SiameseGAN: A Generative Model for Denoising of Spectral Domain Optical Coherence Tomography Images

Nilesh A. Kande, Rupali Dakhane, Ambedkar Dukkipati, *Member, IEEE*, and Phaneendra Kumar Yalavarthy, *Senior Member, IEEE*

***Abstract*— Optical coherence tomography (OCT) is a standard diagnostic imaging method for assessment of ophthalmic diseases. The speckle noise present in the high-speed OCT images hampers its clinical utility, especially in Spectral-Domain Optical Coherence Tomography (SDOCT). In this work, a new deep generative model, called as SiameseGAN, for denoising Low signal-to-noise ratio (LSNR) B-scans of SDOCT has been developed. SiameseGAN is a Generative Adversarial Network (GAN) equipped with a siamese twin network. The siamese network module of the proposed SiameseGAN model helps the generator to generate denoised images that are closer to groundtruth images in the feature space, while the discriminator helps in making sure they are realistic images. This approach, unlike baseline dictionary learning technique (MSBTD), does not require an apriori high-quality image from the target imaging subject for denoising and takes less time for denoising. Moreover, various deep learning models that have been shown to be effective in performing denoising task in the SDOCT imaging were also deployed in this work. A qualitative and quantitative comparison on the performance of proposed method with these state-of-the-art denoising algorithms has been performed. The experimental results show that the speckle noise can be effectively mitigated using the proposed SiameseGAN along with faster denoising unlike existing approaches.**

***Index Terms*— SDOCT denoising, deep learning, SiameseGAN, deep generative model.**

I. Introduction

SPECTRAL domain optical coherence tomography (SDOCT) is a non-invasive, cross-sectional imaging modality that has been widely used in ophthalmology [1], [2]. For clinical analysis, ophthalmologists require high signal-to-noise ratio (SNR) in the SDOCT images. However, speckle noise [3], [4] arising out of the low coherence interferometry degrades the quality of the OCT images. This noise poses a significant challenge to the process of interpretation, in particular, accurate diagnosis of vision-related diseases.

One approach to denoise SDOCT images is to capture a sequence of repeated B-scans from a unique position and then performing registering and averaging to create a less noisy image [5]. The main limitation of this approach is that it drastically increases the image acquisition time. The second approach is model-based SDOCT denoising, where a single B-scan image is denoised using digital filters that depend on statistics and model of signal and noise [6]–[14] or using deep/dictionary learning techniques [15]–[18]. The resulted denoised images using this approach tend to show over-smoothening or missing features.

Further, most model based approaches minimize the mean square error (MSE) between the ground truth image and the denoised image. While MSE based methods improve the peak signal-to-noise ratio (PSNR), they tend to compromise the important structural details. Another drawback of these methods is that they require large computational time to denoise the SDOCT images. In particular, the dictionary learning techniques can require as high as 31 minutes for denoising a single image [17], making them less practical in the clinical setting.

In this work, we propose a new deep generative model called SiameseGAN for denoising the SDOCT images. The proposed model is a Generative Adversarial Network (GAN) equipped with a siamese twin network and hence we refer to it as SiameseGAN. The siamese network in our proposed model forces the GAN to generate denoised images that are closer to the ground-truth images. This module additionally helps the discriminator to fool the generator to produce a denoised image by extracting the discriminative features from the groundtruth high signal-to-noise ratio (HSNR) patch and the denoised patch by passing them through a twin network. Further, this classifies the pair of HSNR patch and denoised patch as a non-matching pair and loss is incurred, if they are classified as matching pair by siamese network. The loss ensures that along with the denoised and groundtruth

Manuscript received June 21, 2020; revised September 8, 2020; accepted September 11, 2020. Date of publication September 14, 2020; date of current version December 29, 2020. This work was supported in part by the IMPRINT under Grant IMP/2019/000383 and in part by the WIPRO GE-CDS Collaborative Laboratory of Artificial Intelligence in Healthcare and Imaging. *(Corresponding author: Phaneendra Kumar Yalavarthy.)*

Nilesh A. Kande, Rupali Dakhane, and Ambedkar Dukkipati are with the Department of Computer Science and Automation, Indian Institute of Science, Bengaluru 560012, India (e-mail: kandea@iisc.ac.in; rupalid@iisc.ac.in; ambedkar@iisc.ac.in).

Phaneendra Kumar Yalavarthy is with the Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru 560012, India (e-mail: yalavarthy@iisc.ac.in).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3024097

patches being perceptually similar, their extracted features are also identical. The speckle noise in the SDOCT images poses great difficulty in differentiating this from the true OCT image formation signals. The statistical correlations that are performed for the feature vectors in the siamese network will be able to provide this differentiation, in turn enabling effective denoising. Also in the proposed model SiameseGAN, the combined restoration loss (perceptual loss and multiscale structural similarity index metric (MS-SSIM) loss) along with siamese loss is minimized to provide improved approximation to ground-truth images.

Here, the variant GAN that we use is Wasserstein GAN (or WGAN) [19] to encourage the denoised SDOCT images to share the same distribution as that of averaged and registered HSNR images. In sequel, by GAN we are referring Wasserstein GAN. These denoised patches along with HSNR patches are then fed to the discriminator module, siamese module and Image restoration loss calculator. The losses computed from all three modules are then combined.

In this work, the proposed SiameseGAN model has been trained end-to-end, that is the combined output loss from the three modules is then back-propagated to learn the weights of generator, discriminator and siamese network. Once the parameters of these three networks has been trained, generator network was utilized to denoise test images. While testing, instead of patches, a whole low signal-to-noise ratio (LSNR) test image was given as input to the generator to obtain the output as the denoised version of the image.

In short, the main contributions of this work is as follows.

- A new generative model called as SiameseGAN was designed/developed specially for denoising of SDOCT images.
- We propose to combine restoration loss with siamese loss and show that this will result in denoised SDOCT images that are closer to ground truth images.
- A systematic comparison with the existing deep learning networks used for denoising, in particular GAN based methods, was performed to show the efficacy of the proposed network model.
- Bench marking of proposed SiameseGAN model with the state-of-the-art deep/dictionary learning models that exist in the literature for performing the denoising task in SDOCT.

It is important to note that main job of discriminator in the traditional GANs is to classify the denoised (generated) image as real of fake. In the proposed SiameseGAN, the twin network performs the task of similarity between the input pairs, thus forcing the generator to produce semantic features close to the expected ground truth. This aspect of SiameseGAN is beneficial for the OCT image denoising task, where the aim is to provide a close estimate of HSNR image from the LSNR version. Also to reiterate, other than deep-learning models for denoising, the standard approach is dictionary-based denoising, which requires HSNR OCT image to build the dictionary. The proposed network is a purely a data-driven network, which overcomes the requirement of having HSNR image and works with only LSNR images for producing the denoised result. Further, the deep learning models, including the proposed one, requires HSNR images only in the training phase, which is performed off-line.

The remainder of the paper is organized as follows: Section II gives the background information on SDOCT denoising. Section III gives the proposed methods and the objective and Section IV describes the network architecture. In Section V, the experiments and results have been described. Section VI mentions the related work and finally conclusion and future work are given in Section VII.

II. Background

The methods that are proposed for removing speckle noise in SDOCT images can be categorized into two groups: (i) model-based single-frame, and (ii) multi-frame averaging techniques. The first group of methods assume an apriori parametric or non-parametric model for the signal and noise. Local statistics-based filtering methods [6]–[8] belong to this category, which are time efficient, but are limited in preservation of the details. Though diffusion-based methods [9]–[12] achieve good results for mitigating the noise, they largely produce over-smooth images. The wavelet-based methods [13], [14] too can give good results, but they are known to introduce some artifacts.

In the multi-frame averaging techniques, a sequence of repeated B-scans are captured from a unique position and then image registration and averaging is performed [5]. While SDOCT imaging systems with built-in image stabilization and averaging systems can produce HSNR images directly, registration and averaging can also be performed after the images are captured. In both these cases, the image acquisition time dramatically increases to produce HSNR images.

The dictionary learning technique has been based on sparse representation using local image patches and is a hybrid of the above two techniques [15]–[17]. The patches are obtained from a averaged and registered less noisy image and a dictionary is created using these patches. This dictionary is then utilized to denoise a nearby noisy image [15]. Some techniques use a dictionary trained on images obtained from few subjects and denoise/improve the noisy scans obtained from other subjects [16]. This follows customized scanning pattern in which B-scans are captured at nominal SNR from adjacent positions. Here, the adjacent positions are closer in terms of azimuthal distance. These azimuthally repeated scans are then averaged and registered to obtain a denoised image.

