Enhancement of Retinal Image From Line-Scanning Ophthalmoscope Using Generative Adversarial Networks

WANYUE LI [1,2], **YI HE** [1], **WEN KONG** [1,2], **FENG GAO** [1], **JING WANG** [1,2], **AND GUOHUA SHI** [1,2,3]

[1]Jiangsu Key Laboratory of Medical Optics, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
[2]Department of Biomedical Engineering, University of Science and Technology of China, Hefei 230041, China
[3]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China

Corresponding author: Guohua Shi (ghshi_lab@126.com)

ABSTRACT A line-scanning ophthalmoscope (LSO) is a retinal imaging technique that has the characteristics of high imaging resolution, wide field of view, and high imaging speed. However, the high-speed imaging with rather short exposure time inevitably reduces the signal intensity, and many factors, such as speckle noise and intraocular scatter, further degrade the signal-to-noise ratio (SNR) of retinal images. To effectively improve the image quality without increasing the LSO system's complexity, the post-processing method of image super-resolution (SR) is adopted. In this paper, we propose a learning-based multi-frame retinal image SR method that directly learns an end-to-end mapping from low-resolution (LR) image sequences to high-resolution (HR) images. This network was validated on down-sampled and real LSO image sequences. We evaluated the method on a down-sampled dataset with the metrics of peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and perceptual distance. Moreover, the power spectra and full width at half maximum (FWHM) were used as the no-reference image quality assessment (NR-IQA) algorithms to evaluate the reconstruction results of the real LSO image sequences. The experimental results indicate that the proposed method can significantly enhance the SNR of LSO images and efficiently improve the resolution of LSO retinal images, which has great practical significance for clinical diagnosis and analysis.

INDEX TERMS Line-scanning ophthalmoscope, retinal images, multi-frame image super-resolution, learning-based.

I. INTRODUCTION

Retinal imaging is one of the most common modalities of clinical practice in diagnosing retinal diseases. The confocal scanning laser ophthalmoscope (CSLO) [1] is a confocal imaging technique that can produce high-resolution retinal images by two-dimensional (2D) scanning illumination and filtration of stray light through a confocal arranged pinhole. However, the 2D scanning of the CSLO results in an imaging speed usually lower than 20 Hz, such a low imaging speed will cause serious intra-frame jitter and severely

blur the retinal images. Thus, an improved line-scanning ophthalmoscope (LSO) technique [2] based on line-beam illumination and probing imaging is used for retinal imaging, which can greatly improve imaging speed, typically above 100 Hz. The most direct solution to increase imaging speed is to reduce the exposure time and control the exposure available. As the exposure decreases, the signal-to-noise ratio (SNR) of the LSO images also decreases. Moreover, affected by speckle noise, intraocular scatter, and other factors, the quality of LSO images will be further degraded, and important fine morphological features will be further obscured, creating a challenge to any follow-up image analysis, such as retinal vessels segmentation and lesion

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

segmentation and recognition. An appropriate image post-processing approach is indispensable to compensate for the limitations of the LSO retinal imaging procedure.

Image super-resolution (SR) is an effective and widely used post-processing method for image-quality improvement that has been utilized for various medical imaging modalities [3]–[6]. Algorithms that tackle the super-resolution problem can be mainly classified as either reconstruction- or learning-based. Reconstruction-based methods can be divided into frequency- and spatial-domain methods. In frequency- domain methods, under the assumption of limited signal bandwidth in the original scene, using the properties of the Fourier transform, formulas for restoring a high-resolution (HR) image can be given [7]–[10]. This method's disadvantage is that it is limited to global motion, hence it only works for planar shifts and rotations [11], [12], so frequency-domain methods are ineffective with multi-frame image SR. In spatial-domain methods, an HR image with an improved SNR is reconstructed from multiple low-resolution (LR) frames by exploiting sub-pixel motion in an image sequence [13], [14]. Spatial-domain methods are commonly used in retinal image SR tasks. Murillo *et al.* [15] presented the first retinal image super-resolution method for scanning laser ophthalmoscopes (SLOs), which is effective in improving image quality. In Köhler's work [13], natural eye movements during an examination were used as cues to reconstruct HR retinal images from LR retinal video sequences. Spatial-domain methods are sensitive to the accuracy of image registration [16]. All of the above methods only super-resolve regions of interest (ROIs) or images with small fields of view (20° or 25°), making the registration of such images much easier. The LSO retinal image has a wide field of view (45°) and is rich in retinal vessels, so the accurate multi-frame registration of the entire LSO retinal image is a complicated and challenging task, and inaccurate registration will greatly affect the quality of image super-resolution. In addition, these methods have limited performance for handling large scaling factors and they incur great computational cost. In summary, reconstruction-based methods are unsuitable for LSO retinal-image super-resolution.

Deep learning [17] is a popular image super-resolution approach with promising results. This method uses a large number of LR and HR images in pairs as a priori information to estimate missing high-frequency details by learning the mapping relationship between LR and HR images. Generative adversarial networks (GANs) [18] have been demonstrated to be powerful models for image generation that can generate the details of images, and have been applied to numerous image super-resolution tasks [19]–[21]. In terms of retinal image super-resolution, Mahapatra et al. [21] presented a single-frame retinal image super-resolution method based on the GAN and local saliency maps. However, to the best of our knowledge, learning-based methods have not been investigated for multi-frame retinal image super-resolution. Similarly, work on learning-based methods for LSO image-quality improvement is still lacking in the current literature.

In this paper, we propose a GAN-based method for LSO retinal image sequence SR. Unlike reconstruction-based methods, ours does not need to register the image sequences, but directly learns an end-to-end mapping from LR image sequences to HR images. In this approach, we first train the initial multi-frame image super-resolution GAN (MSRGAN) model on an LSO retinal image sequence dataset, to let the network learn the mapping relationship from LR sequences to HR images. Next, we use the retinal vessel enhancement method [22] to process the target images (HR images), and the processed images are taken as the new target images (HR images) and put into the pretrained MSRGAN network to further improve the reconstruction performance of retinal vessels. The final model called the multi-frame image enhancement and super-resolution GAN (MESRGAN) model, because it can achieve multi-frame LSO retinal image super-resolution and retinal-vessel enhancement. The rest of this paper is organized as follows. In section II, we discuss the experimental setup, datasets, and proposed method. The results of our method, and its quantitative and qualitative evaluation, are presented in section III. In section IV, we summarize our work and discuss the experimental results.
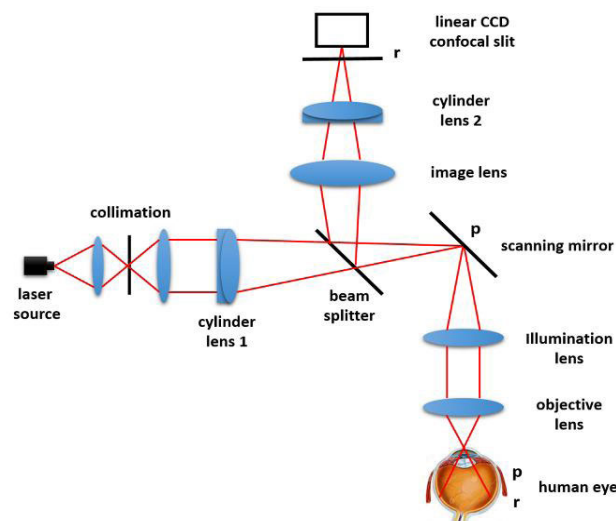
FIGURE 1. Structure diagram of LSO system (p: Pupil conjugate plane, r: Retina conjugate plane).

