



OPEN Retinal fundus image super-resolution based on generative adversarial network guided with vascular structure prior

Yanfei Jia^{1✉}, Guangda Chen¹ & Haotian Chi²

Many ophthalmic and systemic diseases can be screened by analyzing retinal fundus images. The clarity and resolution of retinal fundus images directly determine the effectiveness of clinical diagnosis. Deep learning methods based on generative adversarial networks are used in various research fields due to their powerful generative capabilities, especially image super-resolution. Although Real-ESRGAN is a recently proposed method that excels in processing real-world degraded images, it suffers from structural distortions when super-resolving retinal fundus images are rich in structural information. To address this shortcoming, we first process the input image using a pre-trained U-Net model to obtain a structural segmentation map of the retinal vessels and use the segmentation map as the structural prior. The spatial feature transform layer is then used to better integrate the structural prior into the generation process of the generator. In addition, we introduce channel and spatial attention modules into the skip connections of the discriminator to emphasize meaningful features and accordingly enhance the discriminative power of the discriminator. Based on the original loss functions, we introduce the L1 loss function to measure the pixel-level differences between the segmentation maps of retinal vascular structures in the high-resolution images and the super-resolution images to further constrain the super-resolution images. Simulation results on retinal image datasets show that our improved algorithm results have a better visual performance by suppressing structural distortions in the super-resolution images.

The retina is a thin, transparent, light-sensitive tissue located at the back of the eye, and it is the only organ in the body that can be viewed non-invasively using visible light. Doctors can analyze retinal fundus images to screen for and diagnose ophthalmic and systemic diseases¹. The fundus imaging system or retinal camera is a sophisticated low-power microscope². In telemedicine, medicine images such as retinal fundus are sometimes degraded during upload, transmission, and download³. Besides, most retinal diseases have no clear symptoms in the early stages. As a result, some retinal fundus images do not have the desired clarity to support the ophthalmologist in making an accurate diagnosis. Therefore, it is necessary to improve the quality of the retinal fundus image. Image super-resolution techniques have been introduced to enhance the resolution and clarity of retinal fundus images. Image super-resolution aims to reconstruct the corresponding high-resolution image from a low-resolution image⁴. Image super-resolution also plays an important role in video surveillance⁵, remote sensing technology⁶, and medical image process⁷. With the development of artificial intelligence, deep learning neural network techniques are widely used in image super-resolution⁸. The generative adversarial networks based on deep learning neural networks were proposed as a powerful framework to generate super-resolution (SR) images with better visual perception⁹. Real-ESRGAN¹⁰ is a recently proposed super-resolution method based on generative adversarial networks. It uses a high-order degradation model to synthesize data for training. Compared to previous super-resolution methods based on generative adversarial networks, Real-ESRGAN generates images with better visual perception when dealing with low-resolution (LR) images of general real-world degradation. However, the Real-ESRGAN method suffers from structural distortions when super-resolving images¹¹. The retinal fundus images contain rich structural information and are therefore more prone to structural distortion. The structural distortions refer to the distortion of blood vessel shapes, as well as the misalignment or distortion of details on the generated super-resolution images. To solve the problem, we propose a retinal fundus image super-resolution based on the generative adversarial network guided with vascular structure priors.

¹School of Electrical and Information Engineering, Beihua University, Jilin 132013, China. ²College of electronic science and engineering, Jilin University, Changchun 130015, China. ✉email: jia_yanfei@163.com

In this paper, our main contributions are summarized as follows:

- 1 The retinal vascular structure information is essential for diagnosing retinal diseases, and the structure is distributed regularly on fundus images. Therefore, we first segment the input retinal fundus image using a pre-trained U-Net model to obtain its retinal vascular structure segmentation map. Secondly, we design a condition network and use the segmentation map as its input to introduce the structure to the generator. It can suppress the drawback of structural distortions of the SR images.
- 2 we propose to use the L1 loss function to measure the pixel-level differences between the segmentation map of the retinal vascular structure of the HR and SR images. It increases a quadratic constraint on the SR image to improve the quality of generated SR images.
- 3 we introduce channel and spatial attention modules into the skip connections of the Real-ESRGAN discriminator to make the discriminator pay more attention to important features. It can improve the discriminative power of the discriminator. Based on the above improvements, it can reduce the structural distortions in SR images and improve the quality of super-resolution of retinal fundus images based on Real-ESRGAN.

Related work

In this section, we briefly review traditional super-resolution methods for retinal images and focus primarily on super-resolution techniques based on deep learning. For the traditional super-resolution methods, Thapa et al.¹² reviewed several super-resolution methods for retinal images, including interpolation methods, frequency domain methods, regularization methods, and learning-based methods. They also compared the performance of these approaches in the context of retinal imaging. Jebadurai et al.¹³ proposed a learning-based single-image super-resolution algorithm that leverages the advantages of support vector regression and probability theory to minimize reconstruction errors. Their approach not only learns and develops a nonlinear functional mapping between low and high-resolution retinal images but also achieves minimal super-resolution reconstruction errors. Thomas et al.¹⁴ introduced a super-resolution method for fundus video images by exploiting natural eye movements during the examination. They used affine registration to reconstruct a motion-compensated super-resolution image from lower-resolution video data. Additionally, Thomas et al.¹⁵ developed a fully automatic framework to reconstruct high-resolution retinal images with a wide field of view from low-resolution video data, effectively integrating complementary regions of the retina. Furthermore, Retinex theory¹⁶ and dictionary learning¹⁷ have also been applied to enhance the quality of fundus images.

Although traditional super-resolution reconstruction algorithms can improve the quality of low-resolution fundus images, the results often fall short of expectations due to limitations in the number of available fundus images and the technology employed. Consequently, researchers have turned to deep learning to design network models that achieve better super-resolution reconstruction of fundus images. Dong et al.¹⁸ first proposed a super-resolution convolutional neural network algorithm based on deep learning. It uses only a three-layer convolutional neural network to learn the mapping relationship between the low-resolution (LR) images and the high-resolution (HR) images in an end-to-end manner. This three-layer convolutional neural network is too shallow and makes it difficult to extract deep features¹⁹. To address this problem, Kim et al. later proposed two super-resolution methods: the deep recurrent convolutional neural network (DRCN)²⁰ and the very deep super-resolution network (VDSR)²¹. The VDSR method has achieved excellent performance by introducing residual networks to train deeper network architectures. The DRCN method deepens the network with recursive structures to extract feature information of more layers while using recursive supervision and skip connections to alleviate the problem of gradient disappearance or gradient explosion due to network deepening. However, these methods only focus on high PSNR values and neglect the visual perceptual quality of the super-resolution (SR) images, resulting in blurred images with smoothed edges and textures.

Generative adversarial networks have been introduced into super-resolution methods to generate more realistic images. This method can improve the visual perceptual quality of SR images through adversarial learning. Generative adversarial networks were first proposed by Ian Goodfellow et al.²². Generative adversarial networks consist structurally of a generator and a discriminator. The generator takes an LR image as an input and generates an SR image as an output. The discriminator is to determine whether its input image is the high-resolution image or the SR image generated from the generator. Ledig et al.²³ first proposed a super-resolution approach based on generative adversarial networks (SRGAN). Unlike previous SR works, it introduces a perceptual loss using high-level feature maps of the pre-trained VGG network combined with a discriminator that makes SR images perceptually hard to distinguish from HR images. However, the SRGAN method also introduces artifacts when recovering more detail. To address this issue, ESRGAN²⁴ applies a Residual-in-Residual Dense Block (RRDB) without batch normalization as a basic module in the generator. This basic module deepens the network to improve performance and avoid the artifacts introduced by batch normalization. The ESRGAN method also replaces the original discriminator with a relativistic discriminator²⁵ that tries to predict the probability that a real image is relatively more realistic than a fake one to improve the discriminative power. RankSRGAN²⁶ addresses the problem of being unable to reasonably assess the quality of the SR images in previous methods by using perceptual metrics to assess the perceptual quality of the SR images instead of just PSNR and SSIM metrics. It also proposes a super-resolution generative adversarial network with Ranker²⁷. All of these approaches assume that the image degradation process is an ideal bicubic downsampling process, but this is not the same as real degradation. This degradation mismatch tends to make these above methods unsatisfactory when dealing with real-world degraded LR images.

