A Multi-Perceptual Learning Network for Retina OCT Image Denoising and Classification

Zhe Xiao*, Zongqi He*, Zhuoning Xu*, Yunze Li*, Zelin Song*, Calvin Leighton†, Li Wang*, Shanru Liu*
and Shiun Yee Wong†, Wenfeng Huang†, Wenjing Jia†, and Kin-Man Lam*
* Hong Kong Polytechnic University, Hong Kong
† University of Technology Sydney, Australia

Abstract—**Swept Source Optical Coherence Tomography (OCT), a non-invasive cross-sectional imaging technique, has been widely used in diagnosing and treating various vision-related diseases. However, OCT images often suffer from heavy noise issues, due to the limitations of imaging devices, making analysis and disease classification a great challenge. This paper proposes a Multi-Perceptual Learning Network (MPLN) for retina OCT image denoising and classification. We adopt a triplet cross-fusion GAN approach and use three unpaired OCT images to conduct perceptual learning. In addition, we integrate the Frequency Distribution Loss into GAN to preserve both the structural integrity and perceptual quality of the denoised OCT images, enabling better classification. The method can significantly reduce the noise of highly noisy images. Our proposed method is evaluated on the VIP Cup 2024 dataset in terms of the CNR, MSR, and TP scores. Our model achieves a CNR score of 6.351, and an MSR score of 11.573, which outperforms many existing methods on OCT images. In classification, our MPLN improves accuracy by more than one percent. These results demonstrate that our model can significantly enhance image quality and improve classification accuracy, highlighting its potential for clinical applications.**

I. Introduction

Optical coherence tomography (OCT), a non-invasive cross-sectional imaging technique, plays a crucial role in various medical applications, particularly retinal diagnostics [1]. However, two main issues hinder the development of OCT-based diagnosis. First, OCT images are inevitably corrupted by heavy speckle noise due to the low coherence interferometry imaging modality [2]. Second, OCT image analysis heavily relies on manual labor, which increases the burden on ophthalmologists and leads to potential bias due to subjective opinions. Therefore, an automatic and reliable OCT image analysis algorithm is essential for the efficient diagnosis of eye diseases, which will help alleviate the strain on medical resources.

Over the past decades, several approaches have been proposed to enhance the quality of OCT images. One approach to denoise OCT images involves capturing a sequence of repeated B-scans from the same position and then registering and averaging them to produce a less noisy image [3]. However, this method significantly increases the image acquisition time, making it impractical for clinical use. Another approach is model-based OCT denoising, where a single B-scan image is denoised using digital filters based on the statistical models of signal and noise [4], [5], or through deep/dictionary learning techniques [6], [7]. While these methods can reduce noise, they often result in over-smoothing or loss of fine details. Inspired by the remarkable success of deep learning in various vision tasks, deep learning-based methods for OCT denoising have shown superior performance in both speckle noise reduction and structure preservation. Researchers have improved denoising performance by optimizing deep network architectures [8], [9] and designing structure-sensitive loss functions [10], [11]. However, these methods require paired noisy-clean OCT images for training, which are difficult to obtain due to involuntary eye or body movements during scanning [12].

In addition to image reconstruction, OCT image analysis, especially classification, is also a highly important topic in clinical applications. This field has witnessed a growing number of diverse OCT classification methods over the past years [13]–[15]. Some of them are relatively straightforward, such as the utilization of Support Vector Machines (SVMs) and Random Forests [16], [17], while others are more effective and sophisticated, leveraging significant advancements in the computer vision community [18]. For example, as one of the most prominent deep learning techniques, Convolutional Neural Networks (CNNs) have been widely investigated and explored for OCT image classification. However, when CNNs are applied to noisy OCT images, their classification performance is notably degraded. Applying conventional image processing techniques, including denoising, to OCT images may also lead to the loss or alteration of critical details, ultimately affecting classification results. Additionally, it has been demonstrated that some OCT image classification algorithms can only perform well on specific datasets, and their performances heavily depend on the characteristics of the data. To address the abovementioned issues, we propose a Multi-Perceptual Learning Network (MPLN) for retina OCT image processing, which jointly tackles the challenges in denoising and classification of OCT images.

