in part by the National Natural Science Foundation of China under Grant 62205181; in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF017; and in part by Shandong Provincial Outstanding Youth Science Fund Project (Overseas) under Grant 2023HWYQ-023. The associate editor coordinating the review of this article and approving it for publication was Dr. Anuj Kumar Sharma. *(Corresponding author: Weiye Song.)*

The authors are with the School of Mechanical Engineering and the Key Laboratory of High Efficiency and Clean Mechanical Manufacture, Ministry of Education, Shandong University, Jinan 250061, China (e-mail: st_li2023@163.com; songweiye@sdu.edu.cn).

Digital Object Identifier 10.1109/JSEN.2025.3589585

primary method used in clinical practice to reduce noise involves multiple scans at the same location followed by averaging [12], [13]. While this approach can mitigate some noise, it requires the patient to maintain eye fixation for an extended period in order to acquire sufficient data, which presents a significant challenge in clinical settings [14]. As a result, numerous image processing methods have been proposed and studied for the purpose of denoising OCT images [15], [16].

In the early stages of OCT image denoising [17], most traditional denoising methods rely on filtering techniques and convolution operations. Buades et al. [18] proposed a nonlocal means algorithm that estimates pixel values based on pixel distances, effectively removing noise but with high computational complexity. Median filtering [19] replaces the center pixel with the median value, effective for salt-and-pepper noise but prone to blurring. The bilateral filter [20] reduces noise by considering normalization factors, spatial, and distance weights but has a high computational cost and may retain excessive detail in noisy regions. Due to the limitations of spatial filtering in edge preservation, wavelet-domain methods have gained popularity. Donoho [21] introduced compressed sensing and demonstrated wavelet-based denoising in sparse domains, effective for high noise and low sampling, requiring extensive optimization. Shankar [22] proposed the Visushrink filter, which uses a threshold proportional to the noise standard deviation, effective for additive noise but not for speckle noise. Despite advances, challenges remain in effectively addressing speckle noise. These methods aimed to minimize the mean square error between the original and noisy images. However, such straightforward techniques often struggle to achieve satisfactory results in the complex tasks associated with OCT image processing.

With advancements in computer vision, convolutional neural networks (CNNs), attention mechanisms, and transformer models have been applied to denoising tasks. Zhang et al. [23] investigated feed-forward denoising CNNs (DnCNNs), using residual learning and batch normalization to accelerate training and improve performance for blind Gaussian denoising. However, the lack of paired training data limits further optimization. Chen et al. [24] proposed a two-step framework that leverages GANs for noise modeling, generating noise samples in the first step and using them to train the CNN denoising network in the second. Armanious et al. [25] introduced MedGAN, a framework combining adversarial learning with novel loss functions for medical image translation. To address performance saturation in deep CNN training, BRDNet [26] integrates batch renormalization to mitigate covariate shift and small batch issues. Jiang et al. [27] proposed MalleConv, using dynamic filters generated by efficient predictors to adapt to varying visual patterns in natural images, though bilinear interpolation may introduce artifacts or blurring. Yin and Ma [28] introduced CSformer, a transformer-based denoising method that employs multiscale feature extraction and fusion, enhancing self-attention, but its ability to recover small-scale textures under high noise still needs improvement. Zhao et al. [29] proposed a hybrid denoising model, TECDNet, combining a transformer encoder and a CNN decoder, significantly reducing computational complexity, but the quadratic complexity of self-attention may limit scalability. Fan et al. [30] introduced SUNet, integrating the Swin transformer with UNet, outperforming previous CNN- and UNet-based methods. While these methods demonstrate promising performance, they generally require low-noise images as references for training, which may not always be readily available in clinical medical settings [31].

To address the challenge of obtaining ground-truth data [32], Lehtinen [33] demonstrated that noise data from degraded images alone can restore clear images, introducing the Noise2Noise (N2N) strategy, which requires no clean target images but needs a large number of noisy image pairs. Krull et al. [34] further developed N2N with the Noise2Void (N2V) strategy, where masking pixels and using surrounding information allow training on noisy data without clean target images, though proper selection of masked areas is crucial. Similarly, Batson and Royer [35] proposed Noise2Self (N2S), an unsupervised strategy that assumes noise independence and signal correlation, learning self-similarity in local regions. Furthermore, self-supervised learning has emerged as a promising direction for unsupervised denoising. Wang et al. [36] proposed the Blind2Unblind framework that constructs visible blind spots for self-supervised learning while preserving noise statistics. This paradigm was extended by Lee et al. [37] through the AP-BSN framework using asymmetric projection denoising and blind-spot networks, achieving better performance in complex noise scenarios. Jang et al. [38] further advanced this direction at ICCV 2023 by introducing downsampling invariance loss combined with conditional blind-spot networks to better preserve spatial feature consistency. However, it cannot handle all types of noise. Prakash et al. [39] introduced DivNoising, a method based on fully convolutional variational autoencoders, overcoming the challenge of selecting a single denoised solution. To reduce the computational cost of generating large samples in unsupervised methods, Salmon and Krull [40] proposed a deterministic network strategy that directly predicts the central tendency alongside the VAE. Diffusion models, due to their strong noise estimation capabilities, have become popular in denoising tasks. Zhenjie et al. [41] proposed a doubly physical-regularized denoising diffusion model, which functions as a filter at each layer to remove noise at different scales but requires large datasets for training. Kulikov et al. [42] introduced SinDDM, a framework for training denoising diffusion models on a single image, combining its flexibility with the multiscale structure of SinGAN to effectively guide the denoising process. In the domain of medical imaging, specialized solutions have been developed for OCT speckle noise. Yu et al. [43] adapted the Blind2Unblind paradigm for OCT despeckling, achieving effective noise reduction while preserving tissue structures through self-supervised learning. Li et al. [44] proposed a clean-data-free speckle reduction method demonstrating superior structural preservation in retinal OCT images. Zhou et al. [3] made a breakthrough by integrating transformer architecture with nonlocal means, using self-attention mechanisms to capture global contextual information while maintaining high-contrast features.

While these advancements in unsupervised learning and diffusion models have demonstrated remarkable progress, three critical challenges persist in OCT image denoising: 1) existing diffusion-based methods often sacrifice fine-grained details when handling spatially correlated noise patterns inherent in OCT scans; 2) the inherent conflict between blind-spot training strategies and the global noise modeling requirements of diffusion processes remains unresolved; and 3) most current approaches inadequately address the unique layered structure and anisotropic noise distribution characteristics of OCT images.