Dwarikanath et al.²⁸ proposed progressive generative adversarial networks (P-GANs) to generate a high-resolution retinal fundus image from a low-resolution retinal fundus image. Prajapati et al.²⁹ proposed an unsupervised SR method without explicitly estimating the degradation of LR images using direct mapping of LR to SR in an attempt to mitigate the above limitation. Recently, Wang et al.¹⁰ proposed the Real-ESRGAN method

to recover better real-world degraded LR images based on the ESRGAN method combined with a higher-order degradation model. It proposes to use a high-order degradation model to better model the degradation process in the real world. In addition, it accordingly uses a U-shaped discriminator³⁰ with spectral normalization to enhance the training stability. However, while Real-ESRGAN performs well in super-resolution processing of real-world degraded images, it is prone to structural distortions in the generated images when super-resolution processing is performed on retinal fundus images, which are richer in structural information. Qiu et al.³¹ proposed an improved generative adversarial network (IGAN) for retinal image super-resolution. They designed an attention convolutional neural, constructed the pixel loss function to use the robust Charbonnier, removed the BN layer, and added multiple updated residual blocks to improve the performance of the generative adversarial network. Ma et al.¹¹ proposed a structure-preserving super-resolution method with gradient guidance to alleviate the issue of structural distortions commonly existing in the super-resolution results of GAN-based methods. The additional supervision provided by gradient maps can better capture structural information. Although the gradient map can provide additional structural supervision to the super-resolution training process, it still has limitations. The gradient only reflects the differences between adjacent pixels. For retinal fundus images, where the colorful differences across the whole image are small, the gradient map is prone to noise unrelated to the structural information, and the gradient information is not obvious. In this paper, we replace the gradient map with a segmentation map of the retinal vascular structure for retinal fundus images to alleviate the above limitations.

Methods

Due to the superior image generation capability of generative adversarial networks, it can be applied to improve the quality of retinal fundus imaging to help doctors in retinal image analysis. Although the Real-ESRGAN can generate super-resolution (SR) retinal fundus images, the SR image contains structural distortions. To overcome the problem, we proposed an improved Real-ESRGAN. It contains three parts: our improved generator, our improved discriminator, and our designed new loss function.

Improved generator

The improved generator is shown in Fig. 1. Compared with the original generator of Real-ESRGAN, we added a branch that contains a pre-trained network (U-Net model), condition network, and our designed improved RRDB (Residual in Residual Dense Block). The distribution of retinal vascular structures in fundus images is regular, relatively constant, and semantically simple and clear. The above situation allows retinal vascular structures to be located using low-resolution information. At the same time, the boundaries of different tissues within the retinal fundus image are blurred, and the fine structures are not obvious, requiring more high-resolution(HR) information to distinguish them. Therefore, we first downsample the input image to obtain deep

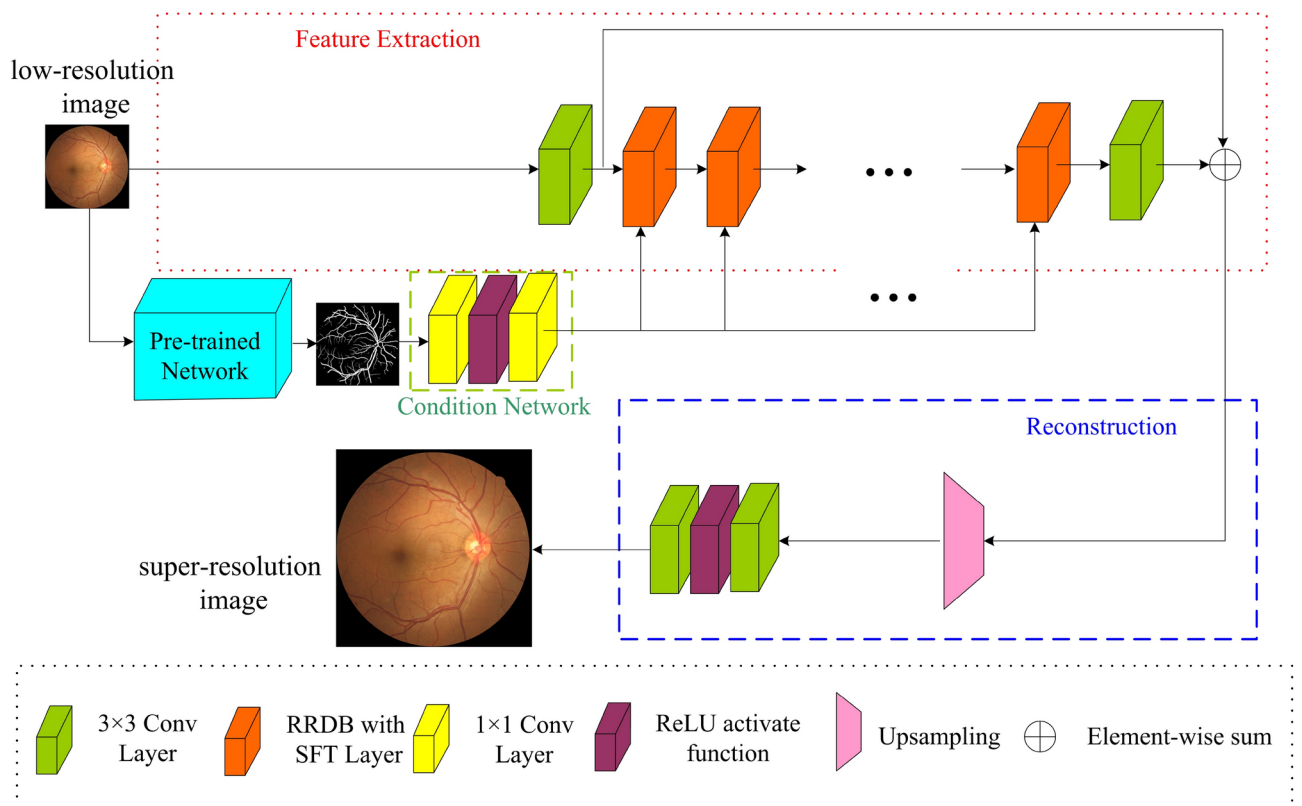


Fig. 1. The architecture of the proposed generator.

features in the U-Net model. The deep features contain LR information that can provide contextual semantic information about the segmentation target in the whole image. Secondly, we use the upsample operation to obtain shallow features that contain more HR information to segment the segmentation target in the U-Net model finely. In addition, the skip connections used in the upsampling process combine the shallow features with the deeper features of the corresponding layer to obtain more accurate segmentation³². We train the U-Net model on the DRIVE (Digital Retinal Images for Vessel Extraction) dataset³³ that is specifically designed for the segmentation of blood vessels in retinal fundus images and use the trained U-Net model as the pre-trained network in Fig. 1.