II. METHODS

A. EXPERIMENTAL SETUP

A schematic of the experimental setup is shown in Fig. 1. This LSO system [23], [24] mainly consists of an illumination path and an imaging light path. In the illumination part, the divergent beam of light emitted from the laser source ($\lambda = 641nm$) changes into parallel light after passing through a collimation lens, and then changes to linear light after cylinder lens 1. The linear light beam passing through the beam splitter will be reflected by the scanning mirror (6320H, Cambridge Corporation) and then focused by the illumination lens, after which it will enter the objective lens and reach the sample. In the imaging light path, the light reflected from the focusing plane

of the sample passes through the objective lens and collection lens, and then is reflected by the scanning mirror and image lens. After passing through cylinder lens 2, the information of the retina can be captured by charge-coupled device (CCD). The field of view of our imaging system is 45°, and the lateral resolution is approximately 10μm in the human retina.

B. DATASETS

Eight normal, healthy subjects (2 female, 6 male, ranging in age from 22 to 35 years) were recruited for this study. Video clips of human retinas at the spectral range of 641 nm were captured sequentially. The captured retinal image sequence datasets were divided into 1,344 subsets, of which 1,200 were used for training and 144 for testing. There were five images in each subset, and the resolution of each image was 512×512. We down-sampled the images by 4 to obtain images of resolution 128×128 as the input images (LR images) of the network, and chose the centering-frame images (512×512) as the target images (HR images). Training data were augmented with random horizontal flips and rotations following common data-augmentation methods [25]. Affected by Gaussian beam and limited exposure available, the images captured from LSO system usually suffer from heterogeneous illumination and low contrast. Therefore, all of the images in our dataset were preprocessed by using contrast-limited adaptive histogram equalization (CLAHE) [26] method, the preprocessed image as shown in Fig. 2(b). CLAHE is a histogram equalization technique, which chooses the clipping level of the histogram flexibility by computing the local histogram mapping function. We used Matlab's *adapthisteq* function to equalize histogram, and the clipping value used for all of the images was 0.02, a suitable value to keep good visual effect of the images simultaneously achieving the image contrast enhancement and heterogeneous illumination compensation.
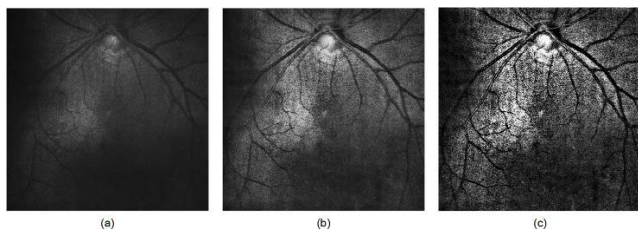
FIGURE 2. Preprocessed image results. (a) Original image. (b) CLAHE enhanced image. (c) Retinal-vessel enhanced image.

Retinal vessels are the most important features in retinal images. To let the network focus more on the learning of retinal vessels, we used a retinal-vessel enhancement method [22] to process the target images, and we made them the new target images [Fig. 2(c)] for retinal-vessel enhancement training.

C. GENERATIVE ADVERSARIAL NETWORKS

Multi-frame image SR estimates a super-resolved image I_{SR} from a sequence of LR input images I_{LR}^m. To achieve this,

we train a generator network as a feed-forward convolutional neural network G_{θ_G} with parameters $\theta_G = \{W_{1:L}; b_{1:L}\}$. Here, $\{W_{1:L}; b_{1:L}\}$ denotes the weights and biases of a deep convolutional neural network that are obtained by:

$$\hat{\theta} = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l_{SR}\left(G_{\theta_G}\left(I_{LR}^{mn}\right), I_{HR}^n\right) \tag{1}$$

where l_{SR} is the loss function, I_{LR}^{mn} is the LR image sequence, and I_{HR}^n is the HR image.

Inspired by GAN [18] and Wasserstein GAN (WGAN) [27], we further define a discriminator network D_{θ_D}. Unlike most training, which optimizes in an alternating manner along with G_{θ_G} to solve the adversarial min-max problem, we introduce the Wasserstein distance as an indicator of the training process. The Wasserstein distance has the desirable property of being continuous and differentiable almost everywhere under mild assumptions [28]. The Wasserstein distance is informally defined as the minimum cost of transforming the distribution of generated images to the distribution of target images. Therefore, using the Wasserstein distance $W(I_{SR}, I_{HR})$ to measure the difference between the super-resolved image I_{SR} and the HR image I_{HR} can prevent gradient vanishing and obtain better super-resolution results. The adversarial min-max problem based on the WGAN is defined as:

$$\min_{\theta_G} \max_{\theta_D \in L} \mathop{\mathrm{E}}_{I_{HR} \sim p_{\text{train}}(I_{HR})} [D(I_{HR})] - \mathop{\mathrm{E}}_{I_{LR} \sim p_G(I_{LZ})} \left[D\left(G_{\theta_G}\left(I_{LR}^m\right)\right)\right] \tag{2}$$

where L is the set of 1-Lipschitz functions.

The general idea behind formula (2) is that it allows one to train a generative model G with the goal of fooling a discriminator D that is trained to distinguish super-resolved images from HR images. With this approach, our generator can learn to reconstruct SR images that are highly similar to HR images and difficult to be distinguished by D.

The architectures of the proposed generator and discriminator networks are shown in Figs. 3 and 4, respectively. The core of our generator network is a series of residual blocks, each composed of two convolutional layers with kernel size 3×3. Each convolutional layer is followed by a rectified linear unit (RELU) activation function. As shown in Fig. 3, the first convolutional layer of the generator network extracts spatial information individually from each of the five frames in the input sequence. The second convolution takes a concatenation of the extracted spatial information from previous steps. Then, the residual blocks learn the transformation function that provides the final solution. The centering frame is the most important of all the input frames [29], so we directly add the input centering frame to the output details to ensure that the low-frequency details are better preserved in the training process.

To discriminate generated SR images from real HR images, we train a discriminator network proposed in [19], as illustrated in Fig. 4. The discriminator has 8 convolutional layers
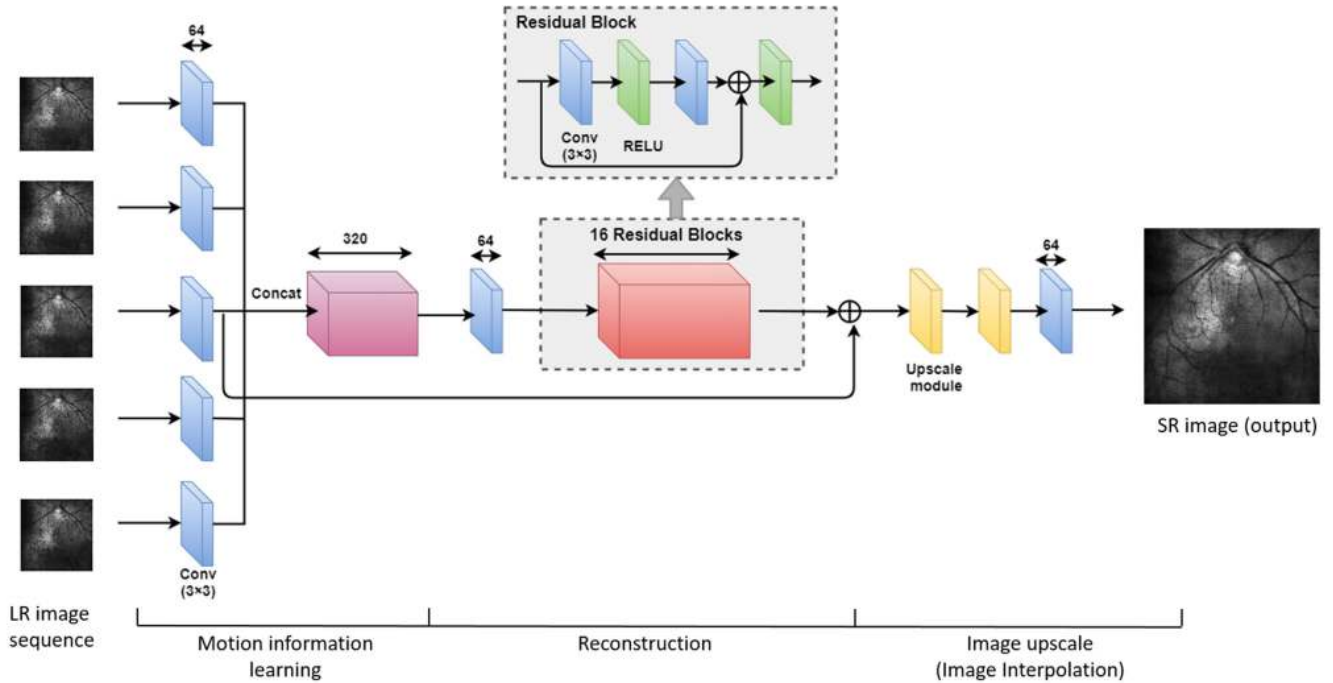
FIGURE 3. Proposed generator architecture. (Conv: Convolutional layer, Concat: Concatenate, RELU: Rectified linear unit activation layer).
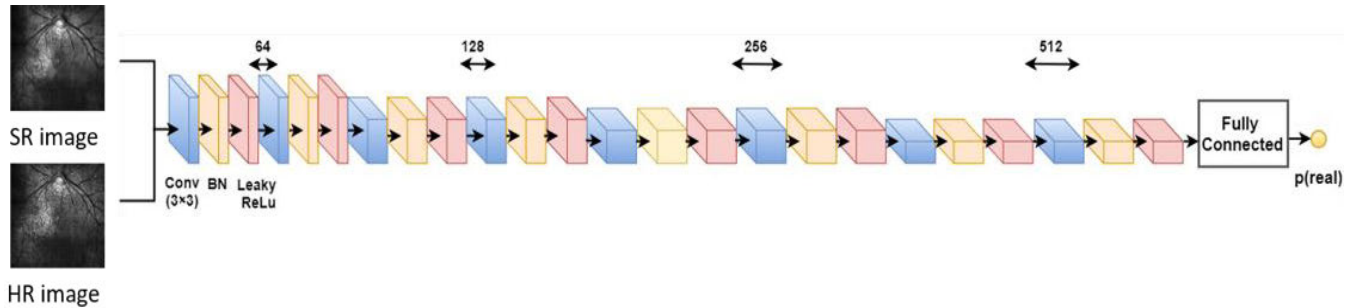