The rationale for this approach is that neighboring B-scans, in common SDOCT volumes, are expected to have similar texture and noise pattern. In the proposed approach, we use the averaged and registered image with high signal-to-noise ratio (HSNR image) as ground-truth image in the training phase and one B-scan from the same subject as noisy, Low signal-to-noise ratio (LSNR) input image while training. A less noisy/HSNR image is not required to denoise a test image from a particular subject while testing. The HSNR image is a requirement for the dictionary-based learning methods while performing denoising [15]–[17].

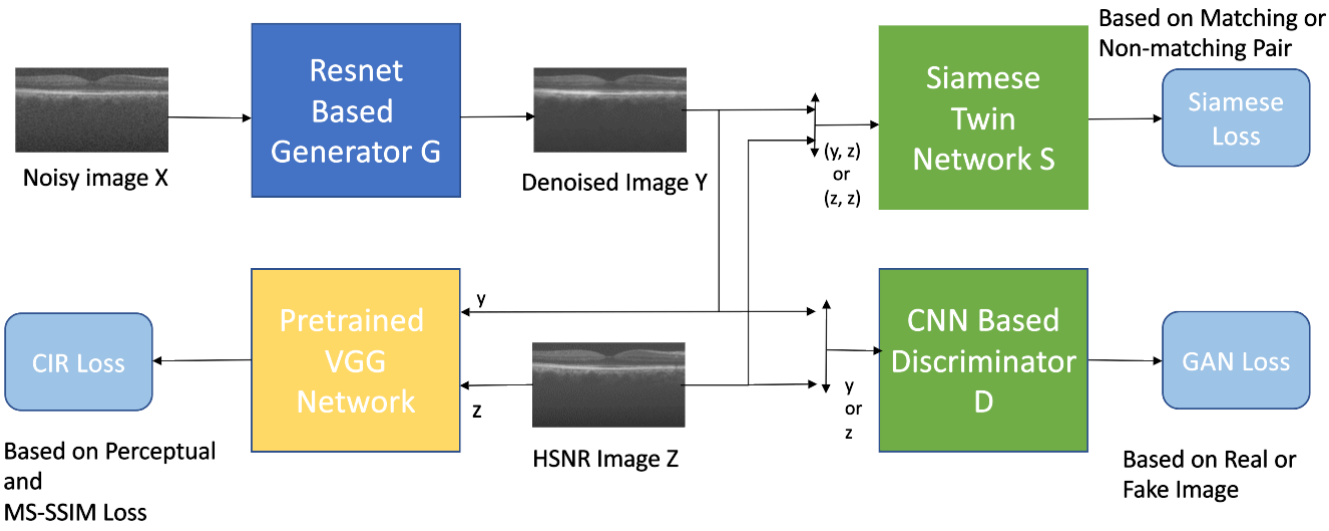

Fig. 1. SiameseGAN architecture. A pictorial depiction of generator *G*, discriminator *D* and siamese network *S*. The details of siamese network *S* has been provided in Fig. 2 and the data-flow in the training is given in Algorithm-1.

III. The Proposed Model: SiameseGAN

The proposed method is based on Wasserstein GAN that uses the Wasserstein distance instead of the Jensen-Shannon (JS) divergence to compare data distributions. The proposed model consists of three networks, Generator G, Discriminator D and a Siamese Network S (Fig. 1). Generator learns a mapping $G : x \mapsto \hat{y}$, where x is the raw, low SNR (LSNR) image and $\hat{y}$ is the denoised image generated by G. The discriminator's objective is to differentiate the fake image pair $\hat{y}$ from the real, averaged high SNR image (HSNR) y.

A. GAN Objective

In the Wasserstein GAN (WGAN), the Earth-Mover (EM) distance or Wasserstein metric between the generated image samples and real data gets minimized. The reason for choosing this metric is that it is continuous and differentiable almost everywhere under some mild assumptions, while Kullback–Leibler (KL) and Jensen-Shannon divergence do not satisfy these conditions. The training of G against D forms the adversarial part of the objective, and can be expressed as

$$\min_G \max_D L_{GAN}(D, G) = -\mathsf{E}_y[D(y))] + \mathsf{E}_x[D(\hat{y})] + \lambda \mathsf{E}_{\hat{x}}[(\|\nabla_{\hat{x}}(D_{\hat{x}})\|_2 - 1)^2] \ , \quad (1)$$

where the first two terms perform a Wasserstein distance estimation; the last term is the gradient penalty term for network regularization; $\hat{x}$ is uniformly sampled along straight lines connecting pairs of generated and real samples; and λ is a constant weighting parameter. The networks D and G are trained alternatively by fixing one and updating the other.

B. Combined Image Restoration Loss

We incorporate two losses in our model: (i) perceptual loss [20] obtained by extracting features from VGG network [21], and (ii) MS-SSIM loss [22].

1) Perceptual Loss: While generator in the GAN transforms the data distribution from high noise to a low noise version, we incorporate perceptaul loss instead of per pixel loss to retain image information content. The use of the perceptual loss for WGAN facilitate producing sharper details with significant reduction in noise.

The pretrained VGG network [21] extracts the features from the denoised image patch and the HSNR patch. The patches have been duplicated to make RGB channels before they are fed to the VGG network. The rational behind this is the pretrained VGG network takes color images as input while our SDOCT images are in grayscale. The VGG-19 network contains 16 convolutional layers followed by 3 fully-connected layers. The output of the 16$^{\text{th}}$ convolutional layer is the feature extracted by the VGG network and used in the computation of perceptual loss function. The perceptual loss is defined as,

$$L_{\text{VGG}}(G) = \mathsf{E}_{(x,y)}\left[\frac{1}{whd}\left\|\text{VGG}(\hat{y}) - \text{VGG}(y)\right\|_F^2\right]$$

where w, h and d are width, height and depth of the feature space, respectively.

2) MS-SSIM Loss: MS-SSIM metric is a multiscale extension of the structural-similarity metric (SSIM). This metric's pixel wise gradient has a simple analytical form and is computationally inexpensive. The SSIM family of metrics compares corresponding pixels and their neighborhoods in two patches, denoted y and $\hat{y}$, with three comparison function: luminance (I), contrast (C), and structure (S) defined as

$$I(y, \hat{y}) = \frac{2\mu_y\mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1}, \ C(y, \hat{y}) = \frac{2\sigma_y\sigma_{\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2} \ \text{and} \ S(y, \hat{y}) = \frac{\sigma_{y\hat{y}} + C_3}{\sigma_y\sigma_{\hat{y}} + C_3},$$

where $\mu_y, \mu_{\hat{y}}$ denote the mean pixel intensities, and $\sigma_y, \sigma_{\hat{y}}$ denote the standard deviations of pixel intensity in a local image patch centered at either y or $\hat{y}$. The variable $\sigma_{y\hat{y}}$ denotes

the sample correlation coefficient between corresponding pixels in the patches centered at y and $\hat{y}$. The constants C_1, C_2 and C_3 are small values added for numerical stability. The three comparison functions are combined to form the SSIM score as

$$\text{SSIM}(y, \hat{y}) = I(y, \hat{y})^{\alpha} C(y, \hat{y})^{\beta} S(y, \hat{y})^{\gamma}. \quad (2)$$

The SSIM score is a single scale measure. The input patches are iteratively down sampled by a factor of two with a low-pass filter, with scale j denoting the original images down sampled by a factor of 2^{j-1}. This multiscale SSIM variant is given as

$$MS-SSIM(y, \hat{y}) = I_M(y, \hat{y})^{\alpha M} \prod_{j=1}^{M} C_j(y, \hat{y})^{\beta_j} S_j(y, \hat{y})^{\gamma_j}$$

The objective is to minimize the loss related to the MS-SSIM score of the patches and is given as follows:

$$L_{\text{MS-SSIM}}(G) = -\mathsf{E}_{(x,y)}[\text{MS} - \text{SSIM}(\hat{y}, y)],$$

where x is the noisy (LSNR) input patch, y is the groundtruth (HSNR) patch and $\hat{y}$ is the denoised patch generated as output by the generator. The combined image restoration objective is

$$L_{\text{CIR}}(G) = L_{\text{VGG}}(G) + L_{\text{MS-SSIM}}(G). \quad (3)$$

C. Siamese Network Loss

The siamese network takes generated and groundtruth image pair $(\hat{y}, y)$ and classifies it as non-matching class and takes groundtruth image (y, y) and classifies it as a matching pair. Now the generator is not only has to fool the discriminator, but also this siamese network. Hence generator is forced to generate output that always gets classified into matching pair. The training of generator also forms the adversarial part of the objective that can be expressed as

$$\min_G \max_S L_{\text{Siamese}}(S, G) = \mathsf{E}_y[\log(1 - S(y, y))] + \mathsf{E}_x[\log(S(\hat{y}, y))] \quad (4)$$

where, these two terms in the above equation perform Wasserstein distance estimation.