To better incorporate the features contained by the pre-trained network into the SR process, we input the vascular structure segmentation map into the condition network. The condition network consists of a 1×1 convolutional layer, a ReLU activation function, and a 1×1 convolutional layer. The condition network is used to extract feature maps from the vascular structure segmentation map without changing the size. The outputs of the condition network are referred to as the prior conditions. In the end, we use the Spatial Feature Transform (SFT) layer to fuse the prior conditions with the feature maps obtained by the original generator network to improve the quality of super-resolution of retinal fundus images.

Compared with the original RRDB (Residual in Residual Dense Block) basic module in Real-ESRGAN, we add a Spatial Feature Transform (SFT) layer before each convolutional layer in the RRDB module. The SFT layer can make the prior conditions adequately combine with each intermediate feature map obtained in the feature extraction phase. Our proposed RRDB module with SFT layers is shown in Fig. 2. Low-level vision tasks such as SR require more spatial information of the image to be considered and require different processing at different spatial locations of the image. Therefore, we use an SFT layer to combine the feature maps obtained in the feature extraction phase with the prior conditions rather than directly concatenating or summing them. The SFT layer is used to learn a mapping function that outputs a modulation parameter pair based on some prior conditions. This learned parameter pair adaptively affects the output spatially using an affine transformation for each intermediate feature map in an SR network. The affine transformation is carried out by scaling and shifting feature maps:

$$SFT(F|\gamma, b) = \gamma \times F + b \quad (1)$$

where F denotes the feature maps, whose dimension is the same as γ and b , and \times represents element-wise multiplication. The SFT layer is shown in Fig. 3, which feeds the prior conditions into two separate combinations of two convolutional layers with a kernel size of 1×1 to obtain γ and b , and then we modulate the input feature maps with γ and b , as shown in (1).

Improved discriminator

The discriminator discriminates whether the input image is the original high-resolution or super-resolution image. In the Real-ESRGAN, it uses a U-shaped network as the discriminative network. This architecture can provide detailed per-pixel feedback to the generator while maintaining the global coherence of super-resolution images by the global image feedback. It consists of three parts: the downsampling part, the upsampling part, and skip connections. However, the convolution operation extracts informative features by blending cross-channel and spatial information in the downsampling part. Therefore, it also generates some redundant feature information.

To reduce the interference of redundant feature information, we introduce the channel and spatial attention modules into each skip connection to emphasize meaningful features. The spatial and channel attention modules are shown in Figs. 4 and 5, respectively. The channel attention module is used to obtain the weight of different channels. The more important the information contained in the channel, the greater the weight of the channel.

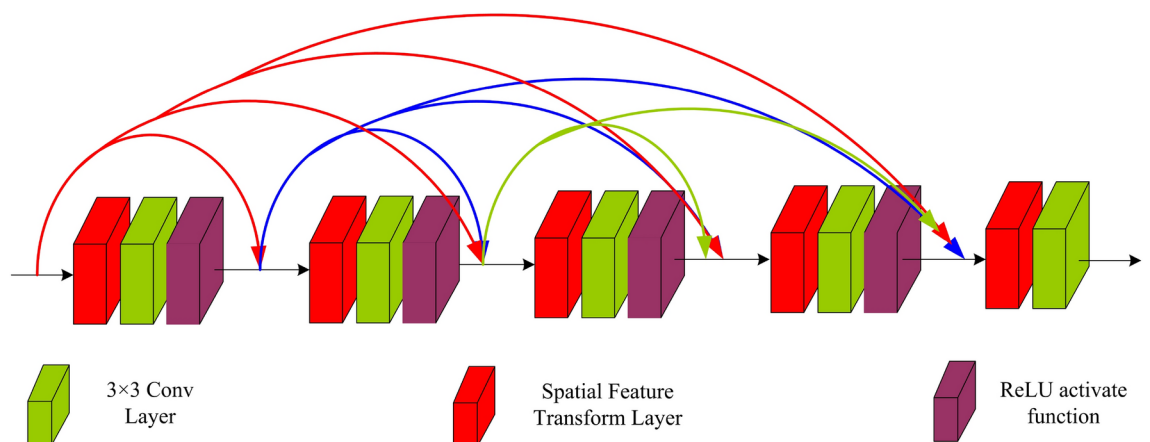


Fig. 2. The RRDB module with spatial feature transform layer.

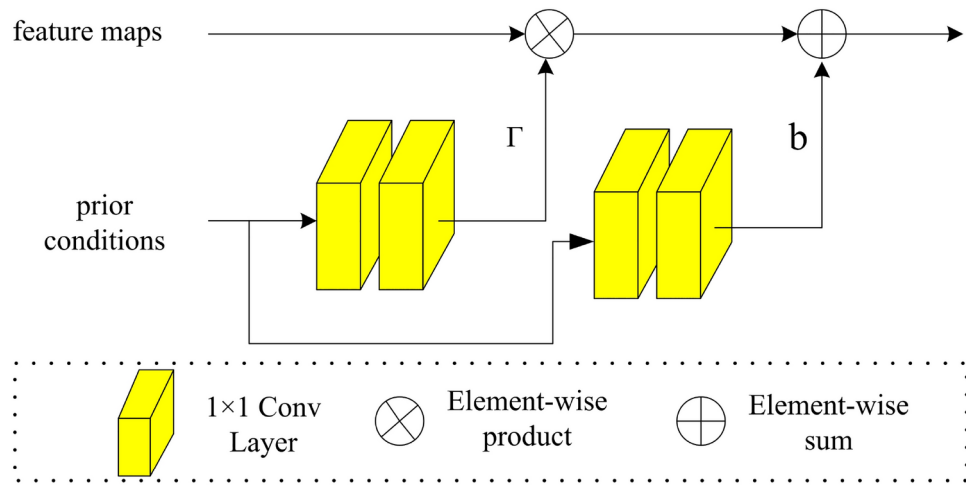


Fig. 3. The architecture of spatial feature transform layer.

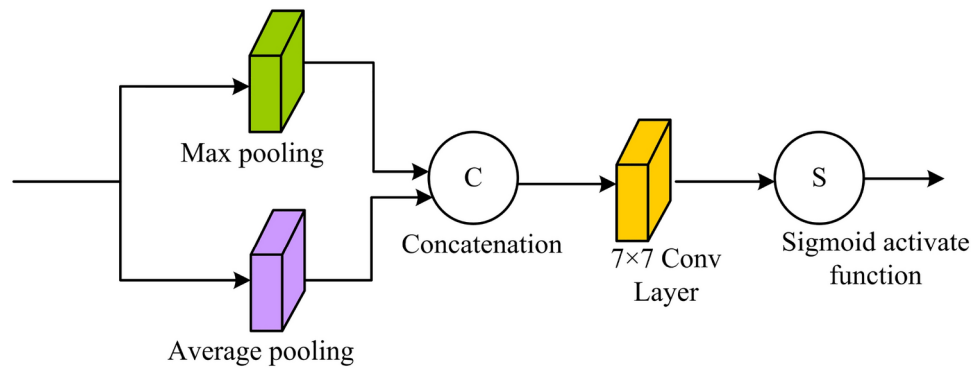


Fig. 4. Spatial attention module.