FIGURE 4. Proposed discriminator architecture. (Conv: Convolutional layer, Concat: Concatenate, RELU: Rectified linear unit activation layer).

with number of kernels increasing by a factor of 2 from 64 to 512. Leaky RELU is used, and strided convolutions reduce the image dimension when the number of features is doubled. A fully connected layer follows the result of 512 feature maps to obtain the final discriminate results.

D. LOSS FUNCTION

To ensure that the generated images resemble the target images, we require that the generator not only minimize the adversarial loss, but also the content loss, which measures the difference between the generated images and the target images. Hence, we defined the function l_{SR}, which is a combination of content loss and adversarial loss, and is given by:

$$l_{SR} = \alpha l_{SR}^{Cont} + (1 - \alpha)l_{SR}^{Adv} \qquad (3)$$

where $\alpha > 0$ is an experimentally determined hyperparameter that controls the contribution of each loss component.

Content loss is composed of feature-space loss and pixel-space loss.

Feature-space loss is based on differences between not pixels but high-level image feature representations extracted from pretrained convolutional neural networks [30]. The feature-space loss used for our model is based on the pretrained 19-layer convolutional neural network (VGG-net) [31], which can help reconstruct high quality images. The feature-space loss, which is also called VGG loss, is calculated as the Euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I_{LR}^m)$ and the target image I_{HR}:

$$l_{SR}^{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I_{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I_{LR}^m))_{x,y})^2$$

$$(4)$$

where $\phi_{i,j}$ is the feature map obtained by the *jth* convolution before the*ith* max pooling layer, and $W_{i,j}$ and $H_{i,j}$ are the dimensions of ϕ.

Pixel-space loss is used to ensure that the generated images do not depart greatly from the content in the corresponding target images. Here, the pixel-space loss is provided by the L1 loss instead of the traditional L2 loss, L1 loss is defined as:

$$l_{SR}^r = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| (I_{HR})_{x,y} - (G_{\theta_G}(I_{LR}^m))_{x,y} \right| \quad (5)$$

We found that our network performed better when trained with L1 loss rather than L2 loss, as discussed in section III and IV.

Adversarial loss based on the Wasserstein GAN is defined as:

$$l_{SR}^{Adv} = \sum_{n=1}^{N} -D_{\theta_D}(G_{\theta_G}(I_{LR}^m)) \quad (6)$$

Wasserstein GAN loss can measure the training process of the generator and discriminator. The smaller the Wasserstein GAN loss the better trained the model. It is defined as:

$$l_{SR}^W = \sum_{n=1}^{N} \left| D_{\theta_D}(G_{\theta_G}(I_{LR}^m) - D_{\theta_D}(I_{HR})) \right| \quad (7)$$

E. TRAINING FOR SUPER-RESOLUTION

Learning an end-to-end SR function G_{θ_G} requires the estimation of network parameters $\theta_G = \{W_{1:L}; b_{1:L}\}$, which is achieved by minimizing the loss function l_{SR} in formula (3) and the Wasserstein GAN loss in formula (7). We train the MSRGAN model for 200 epochs using the RMSprop [32] optimizer and a batch size of 16. The initial learning rate is set to 10^{-4}, and it is reduced by a factor of 0.1 after each 100 epochs. We train the MSRGAN model for SR scale factors of 2, 3, and 4. As illustrated in Fig. 2, all of the LSO images in our dataset used for training and testing are affected by speckle noise that is inherent in coherent imaging systems. The general approach for de-speckling is to take multi-frame LSO images and average the uncorrelated noise [33]. However, this can cause the loss of some intensity information. Therefore, images preprocessed by this method are not conducive to accurate image SR reconstruction. Surprisingly, the MSRGAN model cannot only reconstruct SR retinal images from LR retinal image sequences but can suppress the speckle noise of the LSO retinal images. The details of this finding will be discussed in section IV.

F. TRAINING FOR RETINAL-VESSEL ENHANCEMNET

The MSRGAN model can reduce the detail of retinal vessels while reducing the speckle noise of the image background. Therefore, super-resolved retinal images reconstructed from MSRGAN may blur the retinal vessels and even lose some of their details. We use a retinal-vessel enhancement method [22] to process the target HR images. Then we put the processed HR images and the original LR image sequences into the pretrained MSRGAN model and let the network re-learn the function of retinal-vessel enhancement and re-reconstruct the easily overlooked retinal vessel details. This joint training model, called MESRGAN, achieves multi-frame LSO retinal image super-resolution, speckle noise reduction, and retinal-vessel enhancement.

We train the MESRGAN model for 60 epochs, which we find to be a suitable number to achieve convergence. Similar to the training of the MSRGAN model, the batch size is 16, and we use the RMSprop optimizer to update each gradient.

III. EXPERIMENTS AND RESULTS

The training process took place offline, using the dataset and method described above, implemented in Python using the Tensorflow (Version 1.8.0) library on a desktop computer with an Intel Xeon E5-2620 CPU with 32 GB RAM, and two NVIDIA GeForce GTX 1080Ti GPUs. It took 14 hours to train the MSRGAN model and 6 hours to train the MESRGAN model.

A. MSRGAN MODEL TEST RESULTS

We tested the MSRGAN model on the LSO retinal image sequence dataset. First, we changed the loss function from L2 to L1 and trained the model separately on scales of 2, 3, and 4. The test was conducted on the 16 LSO retinal image sequences of our test set, with the criteria of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), calculated as follows:

$$MSE = \frac{\sum_{n=1}^{N} (x^n - y^n)}{N} \quad (8)$$

$$PSNR = 10 \times lg\left(\frac{255^2}{MSE}\right) \quad (9)$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (10)$$

where, *MSE* is the mean square error; N is the size of image; x^n and y^n are the *nth* pixel of original image x and processed image y; μ_x, μ_y are the average of x, y; σ_x^2, σ_y^2 are the variance of x, y; σ_{xy} is the covariance of x and y.