In the initial stage of our model, a Generative Adversarial Network (GAN), comprising a DnCNN [19] generator and a PatchGAN [20] discriminator, is employed. We incorporate the Frequency Distribution (FD) Loss [21] into the network. This FD Loss function introduces perceptual learning by converting image features into the frequency domain and computing the distribution distance. This process not only enhances the perceptual quality, but also safeguards the structural integrity of images. The generator loss function includes GAN loss, consistency loss, FD loss, PSNR loss, and SSIM loss. In addition, inspired by the Triplet Cross-Fusion Learning (TCFL) scheme [22], our model accepts a triplet combination of inputs, consisting of two noisy images and one clean image, which are unpaired. We enable perceptual learning between the input images to help the network extract the most relevant and significant features. This triplet input strategy significantly expands the training dataset by incorporating an arbitrary number of noisy and clean images. These OCT images do not need to be matched pairs and there is no restriction on the quantity. This broadens the scope and flexibility of our training mechanism.

In the second stage, for retina OCT image classification, we adopted and further improved the Lesion-Aware Convolutional Neural Network (LACNN) [13]. Inspired by the remarkable performance achieved by deep learning models in various vision tasks [23]–[27], we applied the Residual Network (ResNet) to the Lesion Detection Network (LDN) [13] to generate an attention map that highlights the lesion areas in each OCT image. This attention map is then used to compute a weighted feature map, which is subsequently combined with the original convolutional feature maps to obtain the lesion-aware feature map. This map serves as input to the subsequent convolutional layers, allowing the model to focus on clinically relevant lesion regions during classification.

In this study, we advance the LACNN architecture by replacing the original backbone of the LACNN with ResNet, which significantly enhances the model's ability to accurately classify OCT images. This improvement provides a valuable tool for the early detection and differentiation of ocular diseases in clinical practice. In addition, we demonstrate that the MPLN model enhances image perceptual quality. Our model facilitates multi-perceptual learning among images, providing effective denoising and improving the perceptual quality of images. FD loss serves as a crucial element for improving image quality, and this enhancement is validated through classification experiments. The results show that the model trained on denoised images by the MPLN model achieves higher accuracy, confirming the effectiveness of our approach.

In summary, our main contributions are as follows:

- We propose an MPLN framework that effectively addresses both OCT image denoising and classification tasks in a unified approach, leading to improved overall performance in image interpretation.
- We introduce the FD Loss into the GAN architecture, which helps preserve the structural integrity of OCT images during denoising. This facilitates multi-perceptual learning, enhancing both the quality of the denoised images and the classification accuracy.
- We incorporate LDN with ResNet to obtain the refined lesion-aware attention map, which is then combined with LACNN to classify the OCT images more accurately.

II. The Proposed MPLN Method

A. MPLN Denoising

1) Network Architecture: Our proposed MPLN method makes use of a triplet cross-fusion GAN [22] and employs three unpaired OCT images for perceptual learning. In this approach, a speckle noise-corrupted OCT image X can be expressed as the sum of its clean component C and its speckle noise Y, as follows:

$$X = C + Y. \tag{1}$$

Specifically, in our approach, the clean components and the noisy components for training consist of three unpaired images: two noisy OCT images X_1 and X_2, and one clean OCT image C_0. These three unpaired images undergo multi-perceptual learning to help the network extract and learn the speckle noise properties, thereby improving denoising accuracy even with a small dataset. We used a total of 5 clean images and 10 noisy images, all of which are unpaired. The total number of combinations of clean and noisy images is $\binom{5}{1} \times \binom{10}{2} = 225$, which significantly enlarges the training set.

In addition, our proposed approach employs a GAN-based denoising framework, which consists of three key components, as shown in Fig. 1: generating clean images from noisy images, synthesizing noisy images from clean images by adding estimated noises, and generating clean images from synthesized noisy images.