D. Overall Objective Function

By combining above (1), (3) and (4), the final objective function can be written as

$$\min_G \max_{D,S} L_{GAN}(D, G) + \alpha L_{\text{Siamese}}(S, G) + \beta L_{\text{CIR}}$$

where α and β are scaling factors for the siamese network loss and combined restoration loss respectively.

The steps/data-flow involved in training of SiameseGAN has been listed in Algorithm-1.

Algorithm 1 Major Steps in SiameseGAN Training

```
epochs=max_epoch, batches=number of batches
G:Generator(), D:Discriminator(), S:Siamese Network()
CriticUpdates=5, BatchSize=1
FinalModel=FinalModel(G, D, S)
// final model as shown in Fig. 1
for epoch = 1 to epochs do
    for batch = 1 to batches do
        x := raw noisy image from the batch
        // raw noisy image
        y := corresponding averaged HSNR image
            // HSNR image
        z := G.predict(x)
        // generated denoised image
        for critic update to CriticUpdates do
            // discriminator training on real image
            D.train(y)
            // discriminator training on fake image
            D.train(z)
            // siamese training on matching images
            S.train(y, y)
            // siamese training on non-matching images
            S.train(y, z)
        D.trainable:= False
        S.trainable:= False
        // here the generator is trained
        // using losses obtained from
        // other modules keeping them non-trainable
        FinalModel.train(x, y)
        D.trainable = True
        S.Trainable = True
```

IV. Details of Network Architecture

A. Generator

Generator in a GAN can be implemented with various neural network architectures for image-to-image translation tasks. We have performed experiments with U-Net [23] and residual net [24].

1) Residual Network Architecture: Using a very deep neural network can improve the performance, however these networks are very difficult to train and one can observe degradation of generated images. To overcome this we employ residual neural networks [24]. The residual neural network consists of a series of stacked residual units and each residual unit consisted multiple combination of convolutional, batch-normalization and activation layers.

The layer combination of residual unit utilized in this work consists of two 3×3 convolutional layers, each followed by batch-normalization layer with a rectified linear unit (ReLU) activation layer in the middle. The network consists of two 3×3 convolutional layer with stride 2 and filter size 64 and 128, respectively followed by residual units. The number of residual units is considered as hyperparameter in all experiments and it has been tuned accordingly. Residual blocks are followed by two 3×3 deconvolutional layer with stride 2 and

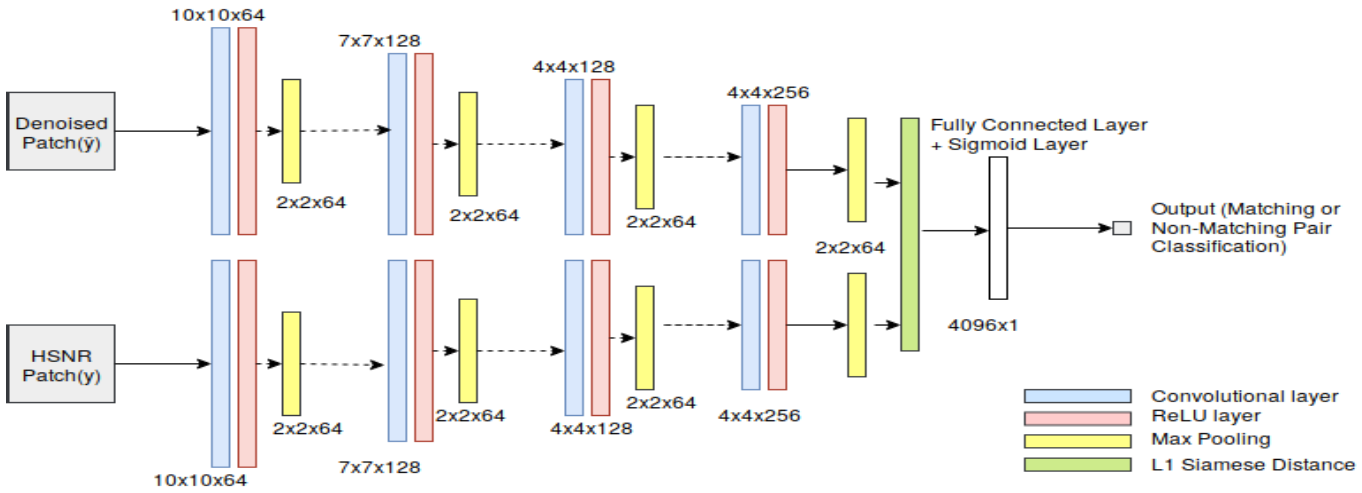

Fig. 2. Siamese network (*S*) architecture that was utilized as part of the proposed denoising network. This forms one block of proposed SiameseGAN model as shown in Fig. 1.

filter size 128 and 64, respectively. The ResNet forms the generator part (G) of the proposed SiameseGAN network model given in Fig. 1.

2) Deep Residual U-Net Architecture: The U-Net architecture [25] also uses skip connection to facilitate the training of the network. The architecture consists of an encoder consisting of convolutional layers and the decoder consisting of deconvolutional layers which uses skip connections by concatenating the output of the deconvolution layers with the feature maps from the encoder at the same level.

Instead of the plain convolutional blocks, residual units have been used in U-Net architecture [25]. The advantage is that the skip connections within a residual unit and between low levels and high levels of the network will facilitate information propagation without degradation. All residual units consist of two convolutional block and each convolutional block has a batch-normalization layer, a ReLU activation layer and a 3×3 convolutional layer with stride 2.

B. Combined Image Restoration Loss Calculator

As mentioned earlier, in the proposed approach, instead of per-pixel loss, a combined loss was minimized in the proposed model. A denoised output image ($\hat{y}$) from the generator G and the ground truth (HSNR) image (y) are fed to the pre-trained VGG network for feature extraction. Then, the perceptual loss was computed using the extracted features from a specified layer. This perceptual loss was combined with a differential MS-SSIM loss function. The computed reconstruction error was then back-propagated to update the weights of G only, while keeping the VGG parameters intact.

C. Siamese Network

The siamese twin network [26] S takes two image patches as inputs. It consists of a sequence of convolutional layers, each of which utilizes a single channel with filters of varying size and a fixed stride of 1. The kernel sizes of the convolutional layers are 10×10, 7×7 and 4×4 respectively. The number of convolutional filters was specified as a multiple of 16 to optimize performance. The network applies a ReLU activation function to the output feature maps, followed by max-pooling with a filter size 2 and stride 2. The twin network joins immediately after the 4096 unit fully-connected layer, where the L1 component-wise distance between vectors is computed. The siamese network architecture has been depicted in Fig. 2.

D. Discriminator

The discriminator (D) in the proposed model consists of a sequence of convolution layers, each of which is followed by batch normalization (BN) and leaky ReLU (LReLU) nonlinearity, except for the first and last layers. For the first layer, there is no batch normalization layer between the convolution layer and its activation. For the last layer, only the convolution layer exists. All convolutional layers have kernel size of 4×4. There are convolutional blocks consisting of a convolutional layer followed by batch normalization layer and leaky ReLU layer between the first and last convolutional layer. The number of these convolutional blocks (n) was considered as a hyper-parameter and was tuned accordingly. The number of filter in each convolutional layer was fixed to 64 for all experiments performed in this work.

V. Experiments

A. Data

Dataset-1: The first SDOCT dataset that was considered in this work is part of the study presented in [15]. These SDOCT images were acquired at 840-nm wavelength Bioptigen, Inc. imaging system. Total of 28 patients, where images from 28 eyes, with and without non-neovascular age-related macular degeneration (AMD) have participated in this study with an axial resolution of $\sim 45\mu m$ per pixel (450×900 (height $\times$ width)) in tissue. Two scans per subject were acquired, first one was a volumetric scan of retinal fovea with 1000 A-scans per B-scan and 100 B-scans. The second scan was also 1000 A-scans per B-scan, but with

40 azimuthally repeated B-Scans centered at the fovea. Test images were chosen to be central foveal B-scan within the first volume. StackReg image registration plug-in [27] for ImageJ (software; National Institutes of Health, Bethesda, MD, USA) was utilized to register azimuthally repeated B-scans to form noiseless HSNR (ground truth) averaged image. More concretely, in each image pair, there was a noisy SDOCT image captured by a Biopitgen SD-OCT imaging system, and a clear OCT image which was acquired by registering and averaging several B-scans obtained at the same position. Since a large number of SD-OCT image pairs were required to train GAN based deep learning models, data augmentation techniques such as flip and rotation were utilized to create more data points. Out of 28 image pairs available, 10 image pairs were selected and rectangular patches with window size 100×200 and stride of 10 were created. After preprocessing these patches of images were used to train all deep learning models.