TABLE 1. Performance comparison with L1 loss and L2 loss on scale factors 2, 3, and 4.

	MSRGAN(L2 loss) PSNR(dB)/SSIM	MSRGAN(L1 loss) PSNR(dB)/SSIM	Difference $\triangle$PSNR(dB)/$\triangle$SSIM
×2	17.13/0.7652	17.65/0.7963	0.52/0.0311
×3	16.42/0.7428	16.85/0.7622	0.43/0.0194
×4	15.83/0.7262	16.15/0.7380	0.32/0.0118
	MSRGAN(L2 loss) PercepDist[34]	MSRGAN(L1 loss) PercepDist[34]	Difference $\triangle$PercepDist[34]
×2	0.0652	0.0608	0.0044
×3	0.1408	0.1356	0.0052
×4	0.1694	0.1649	0.0045

As described in Table 1, since greater magnification makes reconstruction more difficult, the results obtained with larger magnification are decreased to some extent. As shown in
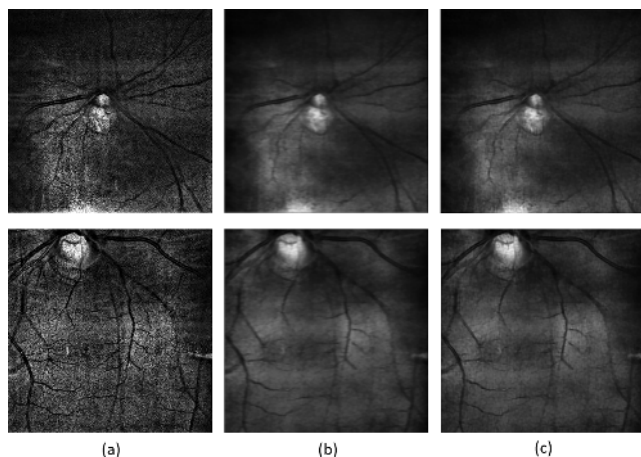
FIGURE 5. SR results comparison with L1 loss and L2 loss for scale factor of 4. (a) Original HR image. Model trained with: (b) L2 loss. (c) L1 loss.

the upper part of Table 1, the MSRGAN model trained with L1 loss clearly performs better across all scale factors than when trained with L2 loss. PSNR and SSIM of the model trained with L1 loss improved by 0.32 dB and 0.0118, respectively, at 4× magnification. In addition, the perceptual distance (PercepDist) [34] was used to measure the perceptual quality of images. Smaller perceptual distance metrics imply better perceptual quality and visual effects [20]. As illustrated in the lower part of Table 1, the perceptual distance of the model trained with L1 loss was 0.0045 lower than that trained with L2 loss, at 4× magnification.

Fig. 5 shows the SR visual results comparison of MSRGAN model when trained with L1 and L2 loss. As expected, the model trained with L1 loss gives sharper quality. This implies that appropriate selection of the loss function can improve image reconstruction.

B. MESRGAN MODEL TEST RESULTS

Having demonstrated that L1 loss outperforms L2 loss in reconstruction, we used L1 loss as the pixel-space loss to train the MESRGAN model on scales of 2, 3, and 4. We also tested the MESRGAN model on the 16 LSO retinal image sequences of the test set.

To demonstrate the usefulness of the MESRGAN model, we compared the results of the proposed model with several

state-of-the-art methods, included Bicubic, SRGAN [19], VDSR [35], VSRnet [36], and SPMC-SR [37]. Among them, Bicubic is one of the well-known classical SR method, SRGAN and VDSR are two state-of-the-art single image SR (SISR) models, VSRnet and SPMC-SR are two state-of-the-art video SR (VSR) models. We trained and tested all these models on our LSO sequence datasets, and the centering frame of sequences were used to train and test the SISR models. The models were evaluated quantitatively with the metrics of PSNR, SSIM, and PercepDist.

It is clear from Table 2 that MSRGAN outperformed other state-of-the-art methods for all scale factors. MSRGAN surpassed Bicubic, SRGAN, VDSR, VSRnet, and SPMC-SR model by 1.45dB, 0.88dB, 0.82dB, 0.46dB, 0.61dB in PSNR and 0.1132, 0.0259, 0.0218, 0.0103, 0.0142 in SSIM when the scale factor was 4. The perceptual distance of MSRGAN model was 0.0326, 0.0263, 0.0248, 0.0055, 0.0008 lower than those state-of-the-art methods, respectively, at 4× magnification. MESRGAN model is the favored model for producing high SNR and visually pleasing frames, as determined by PSNR and perceptual distance. MESRGAN surpassed MSRGAN model by 0.72 dB in PSNR and 0.0201 lower in perceptual distance at 4× magnification. This implies that joint SR and retinal-vessel enhancement training can greatly improve the performance of the network.

For qualitative comparison, Fig. 6 shows the images that were super-resolved by the Bicubic, SRGAN, VDSR, VSRnet, SPMC-SR, MSRGAN, and MESRGAN methods. By zooming in on the regions highlighted with yellow boxes, large differences can be observed. As expected, MSRGAN model can better suppress speckle noise than Bicubic, SRGAN, and VDSR methods. And the visual effect of super-resolved images obtained by MESRGAN had much sharper quality than those obtained by VSRnet, SPMC-SR and MSRGAN, especially in retinal vessels.

### C. RESULTS COMPARISON ON REAL LSO RETINAL IMAGE SEQUENCES

To evaluate the effectiveness of the MESRGAN model in image-quality improvement, it was applied to sets of real LSO images (512 × 512) that had not been subjected to resolution reduction. As illustrated in Fig. 7, compared to the images super-resolved by the state-of-the-art methods, the images super-resolved by MSRGAN look smoother, but their

TABLE 2. Performance comparison with state-of-the-art methods tested on scale factors 2, 3, and 4.