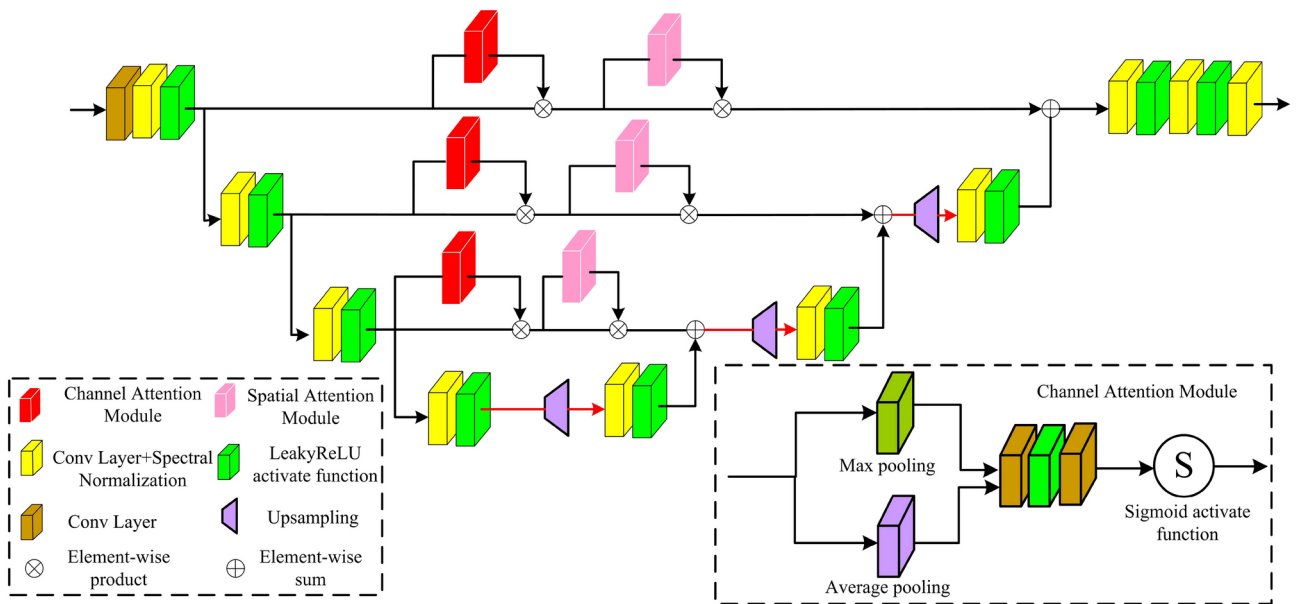


Fig. 5. The architecture of the proposed discriminator.

The output feature maps of each skip connection with attention modules and each cascade module in the upsampling part are fused by an element-wise sum operation. The fused feature maps are used as the input feature maps of the next cascade module in the upsampling part.

The complete discriminator network is shown in Fig. 5. The left part is the downsampling part. The right part is the upsampling part, and the intermediate part is the attention part. We first use a 3×3 convolution to extract features in the downsampling part. Secondly, we use three identical cascade modules to realize the downsampling. Each cascade module includes a 4×4 convolution with a stride of 2, a spectral normalization layer, and a LeakyReLU activation function. The convolution reduces the scale of feature maps and increases the receptive field. Spectral normalization is used to improve training stability. The LeakyReLU activation function is used to improve the fitting ability of the network. The upsampling part includes three identical cascade modules. The cascade module consists of a bilinear upsampling operation, a 3×3 convolution, a spectral normalization layer, and the LeakyReLU function. The bilinear upsampling operation is used to increase the scale of feature maps. The convolution, a spectral normalization layer, and a LeakyReLU activation function are used to extract high-frequency features. At the end of the discriminator, we use two convolutions with a spectral normalization layer and a convolution. It can enhance important features and reduce the interference of the generated noise, which is useful for improving the discriminatory capability of the discriminator.

Construction of new loss function

Considering the retinal vascular structure segmentation map is a two-valued image, we use the L1 loss function to measure the pixel-level differences between the retinal vascular structure segmentation maps of the high-resolution (HR) images and the super-resolution (SR) images. It is expressed as follows:

$$L_{L1-Seg} = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H \|Seg(I^{HR})_{x,y} - Seg(G(I^{LR}))_{x,y}\|_1 \quad (2)$$

where H and W refer to the height and width of the segmentation map of the retinal vascular structure of the HR image or the SR image, respectively. I^{LR} and I^{HR} represent the low-resolution (LR) image and the HR image, respectively. $Seg(\cdot)$ represents the operation of image segmentation using the U-Net model pre-trained on the DRIVE dataset. The introduction of this loss function is a secondary constraint following the introduction of the structural prior in the generator to constrain the SR images.

We use the proposed loss function and the original loss functions to construct a new loss function for the improved generative adversarial network. The new loss function of the generator and discriminator are shown in (3) and (4), respectively.

$$Loss_G = \lambda_{adv} L_{adv_G} + \lambda_{per} L_{per} + \lambda_{L1} L_{L1} + \lambda_{L1-Seg} L_{L1-Seg} \quad (3)$$

$$Loss_D = \lambda_{adv} L_{adv_D} \quad (4)$$

where $Loss_G$ and $Loss_D$ are the loss functions of the generator and discriminator, respectively. They are used to measure the difference between the generated data distribution and the real data distribution. The coefficients in Equations (3) and (4) are set as $\lambda_{adv} = 0.1$, $\lambda_{per} = 1$, $\lambda_{L1} = 1$, $\lambda_{L1-Seg} = 1$. The $Loss_G$ and $Loss_D$ are the adversarial loss functions of the generator and discriminator, respectively. They are expressed as follows:

$$L_{adv_G} = \frac{1}{N} \sum_{n=1}^N -\log(D(G(I^{LR}))) \quad (5)$$

$$L_{adv_D} = \frac{1}{N} \sum_{n=1}^N -\log(D(I^{HR})) - \log(1 - D(G(I^{LR}))) \quad (6)$$

where D and G express the generator and discriminator, respectively. N is the total number of samples in the training set. As shown in (7), this loss function is used to measure the pixel-level differences between the HR and SR images.

$$L_{L1} = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H \|I_{x,y}^{HR} - G(I^{LR})_{x,y}\|_1 \quad (7)$$

The perceptual loss function uses a pre-trained VGG19 network to calculate the gap between the HR and SR images on the feature space. The introduction of the perceptual loss function makes the SR images semantically closer to the HR images.

$$L_{per(i,j)} = \frac{1}{W_{(i,j)} H_{(i,j)}} \sum_{x=1}^{W_{(i,j)}} \sum_{y=1}^{H_{(i,j)}} \|\phi_{(i,j)}(I^{HR})_{x,y} - \phi_{(i,j)}(G(I^{LR}))_{x,y}\|_1 \quad (8)$$

where $\phi_{(i,j)}$ indicates the feature maps obtained by the j -th convolution layer (before activation) before the i -th max-pooling layer within the pre-trained VGG19 network. The values of (i, j) are $(1, 2)$, $(2, 2)$, $(3, 4)$, $(4, 4)$, $(5, 4)$. The perceptual loss function is shown in Equation (9) as follows.

$$L_{per} = \alpha_{(1,2)}L_{per(1,2)} + \alpha_{(2,2)}L_{per(2,2)} + \alpha_{(3,4)}L_{per(3,4)} + \alpha_{(4,4)}L_{per(4,4)} + \alpha_{(5,4)}L_{per(5,4)} \quad (9)$$

where the individual coefficients in (9) are: $\alpha_{(1,2)} = 0.1$, $\alpha_{(2,2)} = 0.1$, $\alpha_{(3,4)} = 1$, $\alpha_{(4,4)} = 1$, $\alpha_{(5,4)} = 1$.

Simulation and discussion

Datasets and metrics

We randomly selected 800 retinal fundus images as training images and 50 retinal fundus images as test images from the Diabetic Retinopathy dataset³⁴ and the Fundus Image Registration (FIRE) dataset³⁵. All experiments are performed between low-resolution images and high-resolution images for $4\times$ enlargement. The resolution of the input images is 256×256 , and the resolution of the output images is 1024×1024 .