In this article, we propose a denoising method called MR-DDPM, which is based on diffusion models and represents an enhanced unsupervised masked region-based DDPM algorithm. As an unsupervised approach, MR-DDPM effectively circumvents the challenge of obtaining ground-truth data. The method incorporates the concept of blind-spot masking, employing a pixel-level occlusion strategy during training. This not only improves the removal of random noise but also directs the model's focus to neighboring regions, preventing oversmoothing and preserving the original image details and edges. In addition, the approach facilitates mutual information exchange between occluded image groups, effectively mitigating information loss due to occlusion. To further address these challenges, we modify the U-Net architecture by introducing a time-step embedding transformer module, which is parallelized to the deepest layer of the U-Net network. This modification enables the network to retain its fine-grained modeling capabilities while incorporating a global receptive field. In addition, by integrating the time-step parameter of the diffusion model, the network becomes more adaptive throughout the diffusion process, thereby enhancing its ability to analyze and handle noise. The key contributions of this work are given as follows.

1) We introduce the MR-DDPM algorithm, a denoising method for OCT images that significantly reduce noise while ensuring the preservation of essential edge and detail information without relying on ground-truth data.
2) We modify the Unet network by incorporating a time-step embedding transformer to better suit the noise characteristics of OCT images. This enhancement enables the network to maintain fine-grained modeling capabilities while possessing a certain global receptive field. The time-step embedding transformer module explicitly encodes the noise time-step information from the diffusion model into dynamic spatial attention biases, replacing the traditional static positional encoding. This helps guide the model to focus on global structures during the early denoising stages (high noise levels) and on local details during the later denoising stages (low noise levels), thereby further improving its ability to analyze and remove noise.
3) Experimental results demonstrate that the proposed MR-DDPM algorithm significantly improves image quality in quantitative metrics such as SNR, CNR, and structural similarity index measure (SSIM), as well as in visual assessments. In addition, further segmentation experiments on the denoised images verify that the algorithm enhances the performance of subsequent segmentation tasks.

II. Method

A. Diffusion Models

MR-DDPM is based on the core principles of DDPM. The main theoretical concepts and mathematical foundations of DDPM are briefly introduced in the following.

Diffusion models generally consist of two stages: diffusion and denoising. During the diffusion stage, noise is progressively added to the input image at different scales, gradually degrading the image until it is transformed into a Gaussian noise distribution, as shown in the following equation:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}z, z \sim N(0,1) \tag{1}$$

where x_t represents the result of the diffusion process at the time step t. $\alpha_t = 1 - \beta_t$, which represents the variance of the noise added at each diffusion step. x_{t-1}, which is similar to x_t, represents the image at time step $t-1$. Finally, z represents the random noise.

Next, by applying the reparameterization trick, the following expression can be derived:

$$q(x_t|x_{t-1}) \sim N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I). \tag{2}$$

Furthermore, if x_0 is known, it is possible to derive x_t at any given time step

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}z, z \sim N(0,1) \tag{3}$$

where $\bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i$. Similarly, the conditional probability distribution of x_t can also be derived as follows:

$$q(x_t|x_0) \sim N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \tag{4}$$

The diffusion process involves the gradual addition of noise to the data, while the reverse process corresponds to a denoising procedure. If the true distribution $q(x_t|x_0)$ of each step in the reverse process is known, it becomes possible to generate a real sample by progressively denoising from an initial random noise $x_T \sim N(0,1)$. Hence, the reverse process can be interpreted as the data-generation process

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T}p_\theta(x_{t-1}|x_t) \tag{5}$$

where $p_\theta \sim N(0,1)$ By leveraging the Bayesian theorem and the Markov chain property, with the addition of infinitesimal Gaussian noise at each step, the following equation is derived:

$$p_\theta(x_{t-1}|x_t, x_0) \sim N(x_{t-1}, \mu_\theta, \sigma_\theta^2 I) \tag{6}$$

where

$$\mu_\theta = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{7}$$

$$\sigma_\theta^2 = \frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t}. \tag{8}$$

Subsequently, the new expression for the mean can be obtained through iteration

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z_t\right). \tag{9}$$

Finally, the noise z_t in formula (9) can be predicted using the neural network, which enables the estimation of the mean μ.

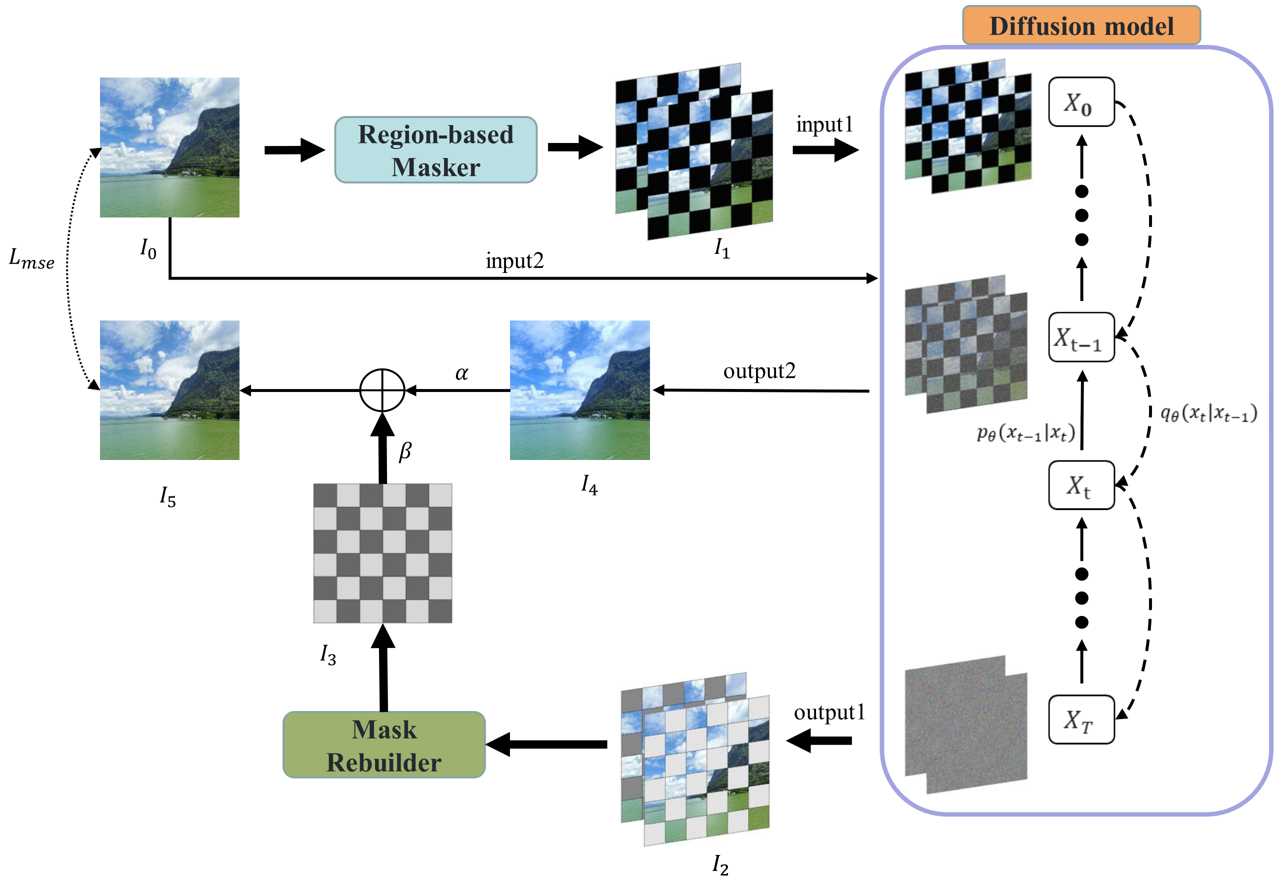


Fig. 1. MR-DDPM algorithm framework diagram. Branch 1: from input1 to output1; Branch 2: from input2 to output2. The results of the two branches are weighted and averaged to produce the network output I_5.