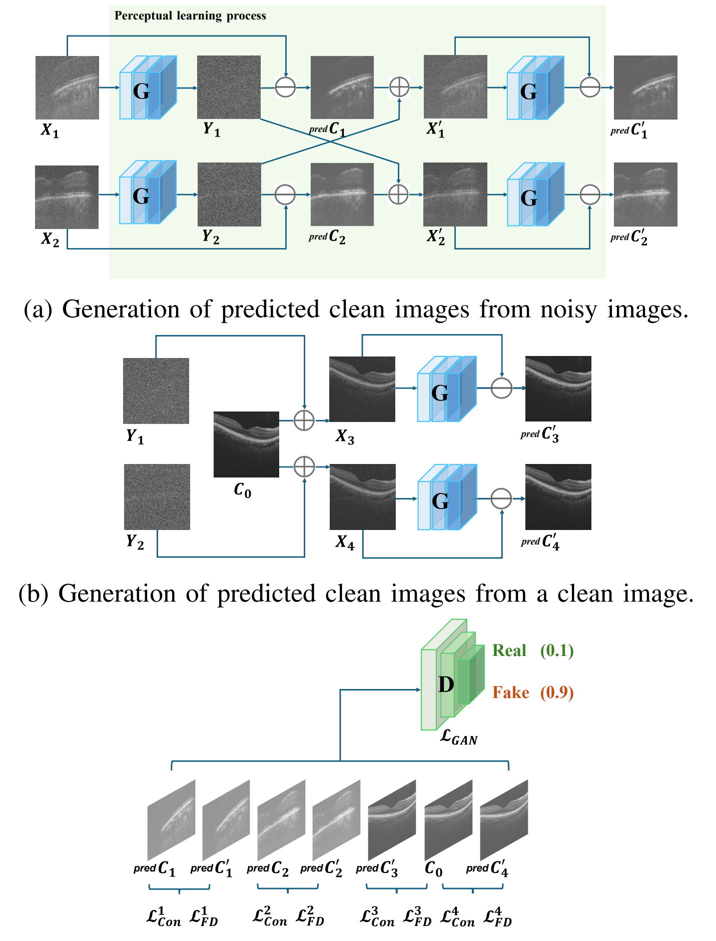
Fig. 1(a) shows the first component of the network, generation of predicted clean images from noisy images, where G represents the Generator. The generator first takes in the two noisy inputs X_1 and X_2. After passing through a perceptual learning process, the first component outputs predicted clean images predC_1' and predC_2'. In the perceptual learning process, the outputs of the generator Y_1 and Y_2 are the estimated noise component of the noisy inputs X_1 and X_2. Subtracting the estimated noise components from themrespectively produces the estimated clean images of the two noisy inputs predC_1 and predC_2. In the cross-fusion mechanism, the generated clean images predC_1 and predC_2 are then added with the estimated noise of the other noise components Y_2 and Y_1 respectively to produce two more synthesized noisy images X_1' and X_2', as the input for the following perceptual learning process.

Similar to the prediction from noisy images, we can add the estimated noise Y to the input clean image to synthesize a fake noisy picture X'. Then, a generator can produce their predicted clean images predC'. Fig. 1(b) illustrates the second component of the network, generation of predicted clean images from a clean image. The noises Y_1 and Y_2, predicted by the generator in the first component, are used in this component. The input is the clean image C_0, and the outputs are predC_3' and predC_4'.

The Discriminator's role is to differentiate between generated images and clean images. As shown in Fig. 1(c), the predicted images from the first two components of the network, together with the clean image C_0, are inputted to the Discriminator D.

2) Loss Functions: The total loss of MPLN is defined as follows:

$$\mathcal{L}_{\text{Total}} = \alpha\mathcal{L}_{GAN} + \beta\mathcal{L}_{Con} + \gamma\mathcal{L}_{FD} + \delta\mathcal{L}_{PSNR} + \epsilon\mathcal{L}_{SSIM}, \tag{2}$$



(a) Generation of predicted clean images from noisy images.

(b) Generation of predicted clean images from a clean image.

(c) The Discriminator and the calculation of Consistency Losses.