	Bicubic PSNR(dB)/SSIM	SRGAN[19] PSNR(dB)/SSIM	VDSR[35] PSNR(dB)/SSIM	VSRnet[36] PSNR(dB)/SSIM	SPMC-SR[37] PSNR(dB)/SSIM	MSRGAN(ours) PSNR(dB)/SSIM	**MESRGAN(ours)** PSNR(dB)/SSIM
×2	16.23/0.7355	16.86/0.7568	17.13/0.7688	17.45/0.7738	17.47/0.7800	17.65/0.7963	**18.36/0.8835**
×3	15.31/0.6842	16.12/0.7328	16.54/0.7407	16.76/0.7569	16.74/0.7426	16.85/0.7622	**17.76/0.8422**
×4	14.70/0.6248	15.27/0.7121	15.33/0.7162	15.69/0.7277	15.54/0.7238	16.15/0.7380	**16.87/0.8098**
	Bicubic PercepDist[34]	SRGAN[19] PercepDist[34]	VDSR[35] PercepDist[34]	VSRnet[36] PercepDist[34]	SPCM-SR[37] PercepDist[34]	MSRGAN(ours) PercepDist[34]	**MESRGAN(ours)** PercepDist[34]
×2	0.0899	0.0853	0.0876	0.0735	0.0694	0.0608	**0.0456**
×3	0.1608	0.1586	0.1593	0.1438	0.1397	0.1356	**0.1109**
×4	0.1975	0.1912	0.1897	0.1704	0.1657	0.1649	**0.1468**
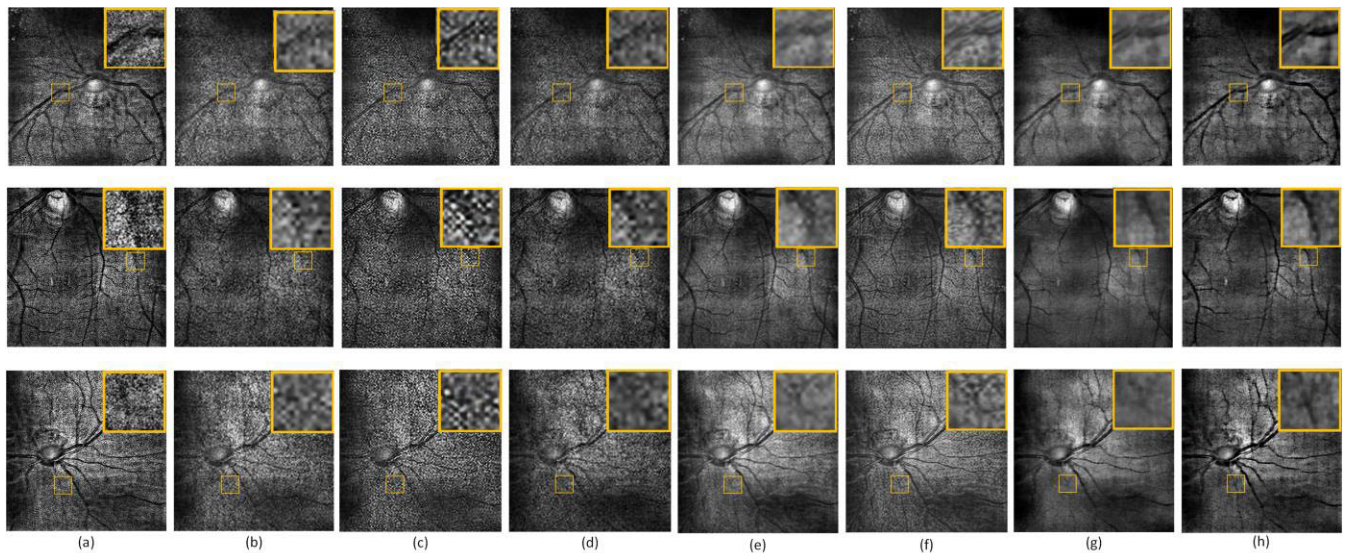
FIGURE 6. SR results comparison for scale factor 4. (a) Original HR images. SR images from: (b) Bicubic method. (c) SRGAN model [19]. (d) VDSR model [35]. (e) VSRnet [36]. (f) SPMC-SR model [37]. (g) MSRGAN (ours). (h) MESRGAN (ours).
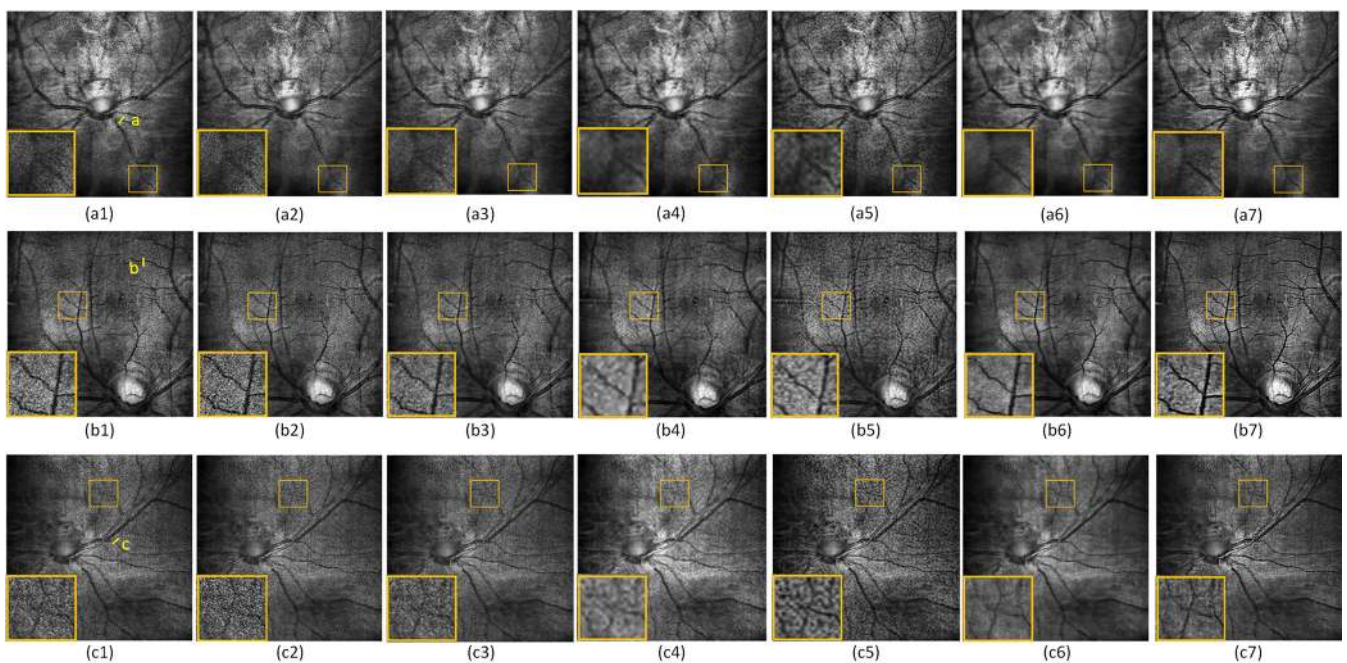

FIGURE 7. SR results for scale factor 4. (a1, b1, c1) bicubic images. SR images from: (a2, b2, c2) SRGAN model [19]. (a3, b3, c3) VDSR model [35]. (a4, b4, c4) VSRnet [36]. (a5, b5, c5) SPMC-SR model [37]. (a6, b6, c6) MSRGAN model (ours). (a7, b7, c7) MESRGAN model (ours).

retinal vessels are also smoothed, which even results in the loss of some small vessel details. Training the MESRGAN model makes the model focus more on the reconstruction of retinal vessels. As seen in Figs. 7(a7), 7(b7), and 7(c7), the background of the super-resolved images was smoothed and the retinal vessels were enhanced.

In the case of real LSO images, there was no ground-truth image for evaluation of the results and a No-Reference Image Quality Assessment (NR-IQA) algorithm was needed. In our work the power spectra and full width at half maximum (FWHM) were used to evaluate the methods' performance.