B. MR-DDPM

To address the issues of image blurring, excessive smoothing, or incomplete noise removal caused by denoising algorithms, this article proposes masked region-based DDPM. The strategy of MR-DDPM is illustrated in Fig. 1.

The clearest possible B-scan images are first obtained as inputs. As shown in Fig. 1, the network consists of two branches: Branch 1 (from input1 to output1) and Branch 2 (from input2 to output2), represented by thick and thin arrows, respectively. The outputs of both branches are weighted and averaged to produce the network output I_5, which is then used to compute the loss relative to the original image I_0.

In Branch 1, the input image is initially processed through the region-based masker module. The functionality of the region-based masker module is illustrated in Fig. 2, where the input image is divided into groups of four pixels, each assigned a label. In Step A, the image is duplicated into two copies. In Step B, all pixels labeled as 1 and 3 in the first image are masked, while all pixels labeled as 2 and 4 in the second image are masked. The image I_1 is then passed through the diffusion model to generate output I_2, which is subsequently processed by the mask rebuilder to produce I_3. The functionality of the mask rebuilder module is shown in Fig. 3, where in Step A, the input image is segmented into individual pixels; Step B selects specific pixels as shown in the figure; and in Step C, the selected pixels are reassembled into a new image.

In Branch 2, the input image I_0 is directly passed through the diffusion model to generate I_4. The outputs from both branches, I_3 and I_4, are then weighted and averaged to obtain the final network output I_5. The coefficients α and β represent the weighting factors for the I_4 image and the I_3 image, respectively. Specifically, the I_3 image undergoes pixel-level feature filtering via the region-based masker module, so increasing the coefficient β enhances the denoising performance of the final output. In contrast, the unmasked I_4 image preserves original details, and a larger α value improves SSIM of the output. Based on experimental validation, the balanced setting of $\alpha = 0.5$ and $\beta = 0.5$ was chosen to achieve an optimal tradeoff between denoising capability and structural preservation. This masking approach enhances the model's capability to infer the current position based on adjacent pixels, which not only improves random noise removal performance but also significantly boosts the structural similarity between samples and ground truth. However, such masking may also result in partial information loss and neglect of global pixel distribution due to localized subregion focus. To address this limitation, we designed an output2 pathway that supplements the missing information by fusing the masked image (I_3 from output1) with the complementary image (I_4 from output2) via weighted summation. This dual-pathway architecture ensures both localized detail preservation and global pixel distribution consideration. In addition, the set of images generated by

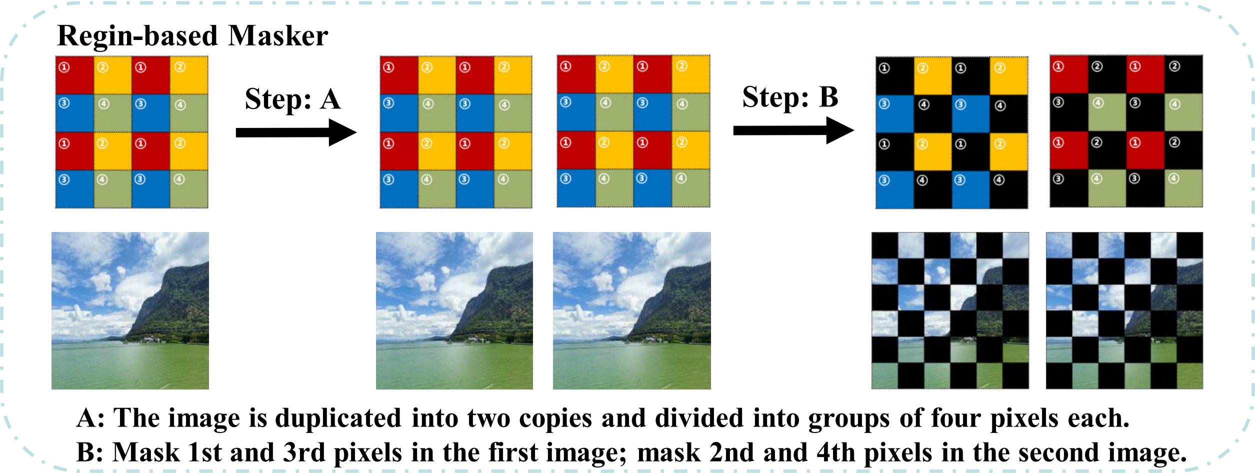

Regin-based Masker
Step: A
Step: B
A: The image is duplicated into two copies and divided into groups of four pixels each.
B: Mask 1st and 3rd pixels in the first image; mask 2nd and 4th pixels in the second image.

Fig. 2. Introduction to the region-based masker module.