Fig. 1. The MPLN triplet cross-fusion mechanism and GAN components.

where α, β, γ, δ, and ϵ are the weights for the respective loss functions. The GAN loss $\mathcal{L}_{GAN}$ attempts to minimize $\mathbb{E}\left[(D(c)-1)^2\right]$ for the generator and minimize $\mathbb{E}\left[(D(C_0)-1)^2+D(c)^2\right]$ for the discriminator, where c denotes an element within the set $\{\text{pred}C_1, \text{pred}C'_1, \text{pred}C_2, \text{pred}C'_2, \text{pred}C'_3, \text{pred}C'_4\}$.

$\mathcal{L}_{Con}$ is the Consistency Loss; $\mathcal{L}_{FD}$ indicates the FD Loss [21] for computing the distribution distance in the frequency domain; $\mathcal{L}_{PSNR}$ and $\mathcal{L}_{SSIM}$ denote PSNR and SSIM losses, respectively.

The spatial Consistency Loss $\mathcal{L}_{Con}$ is designed to differentiate the real clean image C from the generated clean images predC_1, predC_2, predC'_3 and predC'_3, and is computed as follows:

$$\mathcal{L}_{Con} = \mathcal{L}^1_{Con} + \mathcal{L}^2_{Con} + \mu_{Con}(\mathcal{L}^3_{Con} + \mathcal{L}^4_{Con}). \tag{3}$$

where $\mathcal{L}^1_{Con}$, $\mathcal{L}^2_{Con}$, $\mathcal{L}^3_{Con}$ and $\mathcal{L}^4_{Con}$ are computed as follows:

$$\begin{aligned}
\mathcal{L}^1_{Con} &= \parallel \text{pred}C_1 - \text{pred}C'_1 \parallel_1 \\
\mathcal{L}^2_{Con} &= \parallel \text{pred}C_2 - \text{pred}C'_2 \parallel_1 \\
\mathcal{L}^3_{Con} &= \parallel \text{pred}C'_3 - C_0 \parallel_1 \\
\mathcal{L}^4_{Con} &= \parallel \text{pred}C'_4 - C_0 \parallel_1
\end{aligned} \tag{4}$$

where $\parallel \ \parallel_1$ denotes the $L1$ norm.

In addition to the spatial domain consistency, to achieve better perceptual quality, we adopt the FD Loss [21] to compute the distribution distances $\mathcal{L}_{FD}$ of the predicted clean images in the frequency domain, which can be calculated as:

$$\mathcal{L}_{FD} = \mathcal{L}^1_{FD} + \mathcal{L}^2_{FD} + \mu_{FD}(\mathcal{L}^3_{FD} + \mathcal{L}^4_{FD}), \tag{5}$$

where μ_{FD} is the loss weight, and $\mathcal{L}^1_{FD}$, $\mathcal{L}^2_{FD}$, $\mathcal{L}^3_{FD}$ and $\mathcal{L}^4_{FD}$ can be computed as follows:

$$\begin{aligned}
\mathcal{L}^1_{FD} &= \text{fdl}(\text{pred}C_1 - \text{pred}C'_1), \\
\mathcal{L}^2_{FD} &= \text{fdl}(\text{pred}C_2 - \text{pred}C'_2), \\
\mathcal{L}^3_{FD} &= \text{fdl}(\text{pred}C'_3 - C_0), and \\
\mathcal{L}^4_{FD} &= \text{fdl}(\text{pred}C'_4 - C_0),
\end{aligned} \tag{6}$$

where $fdl()$ represents the FD loss function, which uses a pre-trained feature extractor to map the predicted and target images to a feature space. Discrete Fourier Transform (DFT) is applied, followed by Sliced Wasserstein Distance to measure the distance among the target images in the frequency domain. This approach effectively mitigates the interference of spatial misalignment, thereby better capturing the perceptual properties of the images. The combinations of $\mathcal{L}^1_{FD}$, $\mathcal{L}^2_{FD}$, $\mathcal{L}^3_{FD}$, $\mathcal{L}^4_{FD}$, together with $\mathcal{L}^1_{Con}$, $\mathcal{L}^2_{Con}$, $\mathcal{L}^3_{Con}$ and $\mathcal{L}^4_{Con}$ facilitate the multi-perceptual learning of noises from different components.