The power spectra describe the amplitude of the spectral power in the images that is distributed across the spatial frequency. The power spectra of images in Fig. 7 are present in Fig. 8. Figs. 8(a)-(c) show the respective power spectra of Figs. 7(a1)-7(a7), 7(b1)-7(b7), and 7(c1)-7(c7). It can be seen that the high frequencies probably correspond to noise. The power spectra of Figs. 7(a4), 7(b4), 7(c4), 7(a6), 7(b6), 7(c6), 7(a7), 7(b7), and 7(c7) are rapidly decreased, the most
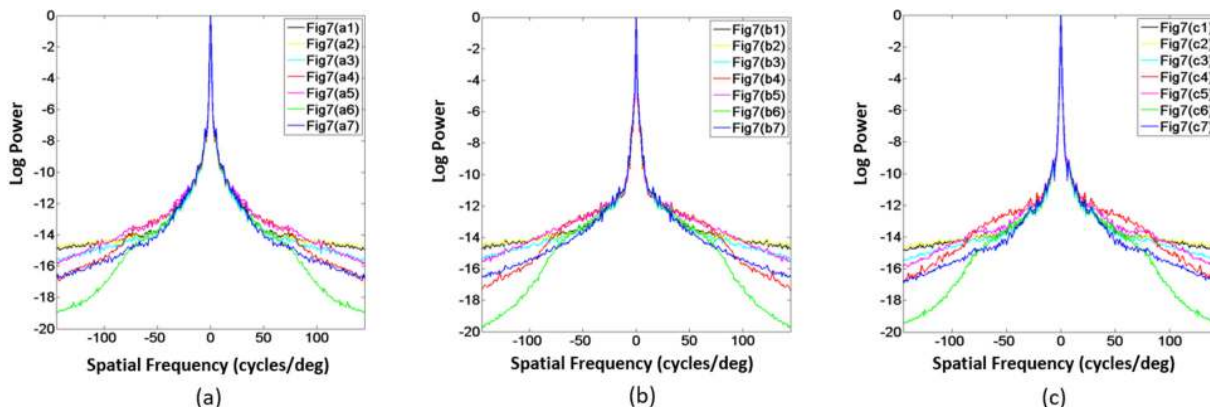
FIGURE 8. Average power spectra of images in Fig. 7. Power spectra curves of (a) Figs. 7(a1)–7(a7), (b) of Figs. 7 (b1)–7(b7), and (c) of Figs. 7(c1)–7(c7).
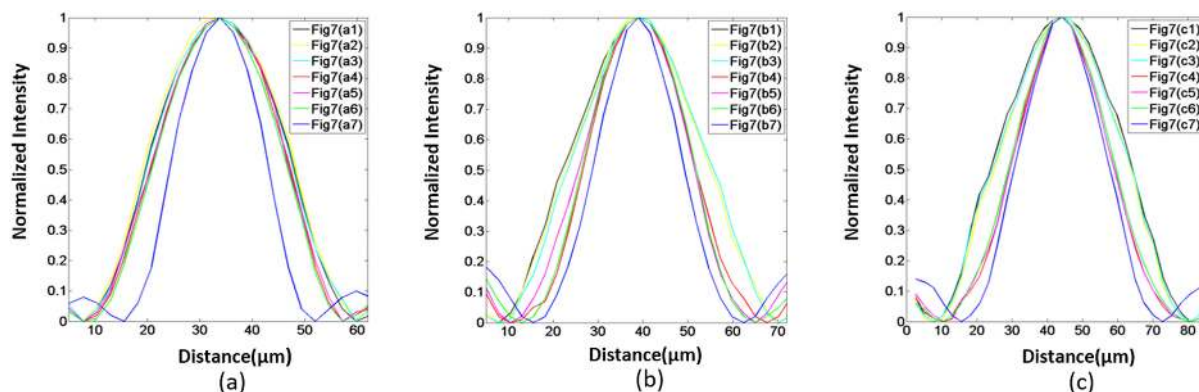
FIGURE 9. Normalized intensity statistics curves of images in Fig.7. Statistics curves of (a) Figs. 7(a1)–7(a7) at the position of line a, (b) Figs. 7(b1)–7(b7) at the position of line b, and (c) Figs. 7(c1)–7(c7) at the position of line c.

TABLE 3. Resolution assessment of reconstructed real LSO retinal images in Fig.7.

	Bicubic Fig.7(a1)	SRGAN Fig.7(a2)	VDSR Fig.7(a3)	VSRnet Fig.7(a4)	SPMC-SR Fig.7(a5)	MSRGAN Fig.7(a6)	**MESRGAN Fig.7(a7)**
FWHM(μm)	35	34	32	31	31	30	**20**
	Fig.7(b1)	Fig.7(b2)	Fig.7(b3)	Fig.7(b4)	Fig.7(b5)	Fig.7(b6)	**Fig.7(b7)**
FWHM(μm)	34	34	33	27	28	26	**22**
	Fig.7(c1)	Fig.7(c2)	Fig.7(c3)	Fig.7(c4)	Fig.7(c5)	Fig.7(c6)	**Fig.7(c7)**
FWHM(μm)	41	40	39	31	31	31	**27**

is Figs. 7(a6), 7(b6), and 7(c6), then Figs. 7(a7), 7(b7), and 7(c7). This indicates that the noise of the image which super-resolved by the proposed method are effectively suppressed and the image quality is improved. Since the high frequencies also correspond to high-frequency details, the power spectra of images reconstructed by the MESRGAN model (blue curves) decrease more slowly than those reconstructed by the MSRGAN model (green curves). This implies that some retinal-vessel details are enhanced by the MESRGAN model.

To demonstrate the validity of the proposed method in the spatial resolution improvement of the reconstructed LSO images, FWHM was used as the measurement metric, since a higher resolution should result in a narrower FWHM. Here,

the spatial resolution was analyzed for the specified line of the LSO images, shown with the yellow line in Figs. 7(a1), 7(b1), and 7(c1). The location of each line is considered such that it across a fine observable detail, and then the pixel intensity of each pixel on this line can be counted. Fig. 9(a) shows the normalized intensity statistics curves for the case in Figs. 7(a1)-7(a7) at the position of line a, Fig. 9(b) shows them for the case in Figs. 7(b1)-7(b7) at the position of line b, and Fig. 9 (c) shows them for the case in Figs. 7(c1)-7(c7) at the position of line c. As the plots and Table 3 illustrate, the MESRGAN model provides a narrower FWHM than the state-of-the-art methods or MSRGAN model. These results demonstrate the ability of the proposed method in increasing the spatial resolution of reconstructed LSO retinal images.
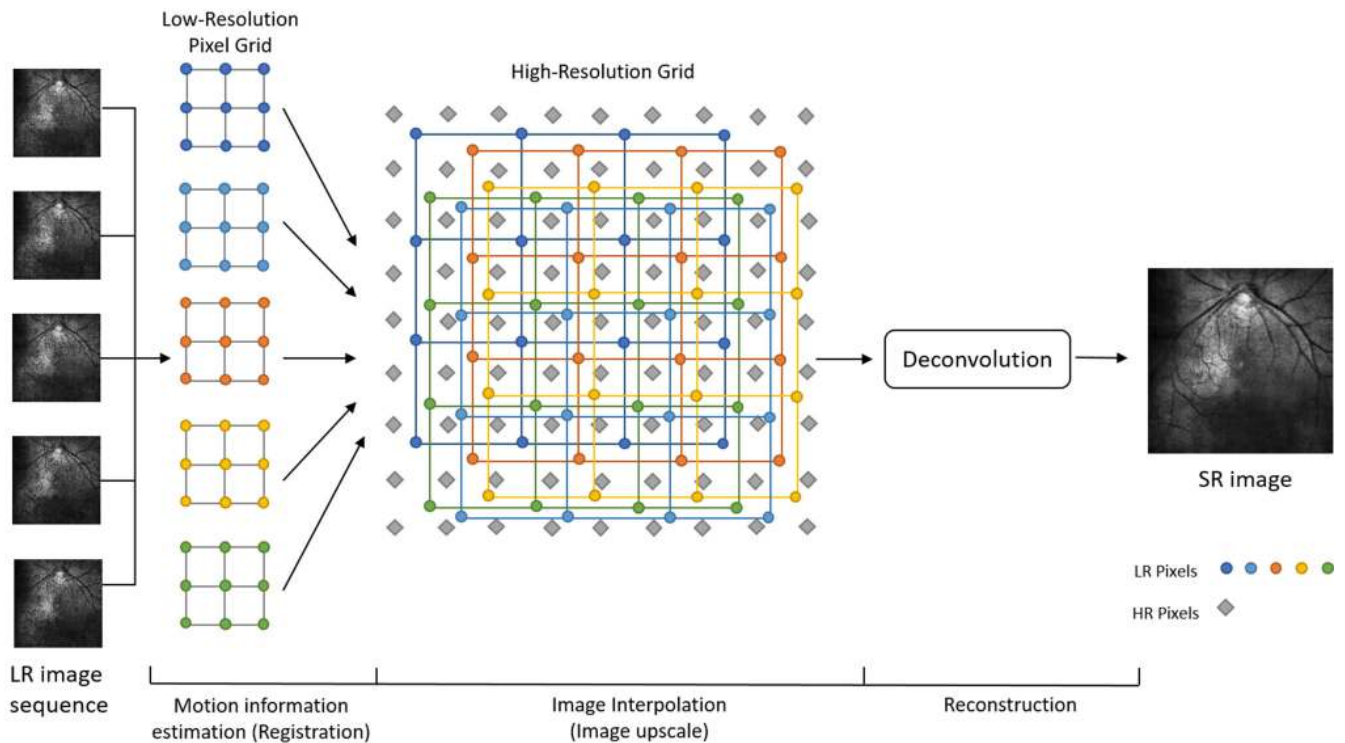
FIGURE 10. SR reconstruction architecture of reconstruction-based method (spatial-domain).