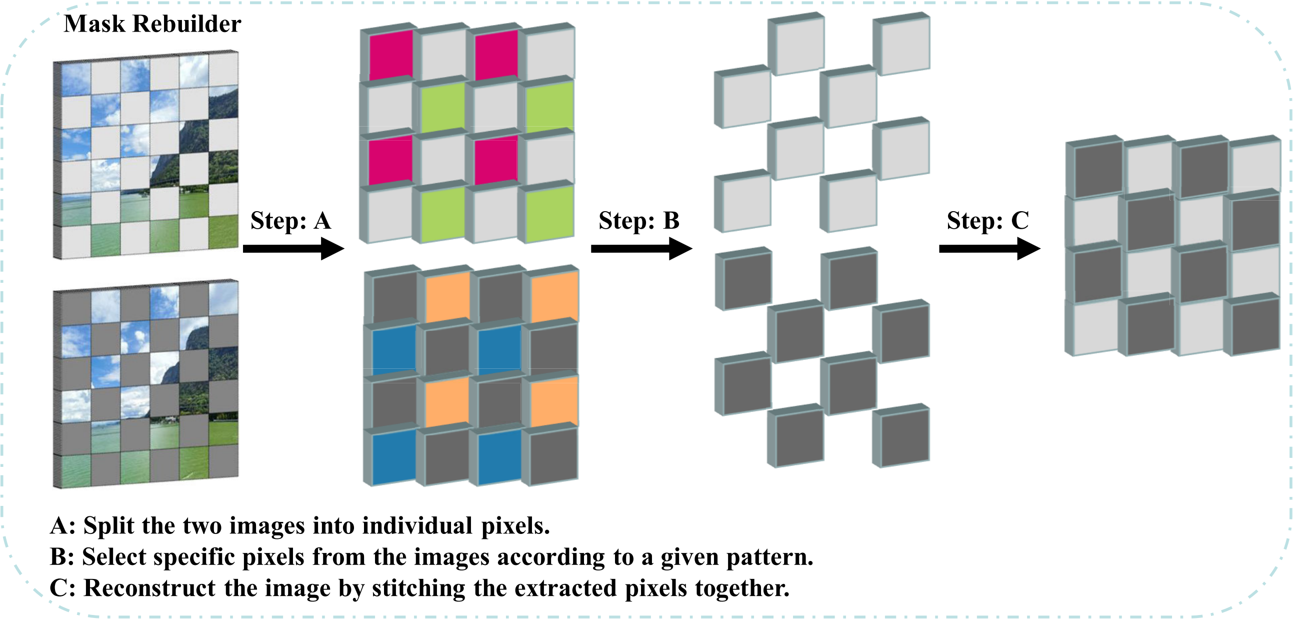

Mask Rebuilder
Step: A
Step: B
Step: C
A: Split the two images into individual pixels.
B: Select specific pixels from the images according to a given pattern.
C: Reconstruct the image by stitching the extracted pixels together.

Fig. 3. Introduction to the mask rebuilder module.

the region-based masker module further mutually reinforces information through their diversified spatial coverage, thereby mitigating occlusion-induced information deficits and enhancing overall reconstruction fidelity.

C. Time-Transformer Block

The diffusion model typically uses the U-Net architecture to fit parameters such as the mean and variance during the reverse process. For the MR-DDPM strategy proposed in this article, we also enhance the Unet network by introducing a time-step embedding transformer module. In our approach, the time-step parameter t does not refer to the dynamic temporal information inherent to OCT images (e.g., the temporal dimension in sequential imaging) but rather represents the *time-step encoding* of the diffusion process in the diffusion model. Specifically, in the diffusion framework, t denotes the current stage of the diffusion process (i.e., the tth step among T steps from the original image to pure noise). This time step is embedded into the network via positional encoding, allowing the model to perceive the current noise level. The time-transformer module facilitates interaction between the temporal encoding and spatial features through a cross-attention mechanism within the transformer architecture. This enables the model to focus on global structure reconstruction during early high-noise stages and refine local details in later low-noise stages. Such a design allows the model to dynamically adapt denoising strategies, addressing potential oversmoothing or underdenoising issues inherent in static architectures like the conventional U-Net. As shown in Fig. 4, at the deepest layer of the U-Net network (denoted position 1), a time-step embedding transformer module is added. The output of the transformer module and the original CNN module of the network is combined through a weighted average before being passed to the next layer. This is a dynamic time-conditioned positional encoding, which explicitly encodes the noise time-step information (timestep t) from the diffusion model into dynamic spatial attention biases, replacing the traditional static positional encoding. This module helps guide the model to focus on global structures during the early denoising stages (high noise levels) and on local

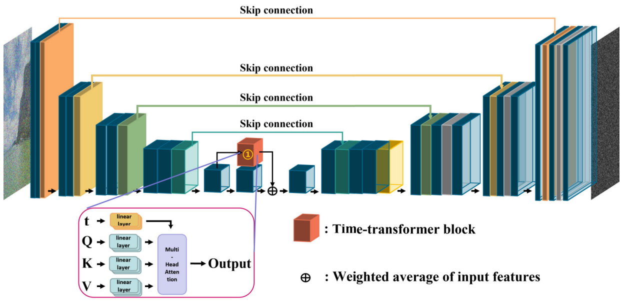


Fig. 4. Improvements of U-Net.

details during the later denoising stages (low noise levels). First, the time step t is mapped to a high-frequency/low-frequency mixed embedding vector

$$e_t = \text{MLP}\left(\left[\sin\left(10^{4i/d}\cdot t\right), \cos\left(10^{4i/d}\cdot t\right)\right]_{i=0}^{d/2}\right). \quad (10)$$

This embedding is injected into the attention layer as a dynamic bias

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q + e_t\mathbf{W}_Q^t, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K + e_t\mathbf{W}_K^t. \quad (11)$$

The attention weights are then computed as

$$\text{Attention} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}_t\right)\mathbf{V}. \quad (12)$$

The addition of the transformer module at the deepest layer not only minimizes computational resource consumption but also combines the fine-grained modeling ability of the CNN with the global receptive field characteristics of the transformer.

III. Experiments

To evaluate the MR-DDPM algorithm and the improved U-Net architecture, experiments are conducted on OCT B-Scan images. The obtained results are compared and analyzed with several methods, including median filtering, Blind2Unblind, BM3D, and Mgan.

A. Dataset

OCT2017 [45] is a publicly available dataset that contains over 80 000 OCT images, including images of various sizes and classifications of both diseased and normal conditions. In this article, 600 high-quality B-Scan images with a size of 496 × 496 are selected from this dataset for the experiments.

B. Implementation Details

All experiments in this article are tested and trained on an NVIDIA A100 Tensor Core GPU. The B-Scan images are normalized to the range of $[-1, 1]$. In the diffusion model, the number of diffusion steps is set to $T = 1000$, meaning that the input is gradually corrupted into Gaussian noise over 1000 steps. The training employs the neural network shown in Fig. 4, with a batch size of 2. The learning rate is initialized to 0.001 and decays every 5 epochs. The Adam optimizer is used for training. The entire training process lasts for 600 epochs to ensure the accuracy of the results.

C. Denoising Results and Analysis

After 600 epochs of training, the model with the lowest loss is saved for inference. The final results are compared with those obtained using several methods, including median filtering, Blind2Unblind, BM3D, and Mgan, as shown in Fig. 5.

As shown in the figure, the original image contains significant random noise, speckle noise, and structural noise due to the inherent properties of OCT and the device structure, which severely interferes with subsequent tasks. Mean filtering, a classical image denoising method, performs well in removing uniform noise; however, it tends to cause blurring of the image. As demonstrated in the figure, while the Blind2Unblind method achieves significant noise reduction, its training strategy involving continuous large-scale occlusion exhibits limitations. During this process, the model is forced to infer occluded information based on adjacent regions' features. When processing OCT images, where signal and noise features are highly overlapped, the true details may be erroneously classified as noise and excessively suppressed. This processing approach leads to unnatural oversmoothing artifacts, resulting in the distortion of critical pathological information. Similarly, BM3D is more effective for Gaussian noise removal, but its performance is limited when dealing with the diverse types of noise in OCT images, which exhibit different spatial and frequency distributions. Although Mgan can generate visually realistic denoised images, it sometimes overrepairs fine details or introduces artifacts, resulting in the loss of true image details and textures.