Dataset-2: To validate the robustness of the proposed model SiameseGAN, a second dataset has been employed in our experiments that was previously utilized in [16]. One of the Bioptigen SDOCT imagers used in the Dataset-1 collection with axial resolution of $\sim 45\mu m$ per pixel in tissue and directly acquired full and subsampled volumes from 13 human subjects with regularly sampled pattern in clinic were utilized. That is, for each subject a square volume centered at the retinal fovea with 500 A-scans per B-scan and 100 B-scans per volume were scanned. For reconstruction comparisons involving real experimental datasets of human subjects, from each dataset, the central foveal B-scan as well as two additional B-scans located approximately 1.5 mm above and below the fovea, have been used. Therefore, 39 raw images were available in this Dataset-2 without their corresponding ground truth (HSNR) images. Since the reference images were not available and all 39 raw images were considered as test cases and these has not been utilized in the training of the proposed model.

For training, a workstation with Dual Intel(R) Xeon(R) CPU $E5-2630$ with 2.40 GHz clock speed along with two NVIDIA GeForce GTX1080 12GB GPUs having 64GB RAM was utilized. The typical training time for each of the deep learning model, including the proposed SiameseGAN, was approximately two hours.

B. Evaluation Metrics

For a comparative evaluation of the the denoising performance of proposed model Siamese GAN, we have utilized the following figures of merit.

1) Peak Signal-to-Noise Ratio (PSNR): The PSNR is a figure of merit that provides the measure of fidelity in the processed image with respect to the reference image. This is defined as

$$\mathrm{PSNR} = 20 \cdot \log_{10}\left(\frac{MAX_R}{\sqrt{\frac{1}{H}\sum_{h=1}^{H}(R_h - \hat{R}_h)^2}}\right),$$

where R_h is the intensity of the hth pixel in the reference HSNR image R, $\hat{R}_h$ represents the same hth pixel of the recovered denoised image $\hat{R}$, H is the total number of pixels, and MAX_R is the maximum intensity value of image R.

2) Structural Similarity Index (SSIM): When comparing images, the mean squared error (MSE) is not highly indicative of perceived similarity. Structural similarity aims to address this shortcoming by taking texture into account. It is a perceptual metric that quantifies image quality degradation caused by processing. It is a full reference metric that requires two images from the same image capture: a reference image and a processed or noisy version of the image. Unlike PSNR, SSIM is based on visible structures in the image. SSIM score can be defined as

$$\mathrm{SSIM}(x, y) = I(x, y)^{\alpha} C(x, y)^{\beta} S(x, y)^{\gamma},$$

where x and y are the two image centers, α, β and γ are the constants and set to 1 (throughout this work) similar to the work presented on [28]. Here, I, C and S refers to luminance, contrast and structure as defined in Eq. 2.

3) Mean Signal to Noise Ratio (MSR) and Contrast to Noise Ratio (CNR): Reference image is not required to compute MSR and CNR of a particular image since they are region based metrics. The metrics MSR and CNR are defined as

$$\mathrm{MSR} = \frac{\mu_f}{\sigma_f}, \quad \text{and} \quad \mathrm{CNR} = \frac{|\mu_f - \mu_b|}{\sqrt{0.5 * (\sigma_f^2 + \sigma_b^2)}},$$

where μ_b and σ_b denote the mean and the standard deviation of the background region; μ_f and σ_f denote the mean and the standard deviation of the foreground regions. To compute MSR or CNR of a single image, mean of MSR or CNR was computed for all foreground region of interests selected in the image. The background and foreground regions were illustrated as blue and red boxes respectively as shown in Figs. 3, 4 and 5.

4) Texture Preservation (TP) Index and Edge Preservation (EP) Index: Similar to MSR and CNR, Texture Preservation (TP) index and Edge Preservation (EP) index do not require reference image for calculation as they are region based metrices. TP can be computed as

$$\mathrm{TP_m} = \frac{\sigma_m^2}{(\vec{\sigma_m})^2}\sqrt{\frac{\mu_{den}}{\mu_{in}}},$$

where, m represents the m-th region of interest (ROI), $(\vec{\sigma_m})^2$ stands for the standard deviation of respective region of the raw image. μ_{den} and μ_{in} denote the mean value of the denoised and noisy image respectively. The TP value considered is the average over all the ROIs. If denoised image features are severely flattened then TP value appears close to 0. We compute EP as

$$\mathrm{EP} = \frac{\Gamma(\Delta I'_m - \overline{\Delta I'_m}, \Delta I_m - \overline{\Delta I_m}}{\Gamma(\Delta I'_m - \overline{\Delta I'_m}, \Delta I'_m - \overline{\Delta I'_m})\Gamma(\Delta I_m - \overline{\Delta I_m}, \Delta I_m - \overline{\Delta I_m})},$$

where m represents the m-th ROI region. I_m, I'_m represents the matrices of the corresponding region of the denoised and input noisy image respectively. $\overline{I}$ represents the empirical mean of I. Δ is a Laplacian operator and ΔI is obtained with a standard 3×3 approximation of the Laplacian operator. Γ shows correlation between two ROIs as