To quantitatively analyze the different methods, we use three evaluation indicators that are PSNR (peak signal-to-noise ratio), SSIM (structural similarity), and LPIPS (learned perceptual image patch similarity)³⁶ to measure the quality of recovered images. The PSNR is expressed as follows:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (10)$$

The MAX_I is the maximum value of image pixel coloration. The MSE is the mean squared error that is expressed as follows:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \|I(i, j) - K(i, j)\|^2 \quad (11)$$

where m and n represent the height and width of the image, respectively. $I()$ and $k()$ represent the high-resolution image and the super-resolution image, respectively. The SSIM is expressed as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (12)$$

where X and y represent the high-resolution image and the super-resolution image, μ_x and μ_y represent the mean values of x and y , σ_x^2 and σ_y^2 represent the variance of x and y , and σ_{xy} represents the covariance of x and y , respectively. The LPIPS is expressed as follows:

$$LPIPS(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|\omega_l \odot (\hat{y}_{hw}^l - \hat{y}_{ohw}^l)\|_2^2 \quad (13)$$

where x and x_0 represent the high-resolution image and the super-resolution image, H_l and W_l represent the height and width of the output feature maps of the l th layer in a pre-trained AlexNet network, respectively. The ω_l represents a learned weight vector. The \odot means the element-wise multiplication operation. The \hat{y}_{hw}^l and \hat{y}_{ohw}^l represent the extracted features of the corresponding x and x_0 of the l th layer at location (h, w) , respectively. The higher the value of PSNR and SSIM, the better the image quality. The smaller the value of LPIPS is, the better the human perception of the image is.

Ablation study

To test the performance of different modules and the loss function proposed in this paper, we use our proposed modules and loss function to replace the corresponding modules and the loss function in the Real-ESRGAN¹⁰ method. We use our designed discriminator to replace the discriminator of Real-ESRGAN and name it "Real-ESRGAN with our designed discriminator". We use our designed generator to replace the generator of Real-ESRGAN and call it "Real-ESRGAN with our designed generator". We add the L1 loss function to measure the pixel-level differences between the retinal vascular structure segmentation maps of the HR and SR images based on the original loss functions. We name it "Real-ESRGAN with our designed new loss function". We use our designed generator and discriminator to replace the generator and discriminator of Real-ESRGAN and replace the original loss function of Real-ESRGAN with our designed new loss function to obtain our complete method. Firstly, We randomly select a retinal fundus image from the test images as the input image to test the performances of our proposed different modules. The low-resolution images obtained by downsampling the high-resolution image with the scaling factor of four and super-resolution images of different methods from the low-resolution image are shown in Fig. 6a. The high-resolution and partially enlarged images of super-resolution images are shown in Fig. 6b. By comparing with the local magnified images of high-resolution images in Fig. 6b, the image generated by Real-ESRGAN exhibits significant distortion in some vessel regions (indicated by arrow). The image generated by Real-ESRGAN with our designed discriminator shows a relatively blurred vessel region (indicated by arrow). Although the image generated by Real-ESRGAN with our designed new loss function has

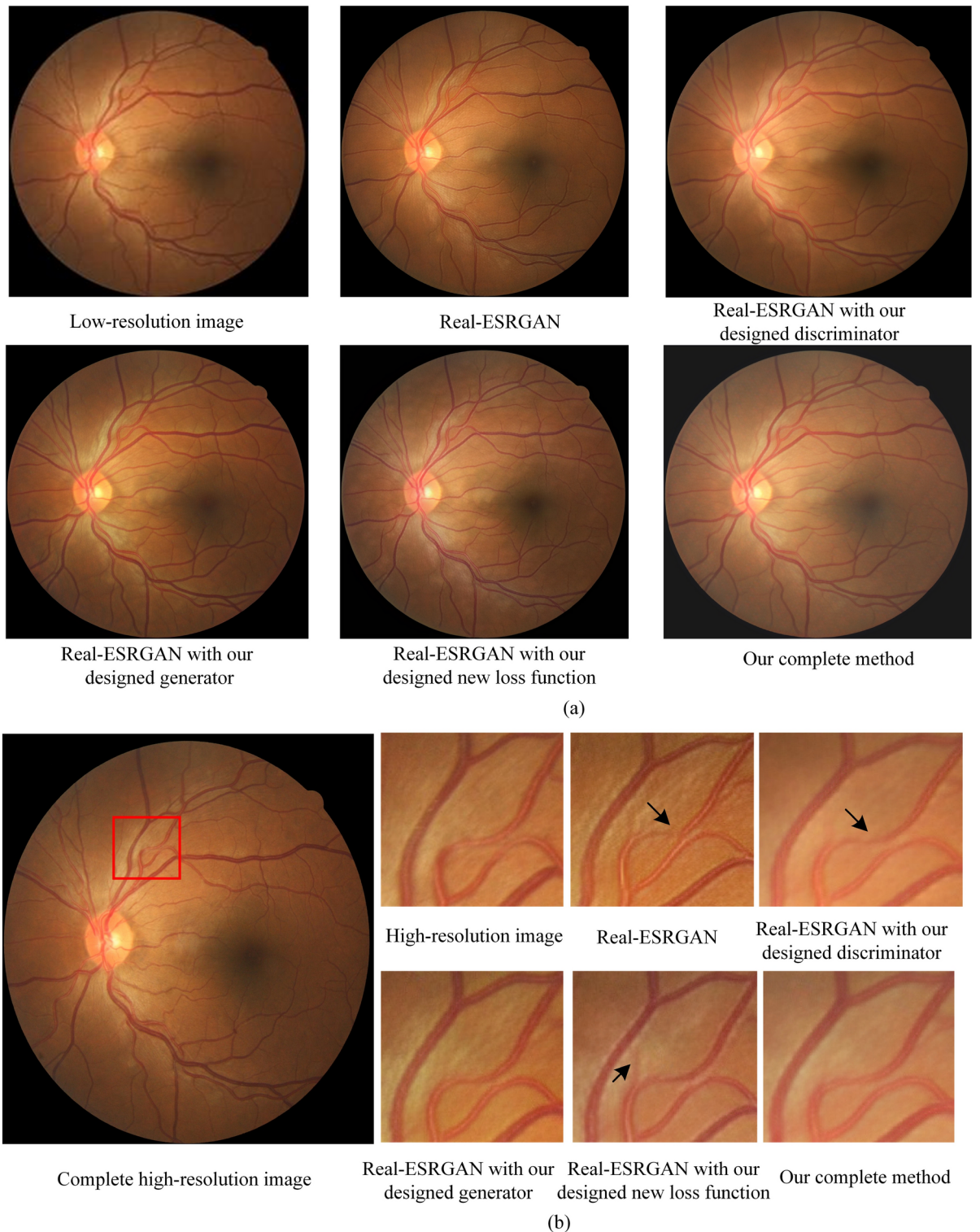


Fig. 6. The low-resolution image, high-resolution image, and super-resolution images. (a) the low-resolution image and super-resolution images; (b) the high-resolution image and partially enlarged images of super-resolution image.

less distortion in the vessel region, it contains a new vessel structure (indicated by arrow). The images generated by Real-ESRGAN with our designed generator and our complete method do not show significant distortion and are closer to the high-resolution images. Secondly, to quantitatively analyze the performance of our proposed modules, we test the above methods on 50 pairs of retinal fundus images. The average PSNR, SSIM, and LPIPS values for Real-ESRGAN and Real-ESRGAN with different modules and loss functions are shown in Table 1.

Methods	Average PSNR	Average SSIM	Average LPIPS
Real-ESRGAN	33.3948	0.8876	0.1554
Real-ESRGAN with our designed discriminator	34.2658	0.8952	0.1371
Real-ESRGAN with our designed generator	34.3494	0.8982	0.1332
Real-ESRGAN with our designed new loss function	34.1635	0.8949	0.1405
Our complete method	35.1015	0.9008	0.1220

Table 1. Average performance for Real-ESRGAN and Real-ESRGAN with different modules.