The MR-DDPM algorithm demonstrates the best performance, as shown in Fig. 5. It is commonly assumed that images dominated by random noise consist of a background and noise, where the background is continuous and the noise is independently distributed. Therefore, the pixel-level masking approach not only significantly enhances the model's ability to remove random noise but also enables the model to focus on preserving the continuity of the image by attending to adjacent regions. In addition, the information within the same group of masked images can complement each other, thereby preventing information loss, a common issue in other similar masking-based algorithms. The improvements made to the network also enable it to retain fine-grained modeling capabilities while incorporating a global receptive field.

D. Evaluation Metrics

This article adopts an unsupervised learning approach, with CNR and SNR selected as evaluation metrics to assess the denoising model's performance. Since OCT images are a type of medical imaging, denoising must not only remove noise but also preserve the original structure of the image. Therefore, the SSIM is also included as an additional evaluation metric. In the absence of ground truth for comparison, CNR and SNR are computed based on the noise and signal regions of the comparison images. These regions are determined using a thresholding method, where two 40 × 60 areas are selected. The metrics presented in Table I are then obtained by averaging over multiple iterations. It can be observed that the MR-DDPM algorithm not only performs the best in terms of the visual quality of the images but also demonstrates supe-

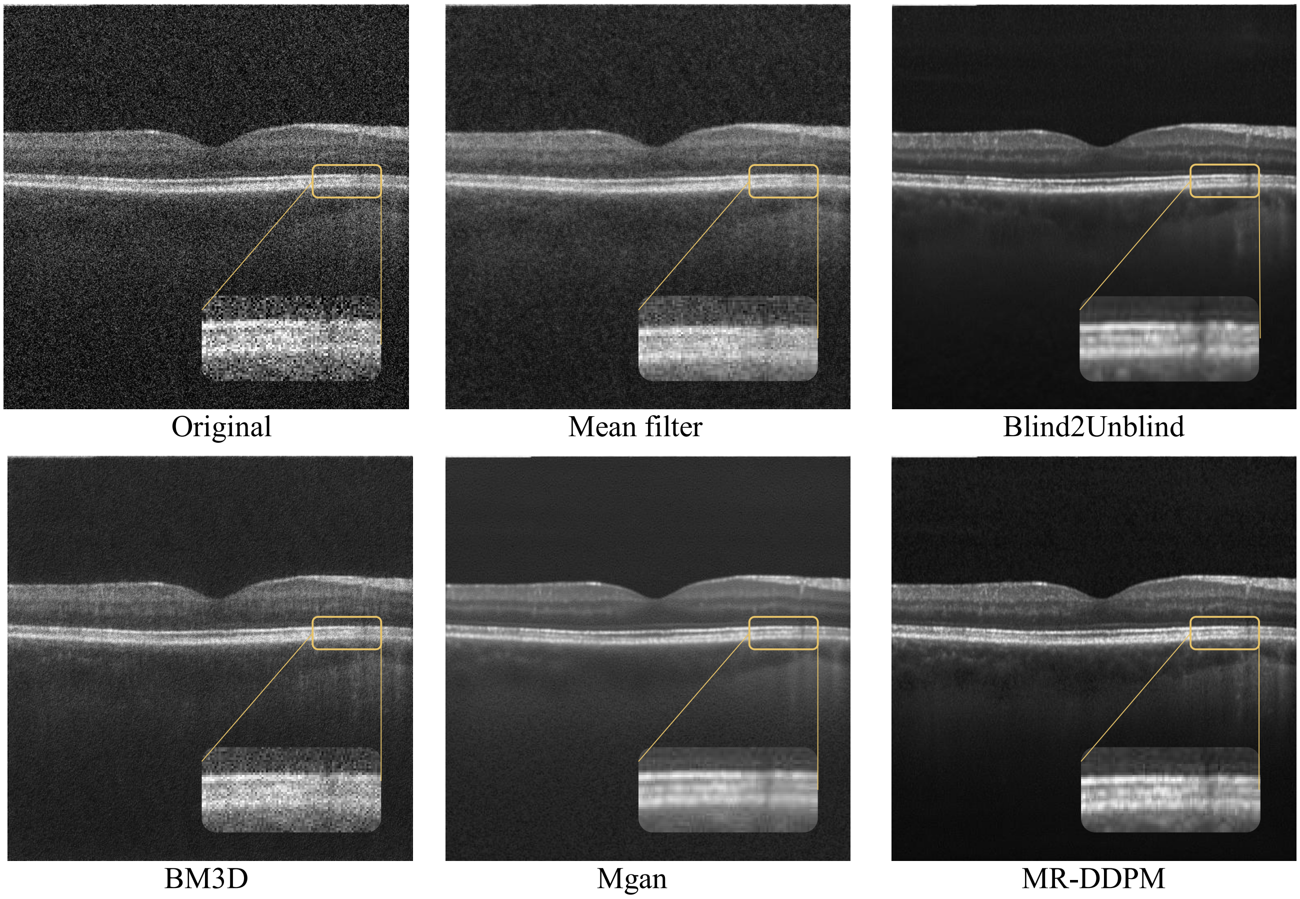


Fig. 5. Comparison of denoising performance across different algorithms.

TABLE I
SNR, CNR, AND SSIM RESULTS OF DIFFERENT DENOISING ALGORITHMS

Methods	CNR (dB)	SNR (dB)	SSIM
Original	1.51	6.22	1.00
Mean filter	1.59	6.31	0.72
Blind2Unblind	3.51	12.1	0.74
BM3D	2.56	10.11	0.85
Mgan	3.48	12.29	0.81
MR-DDPM	3.77	12.41	0.91

TABLE II
MIOU AND ACC RESULTS OF DIFFERENT DENOISING ALGORITHMS

Methods	MIOU(%)	ACC(%)
Original	79.1	87.1
Mean filter	72.8	81.7
Blind2Unblind	82.9	91.3
BM3D	82.3	90.1
Mgan	84.8	90.7
MR-DDPM	85.9	92.6

rior performance across all evaluation metrics. The detailed calculation formulas for CNR, SNR, and SSIM are translated as follows:

$$\mathrm{CNR} = \frac{\mu_{\text{signal}} - \mu_{\text{background}}}{\sigma_{\text{background}}} \tag{13}$$

$$\mathrm{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{background}}} \tag{14}$$

where μ_{signal} represents the mean pixel value of the signal region. $\mu_{\text{background}}$ represents the mean pixel value of the background region. $\sigma_{\text{background}}$ represents standard deviation of pixel values in the background region

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{15}$$

where σ_x^2 and σ_y^2 represent local variances of x and y. σ_{xy} represents the covariance between x and y.