$$\Gamma(\mathrm{I}_1, \mathrm{I}_2) = \sum_{i,j \in ROI} I_1(i, j) I_2(i, j).$$

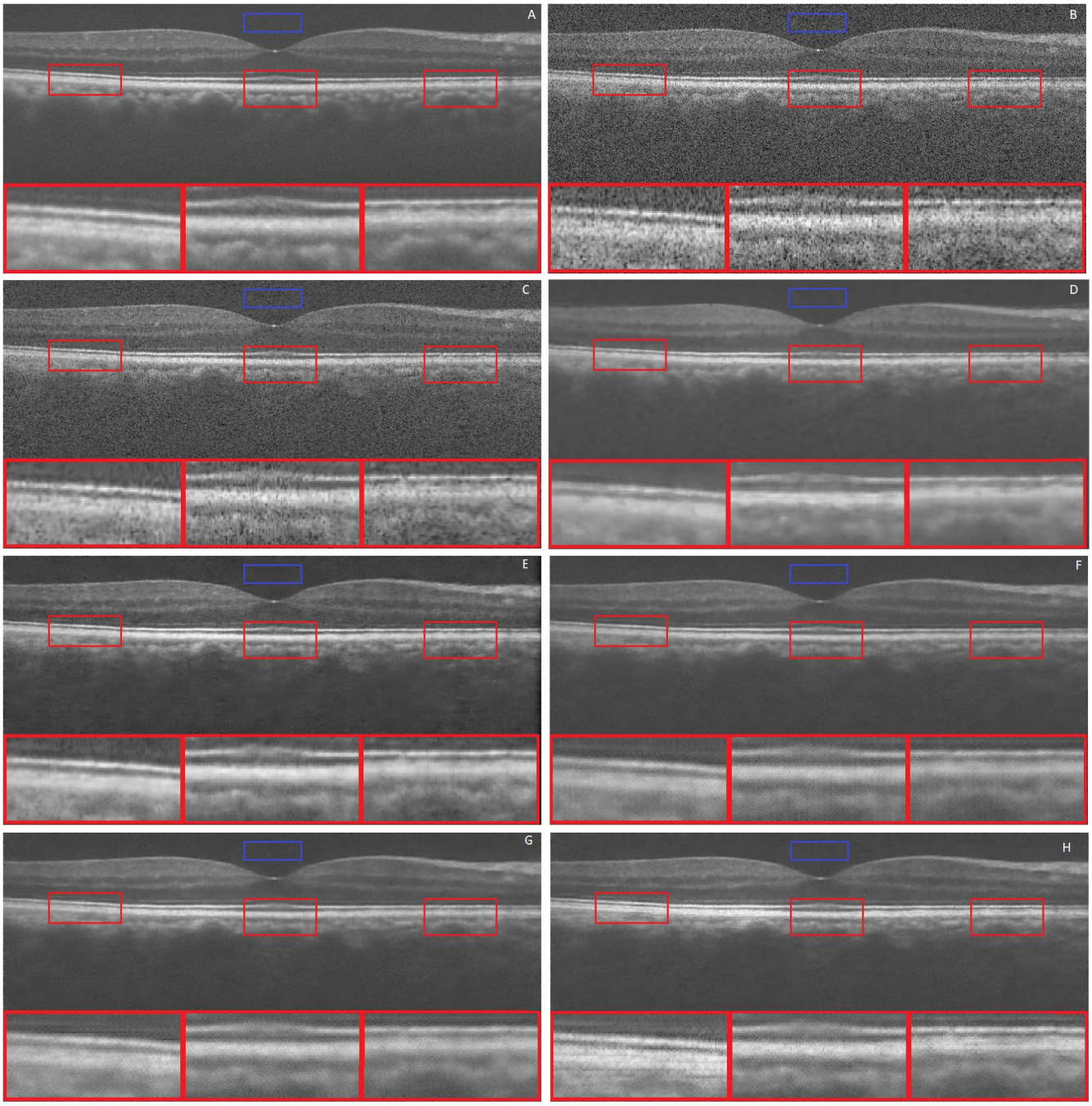

Fig. 3. SDOCT foveal images corresponding to Patient-6 of dataset-1. A: averaged (HSNR) image, B: noisy image, C: MSBTD result, D: MIFCN result, E: SE result, F: WGAN-UNET result, G: WGAN-Resnet result, H: proposed SiameseGAN result. The top most blue box in each image corresponds to the background patch that was utilized in computing figures of merit. The zoomed versions of patches covered by red boxes (foreground) are shown at the bottom of corresponding image to clearly show the improvement obtained in image quality. The PSNR values for these images were given in Table-II.

We average the EP value over the ROIs and its value lies between 0 and 1. When EP value is closer to 0, it means that the edges in the ROI are blurred. The mean of EP and TP was computed over all the foreground regions for an image.

C. Experimental Results

To evaluate the performance of the proposed model SiameseGAN, quantitative and qualitative comparison has been made with state-of-art SDOCT denoising methods, multiscale sparsity based tomographic denoising (MSBTD) [15] and state-of-the-art deep learning based image denoising methods that include multi-input fully-convolutional network (MIFCN) [18], shared encoder (SE) architecture with multiple decoders [29] and Wasserstein GAN with perceptual loss based training (WGAN-ResNet) [30]. Experiments have also been performed by training a Deep Residual UNET architecture as the generator (WGAN-Unet) and combining the perceptual loss with

TABLE I
AVERAGE OF THE PSNR (DB), SSIM, MSR, CNR, TP AND EP FOR EIGHTEEN SDOCT FOVEAL IMAGES FROM DATASET-1 OBTAINED FROM THE MSBTD, MIFCN, SHARED ENCODER (SE), WGAN-UNET, WGAN-RESNET, AND PROPOSED SIAMESEGAN. THE BEST RESULTS ARE SHOWN IN BOLD. * INDICATES THAT RESPECTIVE METRIC DOES NOT REQUIRE GROUND TRUTH

Method	PSNR	SSIM	MSR*	CNR*	TP*	EP*
MSBTD	27.08	0.56	2.68	1.12	0.36	0.29
MIFCN	26.56	0.72	3.57	2.34	**0.74**	0.57
SHARED ENCODER	24.21	0.75	3.20	2.37	0.69	**0.67**
WGAN UNET	26.26	0.61	3.74	2.36	0.60	0.49
WGAN RESNET	25.81	0.80	3.65	2.52	0.63	0.51
SiameseGAN (Proposed)	**28.25**	**0.83**	**4.24**	**2.60**	0.68	0.66

MS-SSIM loss for training. The networks that have been trained for the comparison are:

- MIFCN: Multi-input fully-convolutional network [18],
- SE: Shared encoder architecture with multiple decoders [29],
- WGAN-Unet: Wasserstein GAN with UNET network as Generator with only Perceptual Loss [30],
- WGAN-ResNet: Wasserstein GAN with 16 layer Residual Net network as Generator and perceptual Loss combined with MS-SSIM, and
- SiameseGAN (Proposed): The Proposed Model (network architecture was shown in Fig. 1) with MS-SSIM loss function

Note that for all these deep learning models, same data (10 patients data from dataset-1) was utilized in the training phase. The quantitative analysis of MIFCN provided in [18] shows that it performs better than K-SVD denoising algorithm [31], block matching and 3-D filtering (BM3D), spatially adaptive iterative singular-value thresholding (SAIST), patch group based Gaussian mixture model (PG-GMM), block matching and 4-D filtering (BM4D), and segmentation based sparse reconstruction (SSR). The shared encoder (SE) based model was also proven to be more effective compared to sparsity based simultaneous denoising and interpolation (SBSDI) [16] as well as Complex Wavelet- based Dictionary Learning (CWDL) [32]. For this reason, MIFCN and SE were compared to the proposed SiameseGAN as they provide better performance than standard methods as discussed above.

1) Quantitative Analysis: Table-I shows the quantitative results for the discussed methods on 18 SDOCT test images corresponding to Dataset-1. These results include the mean of PSNR metric, SSIM index, MSR, CNR, EP index and TP index for the 18 SDOCT test images from Dataset-1. The results indicate that the proposed SiameseGAN out performs others clearly on the all figures-of-merit except for the TP index. The MIFCN preserves the texture better than the all other methods. The MIFCN utilizes the weighted averaging through a multi-branch network, performing similar to non-local means method, enabling texture preservation from the neighbors [18].

TABLE II
PSNR (IN DB) CORRESPONDING TO THE METHODS DISCUSSED HERE FOR EIGHTEEN INDIVIDUAL SDOCT FOVEAL IMAGES OF DATASET-1 THAT WERE USED AS TEST CASES. REPRESENTATIVE IMAGES FOR PATIENTS-6 AND 8 ARE AVAILABLE IN FIGS. 3 AND 4 RESPECTIVELY

Image/Method	MSBTD	MIFCN	SE	WGAN-UNET	WGAN-Resnet	SiameseGAN (Proposed)
1	29.58	27.14	25.54	28.41	28.54	**31.73**
2	23.22	29.41	21.34	22.82	22.31	**33.23**
3	23.60	22.70	23.38	25.18	24.17	**26.06**
4	27.78	28.91	25.47	26.64	26.49	**30.78**
5	26.92	26.30	25.18	27.17	27.32	**32.19**
6	27.43	28.34	24.57	26.85	26.20	**28.59**
7	26.47	27.34	22.20	25.39	25.05	**29.90**
8	23.34	28.37	22.21	22.86	22.65	**27.43**
9	**29.44**	23.40	25.92	26.01	27.23	23.19
10	**28.99**	27.07	23.57	27.14	25.83	27.26
11	28.78	27.38	25.22	27.59	27.43	**30.39**
12	**25.23**	21.90	23.60	25.12	24.84	23.02
13	26.95	26.73	21.38	25.16	24.78	**30.85**
14	**29.57**	28.53	28.45	28.80	28.59	26.81
15	23.07	**29.26**	22.42	22.62	21.38	26.90
16	**30.69**	24.43	26.26	29.68	28.38	30.33
17	27.12	22.36	24.94	**27.42**	26.32	21.92
18	**29.20**	28.55	24.08	27.85	27.17	27.95
Mean PSNR	27.08	26.56	24.21	26.26	25.81	**28.25**

One can observe that adding siamese network module and incorporating MS-SSIM loss to WGAN network has improved the performance compared to the baseline model. The SiameseGAN improved the mean PSNR by 1.17 dB and mean SSIM index by 0.27 thus achieving better performance than other methods. The proposed model has also improved the CNR index as well as MSR index proving that it was able to capture low-level finer details as well as high level semantic features. Our model also has better EP than other models showing it preserves edges without making them blurry. Table-IV showing the computational time (for the denoising step) indicates that the deep learning based denoising methods works faster than dictionary based denoising methods. Table II shows the PSNR metric for 18 individual SDOCT test images for the discussed method for knowing performance at the individual image level.

2) Qualitative Analysis: For qualitative and visual analysis, Fig. 3 and 4 illustrate the results obtained for two test images from the Dataset-1. In Fig. 3 and 4, the image A, B are the reference image (ground truth), raw (noisy) image respectively. Images C, D, E, F, G and H are resultant denoised image obtained from MSBTD method, MIFCN method, Shared Encoder (SE) method, WGAN-UNET method, WGAN-ResNet method and proposed SiameseGAN respectively. One background region and three foreground region have been selected from each of the above images for better visual analysis. These same regions have been utilized in computing MSR, CNR, TP and EP for quantitative comparison and averaged results of the same were provided in Table I.