B. Retina OCT Image Classification

To demonstrate the efficacy of the denoising model on the downstream OCT image classification task, we designed a classification model specially for denoised retina OCT images.

The core modules of our classification model, termed "LACNN-ResNet", are illustrated in Fig. 2. We adopted and enhanced the LACNN approach [13] by incorporating residual connections to learn more discriminative features for classification. Additionally, we embedded the Lesion Attention Network (LAN) [13] to strengthen the local lesion-related features while also considering the global structures of the OCT images.

For a lesion attention block, denote x as the input image tensor, $Con_{i,c}(x)$ as the output of a convolutional layer, and A_i as the attention map generated by LAN, where i represents the spatial position and c denotes the channel index. The weighted feature map, represented by $W_{i,c}$, is defined as follows:

$$W_{i,c}(x) = Con_{i,c}(x) \times A_i, \tag{7}$$

where × denotes the element-wise product of corresponding spatial positions from attention maps and feature maps.

In addition, the feature maps generated by convolution are superimposed with the weighted feature map of the corresponding spatial positions. This is based on the consideration that LAN can only mark lesion areas, while unmarked areas may also contain information useful for classification. Thus, the output of an LAN, denoted by $L_{i,c}(x)$, can be represented as an element-wise summation of the corresponding spatial positions from the attention maps and the weighted feature maps, as follows:

$$L_{i,c}(x) = W_{i,c}(x) + Con_{i,c}(x). \tag{8}$$

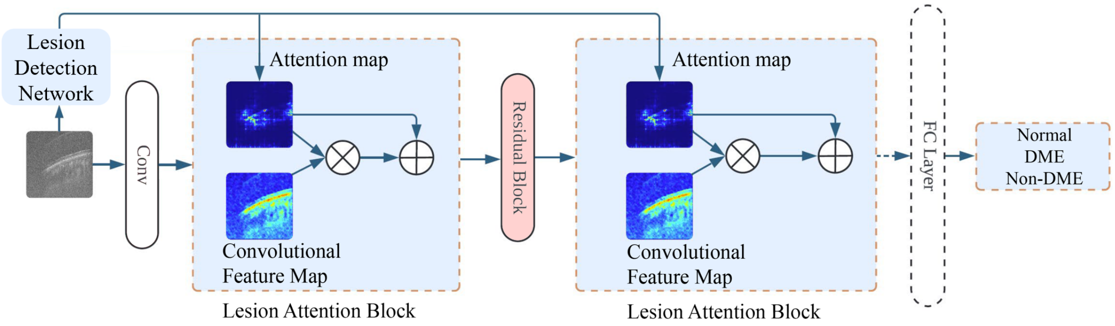


Fig. 2. The core modules of the proposed LACNN-ResNet classification model.

III. Experimental Setup and Results

A. MPLN Denoising

1) Experimental Setup: In the proposed MPLN method, we adopted DnCNN [19] as the generator and PatchGAN [20] as the discriminator. DnCNN is a classical denoising network and serves as a noise predictor in this GAN-based network.

We implemented MPLN with PyTorch and used an NVIDIA GTX 4090 GPU for training. The number of training epochs was set to 50. We opted for cosine annealing learning rate and set the cosine decay period to match the entire training cycle. An initial learning rate was set to 2e-5, gradually decaying to the final value of 1.2e-6. We empirically set α=1, β=3, γ=0.03, δ=0.5, and ϵ=1 as the weights for the generator loss function.

During testing, input images from the VIP Cup 2024 Test Set were normalized to the range $[1, 3]$. A dataloader was established to load a triplet set of input images. We evaluated the proposed MPLN method using the test set, which includes B-scans from 18 subjects. Each subject's dataset consists of 70 to 300 B-scans, which are noisy and have resolutions of either 300 × 150 or 300 × 200 pixels.

2) Evaluation Metrics: To assess the performance of denoising methods, we adopted the Contrast-to-Noise Ratio (CNR), Mean-to-Standard-deviation Ratio (MSR), and Texture Preservation (TP) as evaluation metrics.