Compared with Real-ESRGAN, the Real-ESRGAN with different modules and loss functions still has a larger PSNR value, SSIM value, and smaller LPIPS value. Our complete method has the largest PSNR value and SSIM value with the smallest LPIPS value. The larger the PSNR value and SSIM value, the smaller the LPIPS value, and the better the super-resolution image's quality. Therefore, this shows that every proposed module or loss function is effective.

Performances for different methods

We randomly select two retinal fundus images from the test dataset. We use the Real-ESRGAN¹⁰, ESRGAN²⁴, RankSRGAN²⁶, P-SRGAN²⁸, DUS-GAN²⁹, IGAN³¹, and our proposed method to reconstruct the super-resolution image. The low-resolution and super-resolution images of different methods are shown in Figs. 7 and 8, respectively. The low-resolution images obtained by downsampling the high-resolution images with the scaling factor of four and super-resolution images of different methods from the low-resolution images are shown in Figs. 7a and 8a. The high-resolution and partially enlarged images of super-resolution images are shown in Figs. 7b and 8b. The super-resolution images of ESRGAN, P-SRGAN, and RankSRGAN are blurred. In Fig. 7b, the structures of both the main and branch vessels are significantly blurred in partially enlarged images of super-resolution images of ESRGAN, P-SRGAN, and RankSRGAN. Although the structures of the main vessels are clear, the structures of the branch vessels are not clear enough in the enlarged image of the super-resolution image of DUS-GAN. In Fig. 8b, there are structural distortions in the positions indicated by the arrows in the enlarged images of the super-resolution images of Real-ESRGAN. As shown as we see, our proposed method can inhibit structural distortions effectively. The super-resolution images of our method are clear and close to the original high-resolution images. The super-resolution images of our method are better at the recovery of fine structures than Real-ESRGAN.

In Figs. 8a and b, the structures of the main and branch vessels are significantly blurred in partially enlarged images of super-resolution images of ESRGAN, P-SRGAN, and RankSRGAN. The structures of the main vessels are clear in the super-resolution images of DUS-GAN and IGAN, but the structures of the branch vessels are also not clear enough. Especially, the structures of subtle branch vessels are clearer for our method than Real-ESRGAN.

To quantitatively analyze the performance of different methods, we use three evaluation indexes (PSNR, SSIM, and LPIPS) to compare the ESRGAN method, P-SRGAN method, RankSRGAN method, Real-ESRGAN method, DUS-GAN method, IGAN method, and our proposed method. The average values of different methods are shown in Table 2. The average PSNR values of the ESRGAN method, P-SRGAN method, RankSRGAN method, DUS-GAN method, Real-ESRGAN method, IGAN method, and our proposed method are 32.5327, 32.7224, 33.1178, 33.6180, 33.948, 33.7210, and 35.1015, respectively. The average SSIM values of the ESRGAN method, P-SRGAN method, RankSRGAN method, DUS-GAN method, Real-ESRGAN method, IGAN method, and our proposed method are 0.8840, 0.8821, 0.8802, 0.8879, 0.8876, 0.8882 and 0.9008, respectively. The average LPIPS values of the ESRGAN method, P-SRGAN method, RankSRGAN method, DUS-GAN method, Real-ESRGAN method, IGAN method, and our proposed method are 0.3646, 0.3124, 0.2321, 0.081, 0.1554, 0.1872, and 0.1220, respectively. As shown in Table 2, our method has the largest average PSNR, followed by IGAN and DUS-GAN. Our method has the largest average SSIM value, followed by IGAN and DUS-GAN. Our method has the smallest average LPIPS value, followed by Real-ESRGAN and IGAN. Our method still has the largest average PSNR and SSIM value with the smallest LPIPS value. It shows that our method performs better in reconstructing images than other methods. We also test the parameters, floating-point operations (FLOPs), inference speed that is measured by FPS (Frames Per Second), and memory consumption. They are shown in Table 3. It can be observed that the parameters, floating-point operations, and memory consumption of our method are all relatively high, resulting in a slower inference speed. Although our method has a slower inference speed than others, it has a better performance in reconstructing images.

Conclusion

This paper proposes an improved retinal fundus image super-resolution method based on the generative adversarial network. For the generator, we first train the U-Net model on the DRIVE dataset and use the trained U-net model as the pre-trained network to obtain the retinal vascular structure segmentation map. Secondly, we design a condition network to transform the segmentation map into feature maps. Ultimately, we use the Spatial Feature Transform (SFT) layer to fuse the transformed features by the condition network and extracted features by the original generator network. For the discriminator, we introduce channel and spatial attention modules into the discriminator to improve its discriminative power. Moreover, we also improved the loss function to measure the differences between the generated and original images. We randomly select two low-resolution retinal fundus images as the input images of ESRGAN method, P-SRGAN method, RankSRGAN method, DUS-