As dataset-2 did not have HSNR (ground trutn) images, MSBTD method could not applied to these images. Even in terms of evaluation metrics, PSNR and SSIM can not be computed without the HSNR images, so the quantitative comparison included only MSR, CNR, TP, and EP metrics. In terms of qualitative comparison, two patients

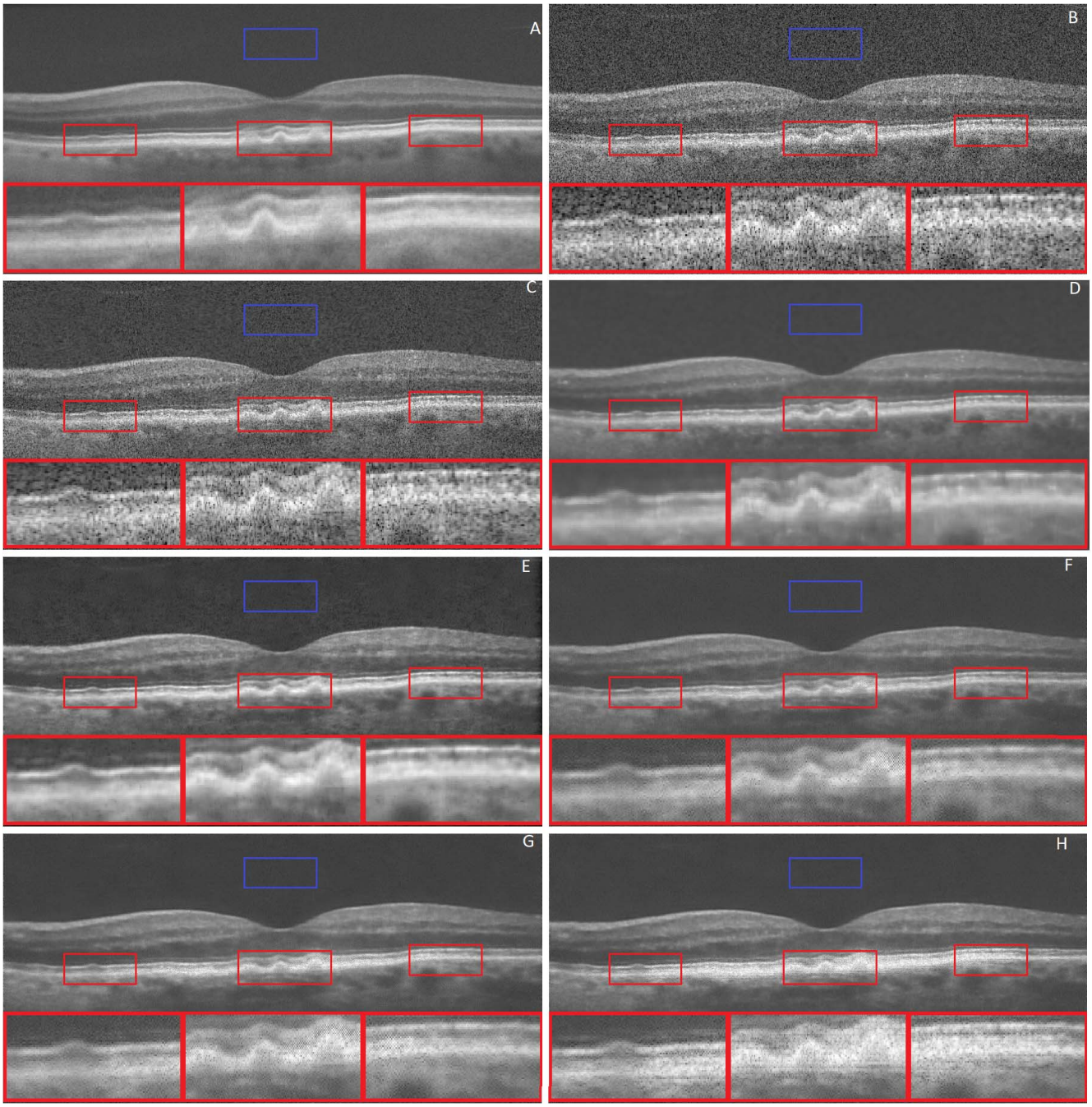

Fig. 4. Same effort as Fig. 3 corresponding to Patient-8 of dataset-1. A: averaged (HSNR) image, B: noisy image, C: MSBTD result, D: MIFCN result, E: SE result, F: WGAN-UNET result, G: WGAN-Resnet result, H: proposed SiameseGAN. The PSNR values for these images were given in Table-II.

results were presented in Fig. 5. Images A and G are two raw (LSNR) images out of thirty nine from the dataset-2 (where reference/ground truth images were absent) of patient-1 and patient-2 respectively. Images B, C, D, E and F are the resultant denoised images for patient-1 obtained using MIFCN, SE, WGAN-UNET, WGAN-ResNet and proposed SiameseGAN respectively. Images H, I, J, K and L are the resultant denoised images for patient-2 obtained using MIFCN, SE, WGAN-UNET, WGAN-ResNet and SiameseGAN respectively. One background and three foreground regions were selected for all these images for computation of MSR, CNR, TP and EP as in previous images. The region of interests in each image have been zoomed and shown below for better comparison. These suggest that the structural integrity was better preserved using proposed SiameseGAN, while other methods result in blurring of edges and missing other finer details in denoised images. Note that even though only two example cases were presented here, the results obtained for other cases followed the same trend as observed in these example cases shown in Fig. 5. The averaged MSR, CNR, TP and EP values obtained using our model for the SDOCT foveal images obtained from dataset-2 have been provided in

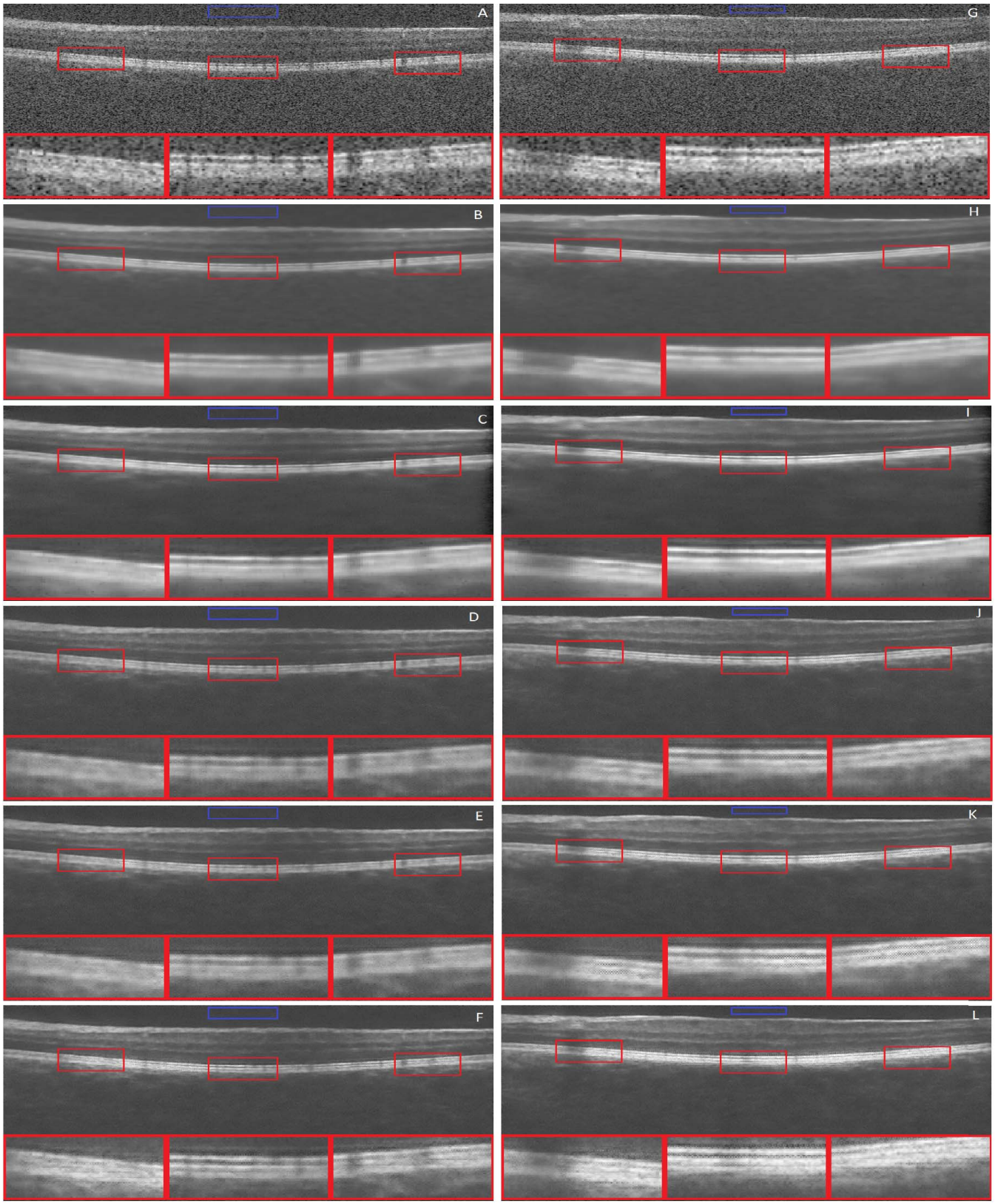

Fig. 5. Example SDOCT foveal images from dataset-2. Images A, B, C, D, E, F belong to patient-1 and images G, H, I, J, K, L belong to patient-2. Row wise (top to bottom) these correspond to the noisy image, MIFCN result, SE result, WGAN-UNET result, WGAN-Resnet result, proposed SiameseGAN result respectively. The top most blue box in each image corresponds to the background patch that was utilized in computing figures of merit. The zoomed versions of patches covered by red boxes (foreground) are shown at the bottom of corresponding image to clearly show the improvement obtained in image quality. The averaged figures-of-merit (MSR, CNR, TP and EP) were presented in Table-III.