CNR evaluates the contrast between the foreground and the background of generated images, which is calculated as follows:

$$S_{\text{CNR}} = 10 \log \left| \frac{\mu_f - \mu_b}{\sqrt{\sigma_f^2 + \sigma_b^2}} \right|, \tag{9}$$

where μ_f and σ_f denote the mean and the standard deviation of the foreground regions, respectively, and μ_b and σ_b denote the mean and the standard deviation of the background region, respectively. The Regions of Interest (ROIs) are chosen between different layers to show the change in contrast.

Similarly, **MSR** evaluates the concentration of pixel intensity values and is calculated as follows:

$$S_{\text{MSR}} = \frac{\mu_f}{\sigma_f}. \tag{10}$$

TP evaluates the conserving of texture between the noisy image and the denoised image and is calculated as follows:

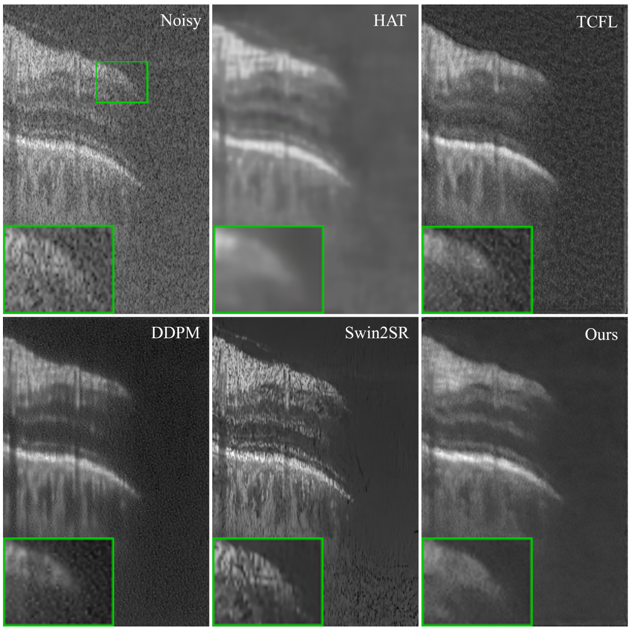


Fig. 3. Visual comparison of the denoising results of HAT [28], TCFL [22], DDPM [29], Swin2SR [30] and our MPLN.

$$S_{\text{TP}}^{(m)} = \frac{\sigma_{m(de)}^2}{\sigma_{m(noisy)}^2} \sqrt{\frac{\mu_{de}}{\mu_{noisy}}}, \tag{11}$$

where $\sigma_{m(noisy)}$ and $\sigma_{m(de)}$ denote the standard deviation of the m-th ROI in the noisy image and denoised image, respectively, and μ_{de} and μ_{noisy} denote their corresponding means. The ROIs encompass the intra-layer regions. We take the average of all the ROIs in the image as its TP score:

$$S_{\text{TP}} = \frac{1}{M} \sum_{m=1}^{M} S_{\text{TP}}^{(m)}. \tag{12}$$

3) Denoising Results and Comparison: We selected a representative retina OCT image from the VIP Cup 2024 Test Set and visually compared its denoising result with other existing approaches in Fig. 3.

Both DDPM and HAT have reduced the speckle noise to a certain extent. However, due to the heavy noise of the original noisy image, some noise still remains, as highlighted by the enlarged area in Fig. 3. While HAT produced smoother images compared to our MPLN method, they suffer from excessive smoothness, which leads to the loss of critical structural details or blood flow structure, as indicated by the box in Fig. 3. The proposed MPLN effectively reduces the noise and improves perceptual quality of the image.

Table I shows the quantitative results comparing the denoising performance of our MPLN with other methods. As shown in this table, our MPLN obtained the best scores across CNR and MSR. Although TCFL, DDPM, and Swin2SR have marginally higher TP scores, they all obtain lower CNR and MSR scores. HAT yields a low TP score, indicating its over-smoothing problem. Our MPLN achieves balanced scores across the evaluation matrices, effectively reducing the noise.