Denoising is a preprocessing task in OCT retinal image processing, with the ultimate goal of enhancing the performance of subsequent tasks. The results from Fig. 5 are then used for retinal layer segmentation. The retinal segmentation is also performed using a deep learning approach, specifically the method developed in previous work [46], [47] by the laboratory. The final retinal layer segmentation results are shown in Fig. 6.

As shown in Fig. 6, the original image suffers from segmentation errors in certain regions due to excessive noise, leading to incorrect pixel classification. The mean filter, on the other hand, results in significant segmentation errors due to boundary blurring. The oversmoothing artifacts caused by the Blind2Unblind algorithm also result in a decline in segmentation accuracy. The performance of the remaining three algorithms is further analyzed through the quantitative metrics presented in Table II. The calculation methods for MIOU and ACC indicators are given as follows:

$$\mathrm{IoU}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \quad \mathrm{MIOU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{IoU}_i \tag{16}$$

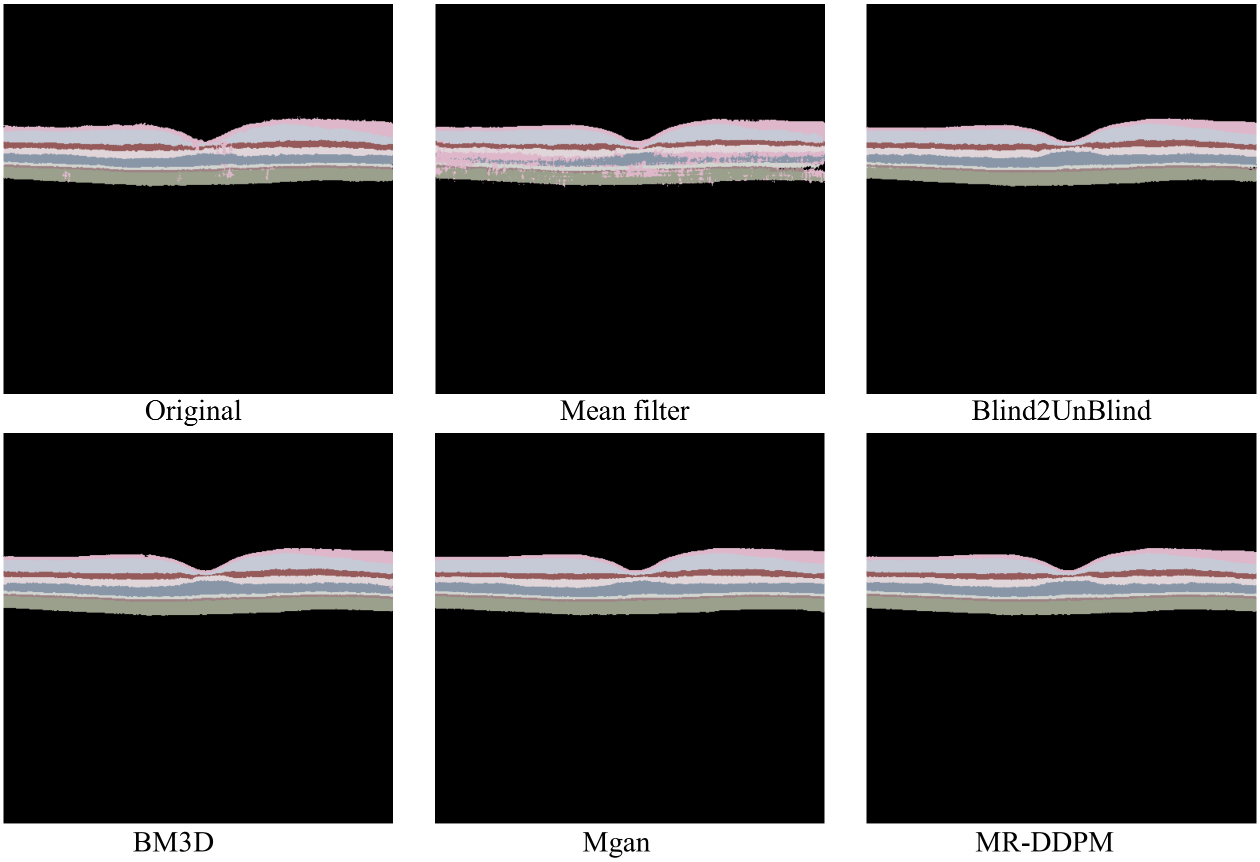


Fig. 6. Comparison of segmentation results across different algorithms.

TABLE III
Ablation Experiment

Methods	CNR (dB)	SNR (dB)	SSIM
MR-DDPM + Time-transformer	3.77	12.41	0.91
MR-DDPM	3.71	12.35	0.88
DDPM	3.58	12.27	0.82

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{17}$$

where TP_i represents true positive pixels of class i. FP_i represents false positive pixels of class i. FN_i represents false negative pixels of class i. As shown in Table II, compared to the BM3D and Mgan algorithms, MR-DDPM achieves the best results in terms of both MIOU and ACC metrics. The segmentation results further validate the superior performance of the MR-DDPM denoising algorithm.

E. Ablation Study

We conduct ablation tests to verify the effectiveness of the key components of our designed model. This article primarily proposes an improved U-Net method with a parallel time-transformer block at the deepest layer of the network, along with the region-based masker module. Next, we sequentially remove the improved U-Net and region-based masker modules to observe the impact on the denoising metrics, as shown in Table III.

F. Discussion

As shown in Fig. 5, we demonstrated denoising performance comparisons across multiple algorithms. The noise configuration in Fig. 5 represents a common pixel-level noise type in OCT imaging caused by light source fluctuations, where the critical challenge lies in the near-identical intensity distributions between noise and signal components. The proposed MR-DDPM algorithm exhibited outstanding denoising performance in this scenario while preserving structural similarity, as evidenced by the SSIM metric. However, Fig. 5 utilizes healthy retinal images where noise-signal boundaries remain relatively distinct. In contrast, Fig. 7 presents pathological retinal images from diseased patients, where signal-noise boundaries are far less distinguishable due to the complex pathological features. Notably, the MR-DDPM algorithm maintained superior denoising performance even under these challenging conditions, demonstrating its robustness across diverse clinical scenarios.