TABLE III
AVERAGE OF THE MSR, CNR, TP, ENL, EP FOR SDOCT FOVEAL IMAGES FROM DATASET-2 (39 SDOCT IMAGES) OBTAINED FROM THE MIFCN, SHARED ENCODER, WGAN-UNET, WGAN-RESIDUALNET, PROPOSED SIAMESEGAN. THE BEST RESULTS ARE SHOWN IN BOLD. ALL THESE METRICS DO NOT REQUIRE GROUNDTRUTH. EXAMPLE (FOR PATIENTS-1 AND 2) RAW AND DENOISED IMAGES WERE PRESENTED IN FIG. 5

Method	MSR*	CNR*	TP*	EP*
MIFCN	3.45	3.10	**0.72**	0.44
SHARED ENCODER	2.79	**2.93**	0.64	0.48
WGAN UNET	4.69	2.26	0.57	0.46
WGAN RESNET	4.42	2.38	0.62	0.40
SiameseGAN (Proposed)	**5.30**	2.36	0.67	**0.57**

TABLE IV
COMPARISON OF THE COMPUTATIONAL TIME (IN SECONDS) FOR DENOISING USING THE DISCUSSED METHODS FOR THE ENTIRE TEST DATA: DATASET-1 (18 IMAGES OF SIZE 450 × 900) AND DATASET-2 (39 IMAGES OF SIZE 450 × 450)

Method	Dataset1	Dataset2
MSBTD	9-10 hours	-
MIFCN	47.85	41.61
SHARED ENCODER	19.64	18.33
WGAN UNET	26.66	26.96
WGAN RESNET	25.68	26.61
SiameseGAN (Proposed)	22.47	28.59

Table-III. These results (from datasets-1 and 2) indicate that the performance of the proposed SiameseGAN was superior compared to the standard MSBTD as well as other deep learning models. The proposed model was able to generalize across the two available datasets and provided improved performance.

3) Ablation Study With Respect to Loss Functions: To understand the contribution of each loss that was utilized in the network, we performed an ablation study [33] utilizing eighteen SDOCT foveal images from dataset-1 as the test data. Table V provides the results from this ablation study. It contains figures of merit, same as presented in Table I, and the first column gives the loss function that was not utilized while training the proposed model. The last row of Table V corresponds to results from proposed SiameseGAN (same as results in Table I) when trained with all loss functions being included. From Table V, it is clear that, inclusion of MS-SSIM loss and Siamese loss significantly improves the edge preservation index of denoised images, preserving structural integrity and also capturing the finer details. Inclusion of all loss functions significantly improved the figures of merit. The MSR is high when Siamese loss was not utilized, as this metric provides signal to noise ratio of the foreground only and does not account noise present for the background region.

TABLE V
RESULTS OF THE ABLATION STUDY WHEN MODEL WAS TRAINED WITHOUT RESPECTIVE LOSSES WITH TEST DATA BEING EIGHTEEN SDOCT FOVEAL IMAGES FROM DATASET-1. THE BOTTOM ROW RESULTS ARE FOR THE PROPOSED MODEL TRAINED WITH ALL LOSS FUNCTIONS INCLUDED. THE BEST RESULTS ARE SHOWN IN BOLD

Without	PSNR	SSIM	MSR*	CNR*	TP*	EP*
MS-SSIM loss	27.10	0.80	4.18	2.24	0.57	0.39
Perceptual loss	26.56	0.78	3.81	1.94	0.67	0.64
Siamese loss	24.21	0.81	**4.8**	1.73	0.62	0.43
SiameseGAN (Proposed)	**28.25**	**0.83**	4.24	**2.60**	**0.68**	**0.66**

VI. DISCUSSION

Though there are several classical denoising methods, like BM3D [34], MSBTD [15], SBSDI [16], curvelet transform based dictionary learning technique [17] have been proposed for denoising spectral domain OCT data, none of these approaches make use of deep neural networks which are powerful function approximators. On the otherhand, a few deep learning based methods have been used in denoising to despeckle OCT images of Optic Nerve Head [35], deblurr retinal images [36]. Edge-sensitive conditional GAN has also been proposed to denoise 3D OCT volumetric data [37].

Deep convolutional neural network [38] and WGAN using perceptual loss [30] based methods have been proposed for low-dose CT image denoising [39]. Above deep learning based methods do not consider denoising of the multiplicative, speckle noise that heavily corrupts SDOCT images. Also, this is the first work in the denoising of SDOCT images that has not only proposed a network that was specifically designed for the job at hand and provided a comparison with popular deep learning models to show the efficacy the proposed SiameseGAN.

The proposed SiameseGAN model has the distinctive advantage of combining restoration loss (perceptual loss and MS-SSIM loss) with siamese loss which forces the denoised image to be close to the expected image. It provides better fidelity as not only the generated denoised image has to go through the discriminator, but also siamese twin network to provide a matching pair. This adds to the robustness of the proposed SiameseGAN to provide improved denoised SDOCT images. The same has been reasserted through ablation study performed in this work (Table- V).

It is important to note that once the model has been trained, the requirement of having HSNR image, unlike MSBTD method, does not arise at all. Even though the model was trained using ten images of Dataset-1, it was able to generalize and provide improved denoised results for Dataset-2 images (Fig. 5 and Table III).

The output of the denoised methods, including proposed SiameseGAN, utilizing a single B-scan noisy image will never be equal to the averaged (HSNR) image. The averaging and registering to form the HSNR image involves several B-scan images with number of such images being in the order of hundred. The denoising methods will only be capable of providing a close estimate to this HSNR image and thus

comparison of the quality of HSNR image with denoised method output should be performed within the context of figures of merit, which provides an objective way of knowing the image quality. The results presented in this work shows that among the presented denoised methods, the proposed SiameseGAN provides the best performance for the task at hand.

Typical analysis of SDOCT images especially for Age-Related Macular Degeneration (AMD) involves the study of drusen layer [40]. The druse entity will be considered as pigment epithelial elevation larger than 25 μm in diameter in these SDOCT images. The typical analysis will involve morphological characteristics of the druse in terms of shape of the epithelial elevation, the reflectivity and homogenity, and the presence or absence of hyperreflective foci above the druse [40]. Even though the detailed analysis is not performed in this work, the results indicate that in all these metrics, the output of proposed SiameseGAN will be able to provide accurate analysis of the morphological characteristics on par with the HSNR image, especially with automated methods. The detailed study including different complex pathologies (different stages of AMD) will be taken up as future work.

The image denoising with Wasserstein distance and perceptual Loss with a GAN has been applied in low-dose CT case [30]. The WGAN UNET and WGAN RESNET utilized in this work are similar to the model utilized for denoising of low-dose CT images [30]. The proposed SiameseGAN also utilizes the Wasserstein Distance and Perceptual Loss in addition to having MS-SSIM loss within the GAN (details are there in Sec. III.B). On top of it, the siamese network loss was also added to final objective function in the proposed SiameseGAN. The preceptual loss provides improved visual perception in the denoised image as can be seen from the presented results. SiameseGAN, with utilization of additional MS-SSIM and siamese loss, provides more robustness to the learning for improved visual perception as well as structure preservation/integrity in the denoised image.

The proposed SiameseGAN can perform on the fly denoising of SDOCT images as shown in Table-IV, average time of 0.7 second per image, making them available to the clinician in real-time, in addition to performing the denoising without the need for repetitive/averaging of SDOCT images. Methods of this type will improve the clinical adoptability of SDOCT images making them universally appealing.