IV. CONCLUSIONS AND DISCUSSIONS

In the work reported in this paper, the GAN was used to extract spatial information from each input LR retinal frame and then learn the mapping relation between LR retinal image sequences and HR retinal images. In training the MSRGAN model, we introduced pixel-space loss (L1 loss) and feature-space loss (VGG loss) to help produce super-resolution images of high perceptual quality. We found that this model cannot only super-resolve multi-frame LSO retinal images but can suppress their speckle noise. To further improve the retinal-vessel reconstruction performance of the MSRGAN model, we proposed an MESRGAN model, which is a joint super-resolution and retinal-vessel enhancement network. The proposed method was validated on down-sampled LSO image sequences as well as real LSO image sequences. The results demonstrate that the proposed method can significantly enhance the SNR of LSO images and efficiently improve the resolution of LSO retinal images, which has great practical significance for clinical diagnosis and analysis.

It should be noted that there are also several points need to be discussed in experimental results.

1) Table 1 and Fig. 6 demonstrate that the model trained with L1 loss outperforms L2 loss. The possible reason is that using an L2 loss produces blurry images [38]–[40]. If the probability distribution for an output pixel has two equally likely modes, the L2 loss over the data will be minimized by the averaged value of these two modes, even though the averaged value has very low probability. In the case of an

L1 loss, this effect decreases, as the output value is the median of the set of equally likely values [38]. Therefore, the model trained with L1 loss has higher PSNR and SSIM, lower PercepDist.

2) In this study, all of the LSO image sequences in our dataset used for training and testing were preprocessed by CLAHE or retinal-vessel enhancement method. Since the images captured by LSO were affected by speckle noise, the preprocessed method can enhance the noise while enhancing the images. Figs. 6(c) and 6(d) are the results of single image SR methods, speckle noise in these images are not suppressed, even more serious. However, the speckle noise in the images that super-resolved by multi-frame SR methods [Figs. 6(e)-6(h)] are suppressed to some extent. In summary, multi-frame SR methods contribute to speckle noise reduction. The possible explanation for the multi-frame learning-based SR method's good performance in de-speckling is that the learning-based method is similar to the reconstruction-based method (spatial-domain). As shown in Figs. 3 and 10, these two kinds of methods can both be said to comprise three key steps: motion-information estimation, image interpolation, and image reconstruction. Image interpolation is a critical stage in transforming LR images into a desired HR grid. In this stage, since the motion transformation between LR images is arbitrary, the registered image always matches to a non-uniformly-spaced HR grid. Thus, non-uniform interpolation is used to obtain uniformly spaced HR images. During this processing, LR images will be fused, and then the uncorrelated speckle noise will be averaged.
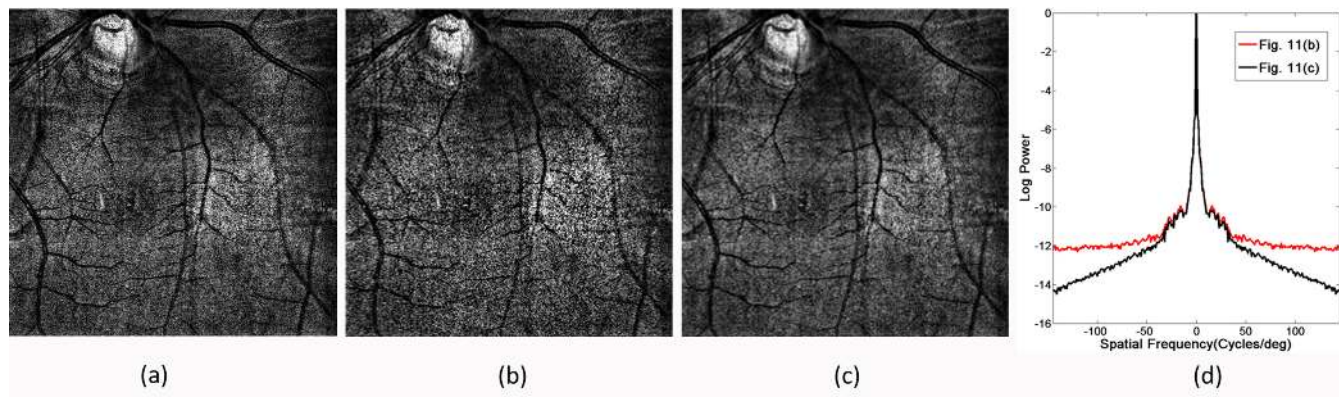
FIGURE 11. Comparison of downsampled images. (a) Original image (512 × 512). (b) Downsampled by "decimation" method with factor of 2. (c) Downsampled by "bicubic" method with factor of 2. (d) Average power spectra of (b) and (c).

3) Floating-point Operations (FLOPs) were used to measure the time complexity of the proposed model. The formulas of FLOPs are shown as follows:

$$\text{Time} \sim \text{O}\left(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{lin} \cdot C_{lout}\right) \quad (11)$$

$$M = \frac{(N - K + 2 \times P)}{S} + 1 \quad (12)$$

In formula (11), D is the number of convolutional layers; M_l is the output feature map of the *lth* layer; K_l is the convolution kernel size of the *lth* layer; C_{lin} is the input channel of the *lth* layer; C_{lout} is the output channel of the *lth* layer. In formula (12), input image size is N; K is the convolution kernel size; Padding is expressed as P, which equals to zero if padding is "valid", otherwise M directly equals to N when padding is "same". S represents the step length of filter in both vertical and horizontal directions in the original image.

The architecture of the proposed model are shown in Figs. 3 and 4, the padding is "same" in this network. After calculating, the FLOPs of the proposed generator and discriminator is $O(3.9 \times 10^6 N^2)$ and $O(7.28 \times 10^7 N^2)$, and the total time complexity of the proposed model is $O(7.67 \times 10^7 N^2)$, where N is the size of input image. Since the MSRGAN and MESRGAN model have same architecture, the total time complexity of them are the same. Note that the training of our models were taken place offline, so the time complexity of the models are acceptable.

4) Although the bicubic method is commonly used in image or video SR tasks to acquired LR images, we used decimation [41] method to down-sample the LSO image sequences. As shown in Fig. 11, we down-sampled an image using decimation and bicubic method by a factor of 2, respectively. As can be seen from Figs. 11(b) and (c), bicubic method is able to smooth the image, while decimation method can reserve even enhance the noise of image. Since the size of down-sampled images and original image are not same, we evaluated down-sampled images using power spectra and Blind Image Quality Index (BIQI) [42] as the NR-IQA algorithms to quantitatively compare the influence of these two

methods on images. Fig. 11(d) describes the power spectra of Figs. 11(b) and (c), and the high-frequencies power of Fig. 11(b) (red curve) is higher than that of Fig. 11(c) (black curve). As high-frequencies probably represent noise, it implies that the image down-sampled by decimation has more noise.

BIQI is a widely-used method for quantitatively evaluating the quality of No-reference images. Given an image, a quality score between 0 and 100 can be obtained using this method. Higher BIQI score indicates poorer image quality. Here, 20 original LSO images were down-sampled by decimation and bicubic method with scale factor of 2, respectively. The average BIQI of the decimation down-sampled images is 60.15 and that of bicubic down-sampled images is 48.75.

In summary, compared with bicubic method, the down-sampling method used in our study introduces more noise and intensity loss. Super-resolving such low-quality and noisy images is a challenging task. However, compared with other state-of-the-art methods [19], [35]–[37], the method proposed in this study performs better, which demonstrate the usefulness of the proposed method. In addition, it is more practical to use the decimation method, for the reason that the images captured from LSO system are inevitably affected by speckle noise. And using the images down-sampled by the decimation method as the input can enhance the noise suppression ability of the network to some extent.

REFERENCES

[1] R. H. Webb, G. W. Hughes, and F. C. Delori, "Confocal scanning laser ophthalmoscope," *Appl. Opt.*, vol. 26, no. 8, pp. 1492–1499, 1987.