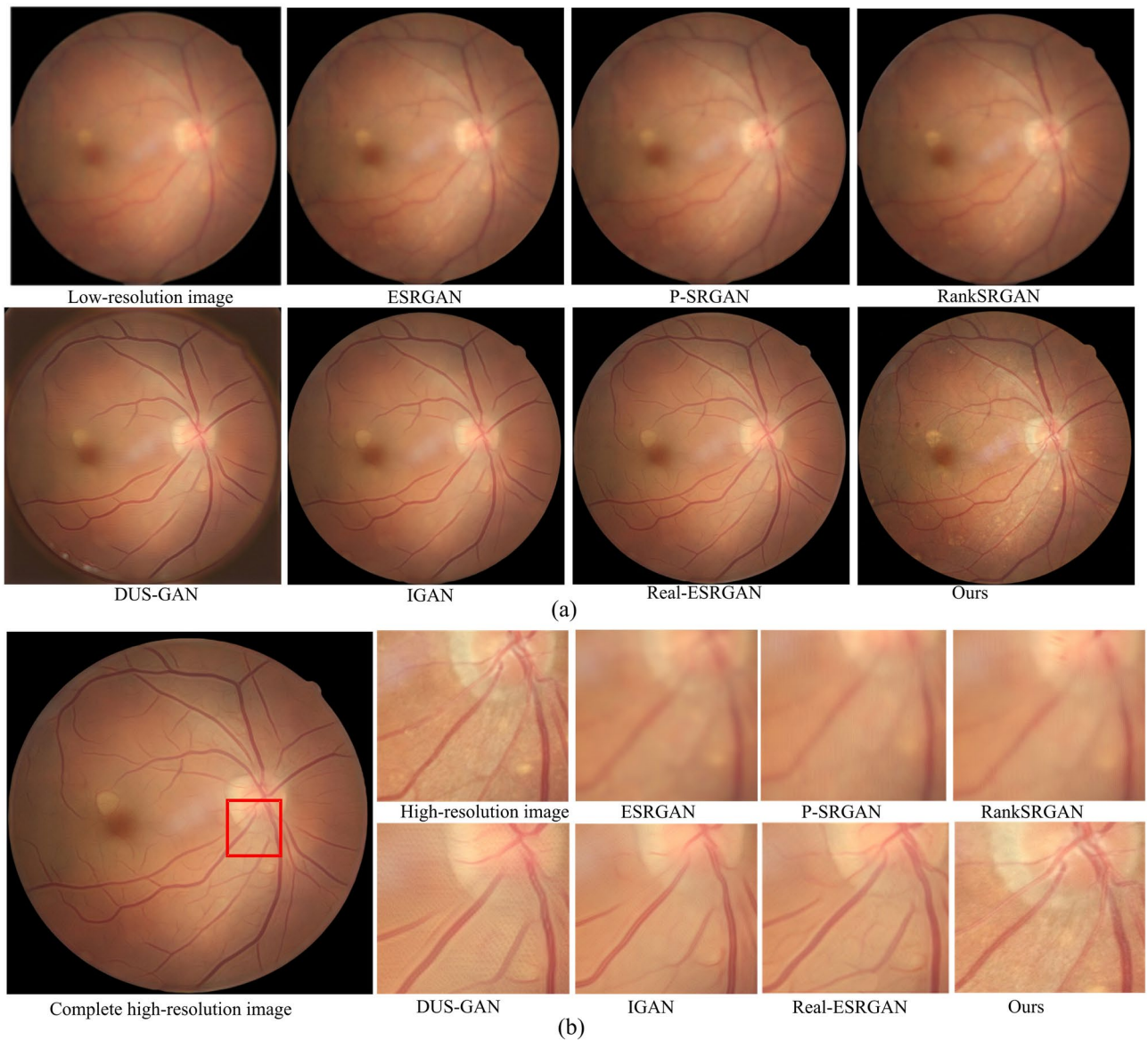


Fig. 7. The first low-resolution image, high-resolution image, and super-resolution images. (a) the low-resolution image and super-resolution images; (b) the high-resolution image and partially enlarged images of super-resolution images.

GAN method, Real-ESRGAN method, IGAN method, and our proposed method. The images obtained by our proposed method are excellent in the recovery of fine structures. Our proposed method has the largest average PSNR value and SSIM value and the smallest average LPIPS value. These show that our proposed method performs better for recovering low-resolution retinal fundus images than other methods.

Although our method demonstrates strong capabilities in image super-resolution reconstruction, the algorithm is relatively complex, leading to a high number of parameters and significant computational requirements. Therefore, in future work, we will focus on reducing network complexity while ensuring the quality of the reconstructed images. This will, in turn, improve image reconstruction speed and make the model easier to deploy on servers with lower computational power.

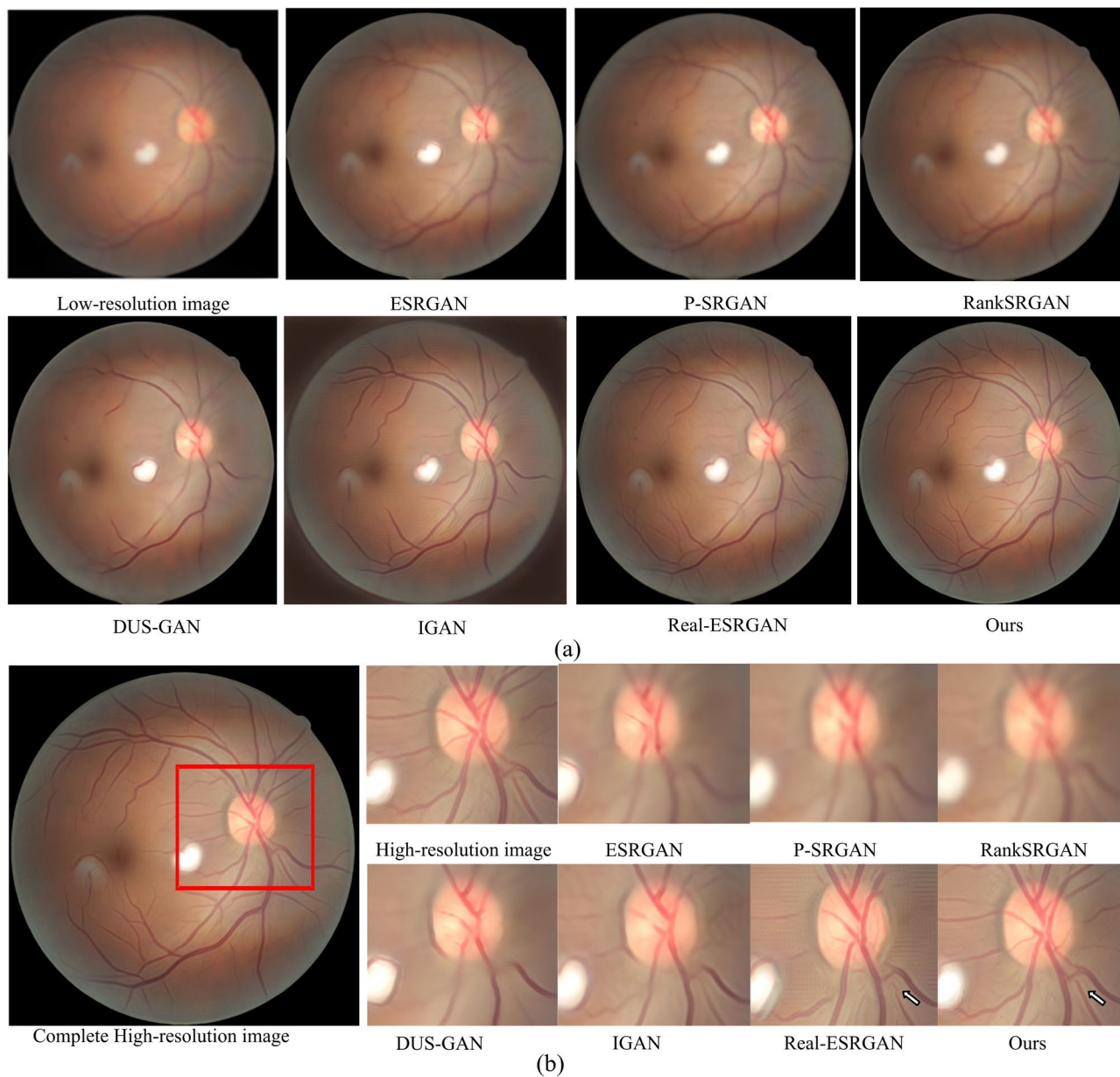


Fig. 8. The second low-resolution image, high-resolution image, and super-resolution images. (a) the low-resolution image and super-resolution images; (b) the high-resolution image and partially enlarged images of super-resolution images.

Methods average	PSNR	Average SSIM	Average LPIPS
ESRGAN	32.5327	0.8840	0.3646
P-SRGAN	32.7224	0.8821	0.3124
RankSRGAN	33.1178	0.8802	0.2321
DUS-GAN	33.6180	0.8879	0.2081
Real-ESRGAN	33.3948	0.8876	0.1554
IGAN	33.7210	0.8882	0.1872
Our method	35.1015	0.9008	0.1220

Table 2. Average values of performance index for different methods.

Methods	Params (M)	FLOPs (G)	FPS	Memory consumption (MB)
ESRGAN	16.76	7	12.8	823.08
P-SRGAN	18.1	9.2	9.7	888.89
RankSRGAN	17.3	10.4	8.6	849.60
DUS-GAN	15.4	12.6	7.08	756.291
Real-ESRGAN	19.6	15.5	5.78	962.55
IGAN	16.9	7.1	12.63	829.96
Our method	21.5	16.8	5.86	1055.86

Table 3. Parameters, floating-point operations, memory consumption, and inference speed.