The MR-DDPM adopts noise prediction rather than direct image reconstruction due to three critical advantages. First, parameterizing the model to estimate $\epsilon_\theta(x_t, t)$ establishes mathematical symmetry with the forward diffusion process $q(x_t|x_{t-1})$, enabling stable training via MSE loss, which directly measures prediction accuracy at each noise level. Second, the incremental noise addition through Markov chains allows progressive denoising: each step's conditional probability $p_\theta(x_{t-1}|x_t)$ only needs to reverse the current noise perturbation, avoiding the instability of directly modeling complex image distributions. Third, this framework naturally handles multimodal medical noise (speckle, structural, and stochastic) by decomposing the denoising task into manageable stochastic steps. The masked region mechanism further refines this process by localizing noise estimation to critical

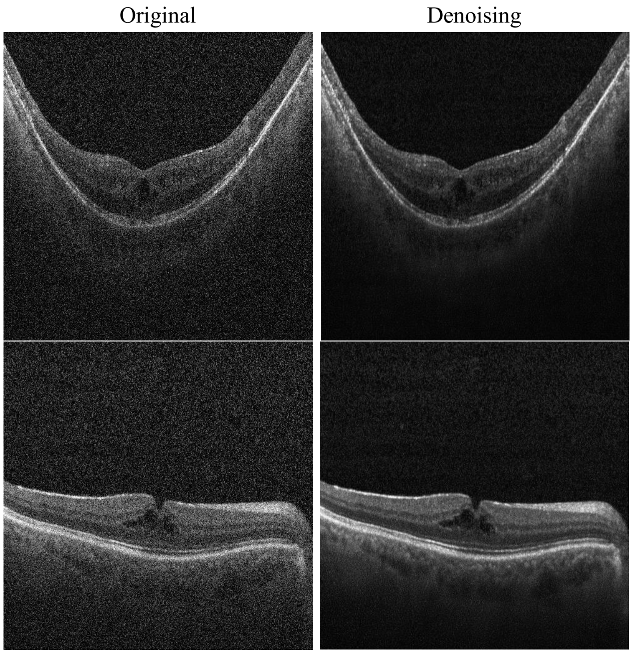


Fig. 7. Denoising performance of retinal diseases.

anatomical areas while preserving global tissue structure through the diffusion's inherent spatial coherence.

As illustrated in Fig. 8, the diffusion model employed in this study consists of two core stages: the forward process (forward process, orange box) and reverse process (reverse process, green box). During the forward stage, the original OCT image I_0 undergoes $T = 1000$ diffusion steps of gradual noise injection, ultimately transforming into a pure noise image. The reverse stage achieves progressive denoising through an inverse process from $t = 1000$ to $t = 1$. Notably, the image generated at $t = 1$ exhibits more pronounced denoising effects compared to the original I_0. This is attributed to the parametric capacity of the model to progressively predict the noise distribution, acquired through training with parameterized reverse mappings. Specifically, during inference, the original image is first mapped to the noise space where inherent noise components are progressively incorporated into the artificially added noise. In the reverse recovery process, the model progressively extracts and preserves critical structural features through backward diffusion steps, ultimately aligning the generated image distribution with the ideal clean image distribution. It is important to emphasize that training data quality significantly impacts model performance. Theoretically, using completely noise-free training images would yield optimal denoising results. However, in OCT imaging applications, obtaining such pristine images is technically challenging due to inherent device limitations. To address this limitation, this study will subsequently develop a dedicated imaging system to implement a multiple-scan averaging method. This approach aims to acquire image data approaching the ideal clean state, thereby further enhancing the denoising capability of the diffusion model.

The innovative aspects of our algorithm primarily focus on the training strategies and neural network architecture optimization. However, during the inference phase, the conventional approach of DDPM remains in use. Notably, the relatively slow inference speed remains a prevalent challenge for diffusion models. Therefore, in future experiments, we aim to further optimize the network's inference efficiency through targeted improvements.

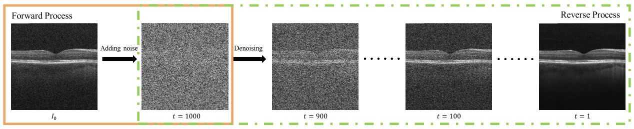


Fig. 8. Visualization display of denoising process.

The article is theoretically grounded in the principles of blind denoising algorithms such as N2V, where local context learning enables the model to infer pixel values based on adjacent regions. By masking pixels at the training stage, the model learns to disentangle noise from structural patterns while preserving edges and details. Our method extends this idea by leveraging region-based masking and intersample complementarity, which further enhance structural preservation. The improvement in SSIM is empirically validated through ablation studies, demonstrating that our approach retains more structural information compared to baselines.

IV. Conclusion

To enhance the denoising of OCT B-Scan images while preserving fine details and avoiding excessive smoothing, this article proposes a region-based masking approach. This method applies pixel-level masking to the input image according to a predefined rule. It effectively removes random noise and enables the model to focus on adjacent regions to infer the value of the current pixel. In addition, the information within the same group of masked images can complement each other, preventing information loss caused by the masking process. Furthermore, an improvement to the U-Net architecture is proposed, where a time-transformer block is integrated in parallel at the deepest layer of the network. This modification allows the model to maintain the fine-grained modeling capabilities of CNNs while leveraging the global receptive field of transformers. In addition, by integrating the time-step parameter of the diffusion model, the network becomes more adaptive throughout the diffusion process, thereby enhancing its ability to analyze and handle noise. Experimental results, both qualitative and quantitative, in terms of SNR, CNR, and SSIM, demonstrate the superiority of the proposed algorithm. Further segmentation experiments, based on the denoising results, also validate the effectiveness of MR-DDPM. Thus, the MR-DDPM denoising algorithm presented in this article significantly contributes to the preprocessing of OCT images, improving the accuracy of subsequent tasks such as segmentation, and holds substantial implications for the advancement of OCT technology.

References

[1] L. Tao, J. Qian, C. Gong, D. Zhang, and Y. Luo, "Cross-domain retinopathy classification based on optical coherence tomography sensors via domain adversarial graph convolutional network," *IEEE Sensors J.*, vol. 25, no. 2, pp. 3473–3483, Jan. 2025.

[2] D. Huang et al., "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, Nov. 1991.

[3] Q. Zhou, M. Wen, B. Yu, C. Lou, M. Ding, and X. Zhang, "Self-supervised transformer based non-local means despeckling of optical coherence tomography images," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104348.

[4] Q. Chen et al., "Automated choroid segmentation of three-dimensional SD-OCT images by incorporating EDI-OCT images," *Comput. Methods Programs Biomed.*, vol. 158, pp. 161–171, May 2018.

[5] X. Li et al., "Multi-scale reconstruction of undersampled spectral–spatial OCT data for coronary imaging using deep learning," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 12, pp. 3667–3677, Dec. 2022.