[2] D. X. Hammer, R. D. Ferguson, T. E. Ustun, C. E. Bigelow, N. V. Iftimia, and R. H. Webb, "Line-scanning laser ophthalmoscope," *Proc. SPIE*, vol. 11, no. 4, 2006, Art. no. 041126.

[3] D. Kouame and M. Ploquin, "Super-resolution in medical imaging: An illustrative approach through ultrasound," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro (ISBI)*, Jun./Jul. 2009, pp. 249–252.

[4] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, 2008.

[5] M. Li, S. Shen, W. Gao, W. Hsu, and J. Cong, "Computed tomography image enhancement using 3D convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 291–299.

[6] S. Zhang, G. Liang, S. Pan, and L. Zheng, "A fast medical image super resolution method based on deep learning network," *IEEE Access*, vol. 7, pp. 12319–12327, 2018.

[7] F. Champagnat and G. Le Besnerais, "A Fourier interpretation of super-resolution techniques," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2005, pp. 4–7.

[8] D. Wei, "Image super-resolution reconstruction using the high-order derivative interpolation associated with fractional filter functions," *IET Signal Process.*, vol. 10, no. 9, pp. 1052–1061, Dec. 2016.

[9] D. Wei and Y.-M. Li, "Generalized sampling expansions with multiple sampling rates for lowpass and bandpass signals in the fractional Fourier transform domain," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4861–4874, Sep. 2016.

[10] D. Wei and Y.-M. Li, "Reconstruction of multidimensional bandlimited signals from multichannel samples in linear canonical transform domain," *IET Signal Process.*, vol. 8, no. 6, pp. 647–657, Aug. 2014.

[11] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP J. Adv. Signal Process.*, vol. 2006, pp. 1–14, Feb. 2006.

[12] D. Thapa, K. Raahemifar, W. R. Bobier, and V. Lakshminarayanan, "Comparison of super-resolution algorithms applied to retinal images," *Proc. SPIE*, vol. 19, no. 5, 2014, Art. no. 056002.

[13] T. Köhler, A. Brost, K. Mogalle, Q. Zhang, C. Köhler, G. Michelson, J. Hornegger, and R. P. Tornow, "Multi-frame super-resolution with quality self-assessment for retinal fundus videos," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2014, pp. 650–657.

[14] S. Chaudhuri, "Super-resolution via image warping," in *Super-Resolution Imaging*, vol. 632. New York, NY, USA: Springer, 2001, pp. 142–148.

[15] S. Murillo, S. Echegaray, G. Zamora, P. Soliz, and W. Bauman, "Quantitative and qualitative image quality analysis of super resolution images from a low cost scanning laser ophthalmoscope," *Proc. SPIE*, vol. 7962, pp. 298–306, Mar. 2011.

[16] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[17] K. Hayat, "Super-resolution via deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–33.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, vol. 2, Apr. 2017, no. 3, pp. 105–114.

[20] A. Lucas, S. López-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 51–55.

[21] D. Mahapatra, B. Bozorgtabar, S. Hewavitharanage, and R. Garnavi, "Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 382–390.

[22] P. Dai, H. Sheng, J. Zhang, L. Li, J. Wu, and M. Fan, "Retinal fundus image enhancement using the normalized convolution and noise removing," *Int. J. Biomed. Imag.*, vol. 2016, Aug. 2016, Art. no. 5075612.

[23] Y. He, Z. Wang, L. Wei, G. Shi, and Y. Zhang, "Compact, high-speed line scanning quasi-confocal ophthalmoscope and retina imaging experiments," *Biomed. Opt.*, vol. 9230, pp. 1–5, 2014.

[24] Y. He, H. Li, J. Lu, G. Shi, and Y. Zhang, "Retina imaging by using compact line scanning quasi-confocal ophthalmoscope," *Chin. Opt. Lett.*, vol. 11, no. 2, pp. 021101–021103, 2013.

[25] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *Proc. AMIA Annu. Symp.*, vol. 2017, 2017, pp. 979–984.

[26] M. Zhou, K. Jin, S. Wang, J. Ye, and D. Qian, "Color retinal image enhancement based on luminosity and contrast adjustment," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 3, pp. 521–527, Mar. 2018.

[27] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–32.

[28] Z. Chen and Y. Tong, "Face super-resolution through Wasserstein GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–8.

[29] D. Li and Z. Wang, "Video superresolution via motion compensation and deep residual learning," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 749–762, Dec. 2017.

[30] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[31] S. Liu and W. Deng, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2014, pp. 730–734.

[32] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[33] G. Gong, H. Zhang, and M. Yao, "Speckle noise reduction algorithm with total variation regularization in optical coherence tomography," *Opt. Express*, vol. 23, no. 19, pp. 24699–24712, 2015.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "Theu nreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[35] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[36] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[37] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4482–4490.

[38] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–14.

[39] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[40] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[41] D. D. Muresan and T. W. Parks, "Prediction of image detail," in *Proc. Int. Conf. Image Process.*, Sep. 2000, pp. 323–326.

[42] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

WANYUE LI was born in Xinjiang, China, in 1993. She received the B.S. degree from Xidian University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree with the University of Science and Technology of China, to do her research work with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. Her research interests include biomedical optical imaging and image processing techniques. Furthermore, she is specializing in the application of deep learning in medical image and optical image reconstruction.

YI HE was born in Sichuan, China, in 1984. He received the B.S. degree in automation from the University of Science and Technology of China, in 2008, and the Ph.D. degree in measurement technology and instrument from the University of Chinese Academy of Sciences, in 2013. From 2013 to 2015, he was a Research Assistant with the Key Laboratory on Adaptive Optics, Chinese Academy of Sciences, Chengdu, China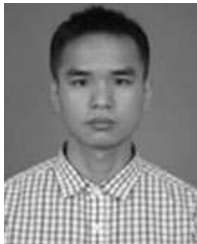. Since 2016, he has been an Associate Professor with the Chinese Academy of Sciences. He has authored of more than 40 articles, and more than 40 inventions. His research interests include biophotonics, biomedical imaging, confocal scanning imaging, optical coherence tomography imaging, and optical wavefront engineering for human eyes high resolution imaging.

WEN KONG was born in Tengzhou, Shandong, China, in 1992. He received the B.S. degree from Nanchang University, and the M.S. degree from China Jiliang University. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China, to do his research work with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. The major of his researching is optical imaging, and he is especially interested in the line scanning imaging method and adaptive optical imaging method for human or model animals' fundus.

JING WANG was born in Kunming, Yunnan, China, in 1995. She received the B.S. degree in biomedical engineering from Beijing Jiaotong University, Beijing, in 2017. She is currently pursuing the master's degree in biomedical engineering with the University of Science and Technology of China, Hefei. She is currently a Researcher with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. Her research interests include image processing and machine learning-based retinal disease automatic diagnosis and treatment systems.

GUOHUA SHI was born in Zhejiang, China, in 1981. He received the B.S. degree from the Department of Optoelectronic Information Engineering, Zhejiang University, Hangzhou, China, in 2003, and the Ph.D. degree in optical engineering from the Institute of Optics and Electronics, Chinese Academy of Sciences. He is currently a Professor and a Doctoral Advisor with the University of Science and Technology of China, Hefei, and the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai. He is also the Director of the Medical Optics Technology Laboratory and the Vice Director of the Key Medical Optics Laboratory of Jiangsu Province, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. His research interests include the *in vivo* optical imaging and detecting method, as well as developing and industrializing the corresponding optical medical instruments.

FENG GAO was born in Jiangxi, China, in 1991. He received the B.S. and M.S. degrees from the College of Optoelectronic Information, University of Electronic Science and Technology of China, in 2013 and 2016, respectively. From 2016 to 2018, he was a Research Trainee with the Key Medical Optics Laboratory of Jiangsu Province, Suzhou Institute of Biomedical Engineering and Technology. Since 2019, he has been a Research Assistant. His research interests include optical coherence tomography imaging and confocal ophthalmoscope research.

● ● ●