Recently, a similar model called SiGAN [41] has been proposed in the context of generative model for human faces. This model is significantly different in terms of functionality from the proposed SiameseGAN model. The SiGAN takes two distinct images as input and pass them through a twin generator network, while in our case one generator network gives output which is fed to Siamese twin network to act as a better discriminator. On the same note, a siamese type network for providing domain adaptation for performing aerial vehicle image categorization was proposed [42]. This was proven to be effective to learn invariant high-level features, when the input data can vary in terms spatial and temporal resolution. This network only enabled domain adaptation (transfer of labels available in one domain to another domain) and it was termed as Siamese network as the encoder-decoder networks to provide feature representations on the labeled and unlabeled images were similar. The shared encoder (SE) that was discussed in this work provides the feature representation similar to siamese type network discussed in [42]. The performance of SE in terms of CNR was better/comparable to other deep learning models, but in all other metrics, it provided sub-optimal results. In short, the Siamese twin network considered in Ref. [42] was part of generator and another network was acting as a discriminator [42]. The discriminators utility, even in SiGAN, was limited to performing matching operation. In the proposed SiameseGAN, the siamese network will provide the similarity and was treated as part of discriminator, forcing the generator to have structures similar to ground truth. The SiGAN [41] is a pairwise learning scheme for providing super-resolution where as domain adaptation based Siamese GAN [42] performs aerial image categorization, these are not aimed at the denoising task at hand.

VII. Conclusion

In this work, we proposed a new deep generative model based on GAN for SDOCT denoising, which is equipped with an additional siamese network module. We have also experimented with state-of-the-art CNN architectures for bench marking the proposed SiameseGAN. The experimental results prove that the proposed approach can be effectively used for fast denoising of SDOCT images and provides improved performance than the traditional dictionary learning as well as other deep learning based techniques. The methods proposed here are universally appealing for mitigating the speckle noise that corrupt images. The developed models along with source code is available as open source [43] for interested users.

Acknowledgment

The authors are thankful to Prof. Sina Farsiu of Duke University for making SD-OCT images, that were utilized in this work, along with implementation code for the MSBTD method available as open-source.

References

[1] M. A. Choma, M. V. Sarunic, C. Yang, and J. A. Izatt, "Sensitivity advantage of swept source and Fourier domain optical coherence tomography," *Opt. Express*, vol. 11, no. 18, pp. 2183–2189, 2003.

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

[3] J. M. Schmitt, S. H. Xiang, and K. M. Yung, "Speckle in optical coherence tomography," *J. Biomed. Opt.*, vol. 4, no. 1, pp. 95–105, 1999, doi: 10.1117/1.429925.

[4] B. Karamata, K. Hassler, M. Laubscher, and T. Lasser, "Speckle statistics in optical coherence tomography," *J. Opt. Soc. Am. A*, vol. 22, no. 4, pp. 593–596, Apr. 2005.

[5] A. Scott, S. Farsiu, L. Enyedi, D. Wallace, and C. Toth, "Imaging the infant retina with a hand-held spectral-domain optical coherence tomography device," *Amer. J. Ophthalmology*, vol. 147, pp. 364–373, Feb. 2009.

[6] M. H. Eybposh, Z. Turani, D. Mehregan, and M. Nasiriavanaki, "Cluster-based filtering framework for speckle reduction in OCT images," *Biomed. Opt. Express*, vol. 9, no. 12, pp. 6359–6373, Dec. 2018.

[7] A. Ozcan, A. Bilenca, A. E. Desjardins, B. E. Bouma, and G. J. Tearney, "Speckle reduction in optical coherence tomography images using digital filtering," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 24, no. 7, pp. 1901–1910, 2007.

[8] J. Rogowska and M. E. Brezinski, "Image processing techniques for noise removal, enhancement and segmentation of cartilage OCT images," *Phys. Med. Biol.*, vol. 47, no. 4, pp. 641–655, Jan. 2002.

[9] S. Aja-Fernandez and C. Alberola-Lopez, "On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2694–2701, Sep. 2006.

[10] R. Bernardes, C. Maduro, P. Serranho, A. Araújo, S. Barbeiro, and J. Cunha-Vaz, "Improved adaptive complex diffusion despeckling filter," *Opt. Express*, vol. 18, no. 23, pp. 24048–24059, Nov. 2010.

[11] P. Puvanathasan and K. Bizheva, "Interval type-II fuzzy anisotropic diffusion algorithm for speckle noise reduction in optical coherence tomography images," *Opt. Express*, vol. 17, no. 2, pp. 733–746, Jan. 2009.

[12] Y. Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1260–1270, Nov. 2002.

[13] M. A. Mayer, A. Borsdorf, M. Wagner, J. Hornegger, C. Y. Mardin, and R. P. Tornow, "Wavelet denoising of multiframe optical coherence tomography data," *Biomed. Opt. Express*, vol. 3, no. 3, pp. 572–589, 2012.

[14] F. Zaki, Y. Wang, H. Su, X. Yuan, and X. Liu, "Noise adaptive wavelet thresholding for speckle noise removal in optical coherence tomography," *Biomed. Opt. Express*, vol. 8, no. 5, pp. 2720–2731, 2017.

[15] 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," *Biomed. Opt. Express*, vol. 3, no. 5, pp. 927–942, May 2012.

[16] L. Fang *et al.*, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE Trans. Med. Imag.*, vol. 32, no. 11, pp. 2034–2049, Nov. 2013.

[17] M. Esmaeili, A. Dehnavi, H. Rabbani, and F. Hajizadeh, "Speckle noise reduction in optical coherence tomography using two-dimensional curvelet-based dictionary learning," *J. Med. Signals Sensors*, vol. 7, pp. 86–91, 04 2017.

[18] A. Abbasi, A. Monadjemi, L. Fang, H. Rabbani, and Y. Zhang, "Three-dimensional optical coherence tomography image denoising through multi-input fully-convolutional networks," *Comput. Biol. Med.*, vol. 108, pp. 1–8, May 2019.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70. Aug. 2017, pp. 214–223.

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV*. Amsterdam, The Netherlands, 2016, pp. 694–711.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[22] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4277–4281.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany, 2015, pp. 234–241.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1–8.

[27] P. Thevenaz, U. E. Ruttimann, and M. Unser, "A pyramid approach to subpixel registration based on intensity," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 27–41, Jan. 1998.

[28] A. N. Avanaki, "Exact global histogram specification optimized for structural similarity," *Opt. Rev.*, vol. 16, no. 6, pp. 613–621, Nov. 2009.

[29] S. A. V. and J. Sivaswamy, "Shared encoder based denoising of optical coherence tomography images," in *Proc. 11th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2018, pp. 1–35, doi: 10.1145/3293353.3293388.

[30] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[31] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[32] R. Kafieh, H. Rabbani, and I. Selesnick, "Three dimensional data-driven multi scale atomic representation of optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1042–1062, May 2015.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[34] M. Lebrun, "An analysis and implementation of the BM3D image denoising method," *Image Process. Line*, vol. 2, pp. 175–213, Aug. 2012.

[35] S. K. Devalla *et al.*, "A deep learning approach to denoise optical coherence tomography images of the optic nerve head," *Sci. Rep.*, vol. 9, no. 1, p. 14454, Dec. 2019.

[36] 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," *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5129–5146, Nov. 2018.

[37] X. Fei, J. Zhao, H. Zhao, D. Yun, and Y. Zhang, "Deblurring adaptive optics retinal images using deep convolutional neural networks," *Biomed. Opt. Express*, vol. 8, no. 12, pp. 5675–5687, Nov. 2017.

[38] H. Chen *et al.*, "Low-dose CT denoising with convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 143–146.

[39] V. S. Kadimesetty, S. Gutta, S. Ganapathy, and P. K. Yalavarthy, "Convolutional neural network-based robust denoising of low-dose computed tomography perfusion maps," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 137–152, Mar. 2019.

[40] F. G. Schlanitz *et al.*, "Identification of drusen characteristics in age-related macular degeneration by polarization-sensitive optical coherence tomography," *Amer. J. Ophthalmology*, vol. 160, no. 2, pp. 335–344, 2015.

[41] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "SiGAN: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6225–6236, Dec. 2019.

[42] L. Bashmal, Y. Bazi, H. AlHichri, M. AlRahhal, N. Ammour, and N. Alajlan, "Siamese-GAN: Learning invariant representations for aerial vehicle image categorization," *Remote Sens.*, vol. 10, no. 3, p. 351, Feb. 2018.

[43] *GitHubRepositry*. Accessed: Jun. 20, 2020. [Online]. Available: https://github.com/sml-iisc/SiameseGAN