B. MPLN-denoised Retina OCT Image Classification

1) Experimental Setup: The proposed MPLN classification was implemented using the same PyTorch framework and GPU as in the earlier stage. We evaluated the proposed LACNN-ResNet method utilizing the VIP Cup training dataset, which includes approximately 18,000 retina OCT B-scan images from 100 patients. We applied our denoising method to the original images and used them as training samples.

We used 90% of the VIP Cup 2024 training dataset for training and the remaining 10% for testing. We evaluated and analyzed the performance of our model under these two configurations. To further enhance the training process, we applied data augmentation to the dataset. Specifically, all input images were resized, horizontally flipped, and rotated within a range of -15 to +15 degrees, followed by normalization.

To further verify the improvement of the MPLN model over mainstream denoising models, we randomly selected a set of noisy images and used them as input for both the TCFL model and the MPLN model, generating two separate test sets. We then tested these sets using the LACNN-ResNet model, which had been trained on noisy images, and compared their accuracy.

2) Classification and Comparison: For each input image, the model outputs the predicted probability distributions across three classes: Normal, DME, and Non-DME. We determine the class label according to the highest proportion of predictions from the patient's OCT images.

Two models were trained for comparison. The first model on the original noisy VIP Cup 2024 training set, and the second model on the denoised version of the same dataset. After each epoch, testing was conducted on a test set approximately 1/9 the size of the training set (around 2,000 images). Results shows that the LACNN-ResNet model trained on the denoised dataset exhibited a 1.45% improvement in accuracy compared to the model trained on noisy data. In addition, the classification accuracy of MPLN denoised dataset is 5.82% higher than that of TCFL. These results demonstrate that the MPLN model effectively enhances classification performance for OCT images by reducing noise interference.

IV. Conclusion

In this paper, we have proposed a Multi-Perceptual Learning Network (MPLN) for denoising retina OCT images. To capture perceptually significant features for denoising, our MPLN model employs a cross-fusion GAN approach, utilizing three unpaired noisy and clean OCT images for multi-perceptual learning, while enforcing image consistency in both spatial and frequency domains by minimizing frequency distribution losses. Our MPLN has shown its ability to effectively reduce noise, enhance OCT image quality, and improve retina OCT image classification for eye disease diagnosis. Experimental results on the VIP Cup 2024 dataset have demonstrated that the proposed method outperforms state-of-the-art image denoising models both visually and quantitatively in this challenge.

TABLE I
Quantitative comparison of the denoising performance of our MPLN with SOTA approaches HAT [28], TCFL [22], DDPM [29] and Swin2SR [30] on the VIP Cup 2024 Test Set.

	Noisy	HAT	TCFL	DDPM	Swin2SR	**MPLN (ours)**
CNR↑	1.966	3.579	3.184	2.513	4.079	**6.351**
MSR↑	6.872	9.702	7.972	7.623	10.302	**11.573**
TP↑	1.000	0.342	**0.685**	0.670	0.578	0.554

Acknowledgment

We would like to extend our heartfelt thanks to Jun Xiao, Cong Zhang, and Cuixin Yang, whose expertise and willingness to share knowledge have greatly contributed to our understanding and application of the methodologies involved in this collaborative project. Additionally, we are immensely grateful to the organizers of the VIP Cup 2024 for providing a key dataset that enabled us to test our methods on real-world data. Finally, we acknowledge the support of our colleagues and peers who have provided constant encouragement and assistance throughout this journey.

References

[1] W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," *Progress in retinal and eye research*, vol. 27, no. 1, pp. 45–88, 2008.

[2] G. Gong, H. Zhang, and M. Yao, "Speckle noise reduction algorithm with total variation regularization in optical coherence tomography," *Optics express*, vol. 23, no. 19, pp. 24 699–24 712, 2015.

[3] A. W. Scott, S. Farsiu, L. B. Enyedi, D. K. Wallace, and C. A. Toth, "Imaging the infant retina with a hand-held spectral-domain optical coherence tomography device," *American journal of ophthalmology*, vol. 147, no. 2, pp. 364–373, 2009.