[6] S. Liao et al., "Dual-spatial domain generalization for fundus lesion segmentation in unseen manufacturer's OCT images," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 9, pp. 2789–2799, Sep. 2024.

[7] V. Das, S. Dandapat, and P. K. Bora, "Unsupervised super-resolution of OCT images using generative adversarial network for improved age-related macular degeneration diagnosis," *IEEE Sensors J.*, vol. 20, no. 15, pp. 8746–8756, Aug. 2020.

[8] M. Rowedder, T. Kepp, T. Neumann, H. Sudkamp, G. Hüttmann, and H. Handels, "Denoising of home OCT images using noise2noise trained on artificial eye data," in *Medical Imaging 2024: Image Processing*, vol. 12926. Bellingham, WA, USA: SPIE, 2024, pp. 583–589.

[9] G. R. Wilkins, O. M. Houghton, and A. L. Oldenburg, "Automated segmentation of intraretinal cystoid fluid in optical coherence tomography," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 1109–1114, Apr. 2012.

[10] X. Wei, X. Liu, A. Yu, T. Fu, and D. Liu, "Clustering-oriented multiple convolutional neural networks for optical coherence tomography image denoising," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–5.

[11] B. N. Anoop et al., "A cascaded convolutional neural network architecture for despeckling OCT images," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102463.

[12] J. J. Rico-Jimenez, D. Hu, E. M. Tang, I. Oguz, and Y. K. Tao, "Real-time OCT image denoising using a self-fusion neural network," *Biomed. Opt. Exp.*, vol. 13, no. 3, pp. 1398–1409, 2022.

[13] G. Ni et al., "Toward ground-truth optical coherence tomography via three-dimensional unsupervised deep learning processing and data," *IEEE Trans. Med. Imag.*, vol. 43, no. 6, pp. 2395–2407, Jun. 2024.

[14] Z. Chen, Z. Zeng, H. Shen, X. Zheng, P. Dai, and P. Ouyang, "DN-GAN: Denoising generative adversarial networks for speckle noise reduction in optical coherence tomography images," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101632.

[15] Y. Huang, N. Zhang, and Q. Hao, "Real-time noise reduction based on ground truth free deep learning for optical coherence tomography," *Biomed. Opt. Exp.*, vol. 12, no. 4, pp. 2027–2040, 2021.

[16] Y. Zhou et al., "Speckle noise reduction for OCT images based on image style transfer and conditional GAN," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 139–150, Jan. 2022.

[17] W. Li, J. Zou, N. Meng, Y. Fang, and Z. Huang, "Evaluation of different denoising algorithms for OCT image denoising," in *Optics in Health Care and Biomedical Optics X*, vol. 11553. Bellingham, WA, USA: SPIE, 2020, pp. 352–357.

[18] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 60–65.

[19] S. Perreault and P. Hebert, "Median filtering in constant time," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2389–2394, Sep. 2007.

[20] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Found. Trends Comput. Graph. Vis.*, vol. 4, no. 1, pp. 1–73, 2009, doi: 10.1561/0600000020.

[21] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[22] P. M. Shankar, "Quantitative measures of boundary and contrast enhancement in speckle reduction in ultrasonic B-mode images using spatial Bessel filters," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 56, no. 10, pp. 2086–2096, Oct. 2009.

[23] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[24] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3155–3164.

[25] K. Armanious et al., "MedGAN: Medical image translation using GANs," *Computerized Med. Imag. Graph.*, vol. 79, Jan. 2020, Art. no. 101684. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611119300990

[26] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Netw.*, vol. 121, pp. 461–473, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019302394

[27] Y. Jiang, B. Wronski, B. Mildenhall, J. T. Barron, Z. Wang, and T. Xue, "Fast and high quality image denoising via malleable convolution," in *Computer Vision–(ECCV)*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham, Switzerland: Springer, 2022, pp. 429–446.

[28] H. Yin and S. Ma, "CSformer: Cross-scale features fusion based transformer for image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1809–1813, 2022.

[29] M. Zhao, G. Cao, X. Huang, and L. Yang, "Hybrid transformer-CNN for real image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1252–1256, 2022.

[30] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 2333–2337.

[31] A. Guo, L. Fang, M. Qi, and S. Li, "Unsupervised denoising of optical coherence tomography images with nonlocal-generative adversarial network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[32] R. Wu et al., "Unsupervised OCT image despeckling with ground-truth-and repeated-scanning-free features," *Opt. Exp.*, vol. 32, no. 7, pp. 11934–11951, 2024.

[33] J. Lehtinen et al., "Noise2Noise: Learning image restoration without clean data," 2018, *arXiv:1803.04189*.

[34] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void–learning denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2124–2132.

[35] J. Batson and L. Royer, "Noise2Self: Blind denoising by self-supervision," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 524–533.

[36] Z. Wang, J. Liu, G. Li, and H. Han, "Blind2unblind: Self-supervised image denoising with visible blind spots," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2027–2036.

[37] W. Lee, S. Son, and K. M. Lee, "AP-BSN: Self-supervised denoising for real-world images via asymmetric PD and blind-spot network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17725–17734.

[38] Y. I. Jang, K. Lee, G. Y. Park, S. Kim, and N. I. Cho, "Self-supervised image denoising with downsampled invariance loss and conditional blind-spot network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12196–12205.

[39] M. Prakash, A. Krull, and F. Jug, "Fully unsupervised diversity denoising with convolutional variational autoencoders," 2020, *arXiv:2006.06072*.

[40] B. Salmon and A. Krull, "Direct unsupervised denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3838–3845.

[41] Z. Zheng, Z. Wang, Z. Hu, Z. Wan, and W. Ma, "Recovering traffic data from the corrupted noise: A doubly physics-regularized denoising diffusion model," *Transp. Res. C, Emerg. Technol.*, vol. 160, Mar. 2024, Art. no. 104513.

[42] V. Kulikov, S. Yadin, M. Kleiner, and T. Michaeli, "SinDDM: A single image denoising diffusion model," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 17920–17930.

[43] X. Yu et al., "Self-supervised Blind2Unblind deep learning scheme for OCT speckle reductions," *Biomed. Opt. Exp.*, vol. 14, no. 6, pp. 2773–2795, 2023.

[44] Y. Li, Y. Fan, and H. Liao, "Self-supervised speckle noise reduction of optical coherence tomography without clean data," *Biomed. Opt. Exp.*, vol. 13, no. 12, pp. 6357–6372, 2022.

[45] D. Kermany, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification," *Mendeley data*, 2018.

[46] X. He et al., "Exploiting multi-granularity visual features for retinal layer segmentation in human eyes," *Frontiers Bioeng. Biotechnol.*, vol. 11, Jun. 2023, Art. no. 1191803.

[47] X. He et al., "Lightweight retinal layer segmentation with global reasoning," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.