Data availability

The Diabetic Retinopathy dataset is publicly available and can be accessed at <https://tianchi.aliyun.com/dataset/93926>. (DOI: 10.1109/ACCESS.2020.2980055). The Fundus Image Registration (FIRE) dataset used in this article is publicly available and can be accessed at <https://projects.ics.forth.gr/cvrl/fire/>. (DOI: 10.35119/maio.v1i4.42)

Received: 26 May 2024; Accepted: 24 September 2024

Published online: 01 October 2024

References

- Masayoshi, K. *et al.* Deep learning segmentation of non-perfusion area from color fundus images and ai-generated fluorescein angiography. *Sci. Rep.* **14**, 10801, <https://doi.org/10.1038/s41598-024-61561-x> (2024).
- Iqbal, S. *et al.* Recent trends and advances in fundus image analysis: A review. *Comput. Biol. Med.* **151**, 106277, <https://doi.org/10.1016/j.compbiomed.2022.106277> (2022).
- Shi, C. *et al.* Assessment of image quality on color fundus retinal images using the automatic retinal image analysis. *Sci. Rep.* **12**, 10455, <https://doi.org/10.1038/s41598-022-13919-2> (2022).
- Ahmad, W. *et al.* A new generative adversarial network for medical images super resolution <https://doi.org/10.1038/s41598-022-13658-4> (2022).
- Guarnieri, G. *et al.* Perspective registration and multi-frame super-resolution of license plates in surveillance videos. *Forens. Sci. Int. Dig. Investig.* **36**, 301087, <https://doi.org/10.1016/j.fsidi.2020.301087> (2021).
- Wang, Y. *et al.* Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Exp. Syst. Appl.* **197**, 116793 <https://doi.org/10.1016/j.eswa.2022.116793> (2022).
- de Farias, E. *et al.* Impact of gan-based lesion-focused medical image super-resolution on the robustness of radiomic features. *Sci. Rep.* **11**, 21361 <https://doi.org/10.1038/s41598-021-00898-z> (2021).
- Dong, S., Wang, P. & Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **40**, 100379 <https://doi.org/10.1016/j.cosrev.2021.100379> (2021).
- Ahmad, W. *et al.* A new generative adversarial network for medical images super resolution. *Sci. Rep.* **12**, 9533. <https://doi.org/10.1038/s41598-022-13658-4> (2022).
- Wang, X. T., Xie, L. B., Dong, C. & Shan, Y. Realesrgan: Training real-world blind super-resolution with pure synthetic data supplementary material. *IEEE/CVF International Conference on Computer Vision Workshops* 1905–1914 (2021).
- Ma, C., Rao, Y. M., Lu, J. W. & Zhou, J. Structure-preserving image super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7898–7911. <https://doi.org/10.1109/TPAMI.2021.3114428> (2022).
- Thapa, C., Raahemifar, k & Bobier, W. R. Comparison of super-resolution algorithms applied to retinal images. *J. Biomed. Opt.* **19**, 056002 <https://doi.org/10.1117/1.JBO.19.5.056002> (2014).
- Jebadurai, J. & Peter, Jd. Super-resolution of retinal images using multi-kernel svr for iot healthcare applications. *Future Gen. Comput. Syst.* **83**, 338–346. <https://doi.org/10.1016/j.future.2018.01.058> (2018).
- Thomas, K., Alexander, B., Katja, M., Zhang, K. & Christiane, Q. Multi-frame super-resolution with quality self-assessment for retinal fundus videos. *Med. Image Comput. Comput.-Assisted Interv.-MICCAI 2014* **8673**, 650–657. https://doi.org/10.1007/978-3-319-10404-1_81 (2014).
- Thomas, K., Heinrich, A., Maier, A. & Hornegger, J. Super-resolved retinal image mosaicing. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1063–1067 <https://doi.org/10.1109/ISBI.2016.7493449> (2016).
- Wang, W., Dong, J., Niu, S. & Chen, Y. Edge-guided semi-coupled dictionary learning super resolution for retina image. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* 1631–1634 <https://doi.org/10.1109/ISBI.2019.8759425> (2019).
- Liu, Y. & Yang, H. The color fundus image enhancement algorithm based on retinex. *Chin. J. Biomed. Eng.* **37**, 257–265. <https://doi.org/10.3969/j.issn.0258-8021.2018.03.001> (2018).
- Dong, C., Loy, C. C., He, K. & Tang, X. Learning a deep convolutional network for image super-resolution. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part IV* 13 184–199 (2014).
- Dong, C., Loy, C. C., He, K. & Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 295–307. <https://doi.org/10.1109/TPAMI.2015.2439281> (2016).
- Kim, C., Lee, J. K. & Lee, K. M. Deeply-recursive convolutional network for image super-resolution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1637–1645 <https://doi.org/10.1109/CVPR.2016.181> (2016).
- Kim, C., Lee, J. K. & Lee, K. M. Accurate image super-resolution using very deep convolutional networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1646–1654 <https://doi.org/10.1109/CVPR.2016.182> (2016).
- Ian, G. *et al.* Generative adversarial networks. *Communications of the ACM* **63**, 139–144 <https://doi.org/10.1145/3422622> (2020).
- Ledig, C., Theis, F., Huszar, L. & Caballero, J. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition* 4681–4690 (2017).
- Wang, X., Yu, K., Wu, S. & Gu, J. Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European conference on computer vision (ECCV) workshops* 63–79 (2018).
- Alexia, J. M. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734* <https://doi.org/10.48550/arXiv.1807.00734> (2018).
- Zhang, W. L., Liu, Y. H., Dong, C. & Qiao, Y. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. *Proceedings of the IEEE/CVF international conference on computer vision* 3096–3105 (2019).

27. Zhang, W. L., Liu, Y. H., Dong, C. & Qiao, Y. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7149–7166. <https://doi.org/10.1109/TPAMI.2021.3096327> (2022).
28. Dwarikanath, M., Behzad, B. & Rahil, G. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Comput. Med. Imaging Graphics* **71**, 30–39. <https://doi.org/10.1016/j.compmedimag.2018.10.005> (2019).
29. Prajapati, K., Chudasama, V. & Patel, H. Direct unsupervised super-resolution using generative adversarial network (dus-gan) for real-world data. *IEEE Trans. Image Process.* **30**, 8251–8264. <https://doi.org/10.1109/TIP.2021.3113783> (2021).
30. Schonfeld, E., Schiele, B. & Khoreva, A. A u-net based discriminator for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 8207–8216 (2020).
31. Qiu, D. F., Chen, Y. H. & Wang, X. S. Improved generative adversarial network for retinal image super-resolution. *Comput. Methods Programs Biomed* **225**, 106995, <https://doi.org/10.1016/j.cmpb.2022.106995> (2022).
32. Du, G., Cao, X., Liang, J. M., Chen, X. L. & Zhan, Y. H. Medical image segmentation based on u-net: A review. *J. Imaging Sci. Technol.* **64**, <https://doi.org/10.2352/J.ImagingSci.Technol.2020.64.2.020508> (2020).
33. He, Z. X., Li, X. X., Lv, N. Z., Chen, Y. L. & Cai, Y. Retinal vascular segmentation network based on multi-scale adaptive feature fusion and dual-path upsampling. *IEEE Access* **12**, 48057–48067. <https://doi.org/10.1109/ACCESS.2024.3383848> (2024).
34. Mateen, M. *et al.* Automatic detection of diabetic retinopathy: A review on datasets, methods and evaluation metrics. *IEEE Access* **8**, 48784–48811. <https://doi.org/10.1109/ACCESS.2020.2980055> (2020).
35. Hernandez M, C. *et al.* Fire: fundus image registration dataset. *Model. Artif. Intell. Ophthalmol.* **1**, 16–28. <https://doi.org/10.35119/maio.v1i4.42> (2017).
36. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition* 586–595 (2018).

Acknowledgements

We thank the National Natural Science Foundation of China (61271115) and the Research Foundation of Education Bureau of Jilin Province (JJKH20210042KJ)

Author contributions

Y.J. was responsible for experimental conceptualization and design, and was the main contributor to writing the manuscript; G.C. verified the experimental design; H. C analyzed and explained the experimental data. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests

Additional information

Correspondence and requests for materials should be addressed to Y.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024