[4] M. H. Eybposh, Z. Turani, D. Mehregan, and M. Nasiriavanaki, "Cluster-based filtering framework for speckle reduction in OCT images," *Biomedical optics express*, vol. 9, no. 12, pp. 6359–6373, 2018.

[5] S. Aja-Fernández and C. Alberola-López, "On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2694–2701, 2006.

[6] L. Fang, S. Li, Q. Nie, J. A. Izatt, C. A. Toth, and S. Farsiu, "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomedical optics express*, vol. 3, no. 5, pp. 927–942, 2012.

[7] L. Fang, S. Li, R. P. McNabb, *et al.*, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE transactions on medical imaging*, vol. 32, no. 11, pp. 2034–2049, 2013.

[8] M. Xu, C. Tang, F. Hao, and et al., "Texture preservation and speckle reduction in poor optical coherence tomography using the convolutional neural network," *Medical Image Analysis*, vol. 64, p. 101 727, 2020.

[9] Z. Shen, M. Xi, C. Tang, M. Xu, and Z. Lei, "Double-path parallel convolutional neural network for removing speckle noise in different types of OCT images," *Applied Optics*, vol. 60, no. 15, pp. 4345–4355, 2021.

[10] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomedical optics express*, vol. 9, no. 11, pp. 5129–5146, 2018.

[11] B. Qiu, Z. Huang, X. Liu, *et al.*, "Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function," *Biomedical optics express*, vol. 11, no. 2, pp. 817–830, 2020.

[12] Y. Huang, W. Xia, Z. Lu, *et al.*, "Noise-powered disentangled representation for unsupervised speckle reduction of optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2600–2614, 2020.

[13] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[14] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," *IEEE Trans. Medical Imaging*, vol. 37, no. 4, pp. 1024–1034, 2017.

[15] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. on Medical Imaging*, vol. 40, no. 6, pp. 1591–1602, 2021.

[16] B. Zagajewski, M. Kluczek, E. Raczko, A. Njegovec, A. Dabija, and M. Kycko, "Comparison of random forest, support vector machines, and neural networks for post-disaster forest species mapping of the krkonoše/karkonosze transboundary biosphere reserve," *Remote Sensing*, vol. 13, no. 13, 2021.

[17] A. Lang, A. Carass, E. Sotirchos, P. Calabresi, and J. L. Prince, "Segmentation of retinal OCT images using a random forest classifier," in *Medical Imaging 2013: Image Processing*, SPIE, vol. 8669, 2013, pp. 199–205.

[18] X. Huang, Z. Ai, H. Wang, *et al.*, "Gabnet: Global attention block for retinal OCT disease classification," *Frontiers in Neuroscience*, vol. 17, p. 1 143 422, 2023.

[19] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," en, *2017 ICCV*, Oct. 2017.

[21] Z. Ni, J. Wu, Z. Wang, W. Yang, H. Wang, and L. Ma, "Misalignment-robust frequency distribution loss for image transformation," *arXiv preprint arXiv:2402.18192*, 2024.

[22] M. Geng, X. Meng, L. Zhu, *et al.*, "Triplet cross-fusion learning for unpaired image denoising in optical coherence tomography," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3357–3372, 2022.

[23] J. Xiao, W. Jia, and K.-M. Lam, "Feature redundancy mining: Deep light-weight image super-resolution model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1620–1624.

[24] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Self-feature learning: An efficient deep lightweight network for image super-resolution," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4408–4416.

[25] J. Xiao, X. Jiang, N. Zheng, *et al.*, "Online video super-resolution with convolutional kernel bypass grafts," *IEEE Transactions on Multimedia*, vol. 25, pp. 8972–8987, 2023.

[26] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[27] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2995–3004.

[28] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *2023 CVPR*, 2023, pp. 22 367–22 377.

[29] D. Hu, Y. K. Tao, and I. Oguz, "Unsupervised denoising of retinal OCT with diffusion probabilistic model," in *Medical Imaging 2022: Image Processing*, O. Colliot and I. Išgum, Eds., International Society for Optics and Photonics, vol. 12032, SPIE, p. 1 203 206.

[30] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," in *ECCV*, Springer, 2022, pp. 669–687.