IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 12 February 2023, accepted 27 February 2023, date of publication 10 March 2023, date of current version 12 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3256086

RESEARCH ARTICLE

IRE: Improved Image Super-Resolution Based on Real-ESRGAN

ZHENGWEI ZHU[1], YUSHI LEI[1], YILIN QIN[2], CHENYANG ZHU[3], AND YANPING ZHU[1]

[1]School of Microelectronics and Control Engineering, Changzhou University, Changzhou 213164, China
[2]School of Computer Science, Changzhou Technical Institute of Tourism and Commerce, Changzhou 213032, China
[3]School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, China

Corresponding author: Chenyang Zhu (zcy@cczu.edu.cn)

This work was supported in part by the Changzhou Science and Technology Program Research Project Applied Basic Research under Grant CJ20210123.

ABSTRACT Image super-resolution (SR) is a research field focusing on image degradation techniques. The High-order Deterioration Model (HDM) implemented in Real-ESRGAN has proven more effective in simulating the degradation of real-world images compared to conventional bicubic kernel interpolation. However, images reconstructed by Real-ESRGAN suffer from two significant weaknesses. Firstly, the rebuilt image is overly smooth and suffers from substantial texture information loss, resulting in a worse performance than classical models such as SRGAN and ESRGAN. Secondly, the reconstructed images exhibit better visualization effects but are entirely different from the original image, violating the principle of image reconstruction. To address these issues, this paper presents an improved image SR model based on the HDM implemented in Real-ESRGAN. The first-order degradation modeling of HDM was removed, and only the second-order degradation modeling was kept to reduce the degree of visual deterioration. PatchGAN was used as the fundamental structure of the discriminator, and a channel attention mechanism was added to the generator's dense block to enhance texture details in the reconstructed images. The L1 loss function was also replaced with the SmoothL1 loss function to improve convergence speed and model performance. The proposed model, IRE, was evaluated on various benchmark datasets and compared to Real-ESRGAN. The results show that the proposed model outperforms Real-ESRGAN regarding visual quality and measures such as RankIQA and NIQE. The study also indicates that PatchGAN, as the discriminator, reduces the average training time by approximately 28%.

INDEX TERMS Super resolution, real-ESRGAN, SmoothL1, channel attention, PatchGAN.

I. INTRODUCTION

Machine learning is a critical subfield in artificial intelligence. Deep learning is a prominent approach that aims to extract high-level abstract features from data and comprehend the underlying distribution patterns through multiple non-linear transformations. This approach generates unbiased judgments or predictions about incoming data. Image SR, a technique based on deep learning, involves converting low-resolution (LR) images into high-resolution (HR) images using specific algorithms [1], [2], [3]. This process addresses issues such as blurry or poor-quality images due to the limitations of the image acquisition environment.

Previous classical image SR models, such as SRGAN and ESRGAN, utilized bicubic kernel interpolation to generate LR images from HR images [4], [5]. However, in reality, image resolution typically suffers from various degradations in diverse combinations, and using bicubic kernel interpolation alone can only partially simulate the degradation of real-life images. To address this issue, blind SR was developed to recover unknown and complicated damaged LR images [6]. Blind SR can be separated into explicit and implicit modeling, depending on the downsampling technique employed. In this article, we only focus on explicit modeling.

The most widely used form of explicit modeling is classical degradation modeling, which employs a simple combination of four degradation processes - Blur, Resize, Noise, and JPEG Compression - to degrade an image [7], [8]. Unlike bicubic

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

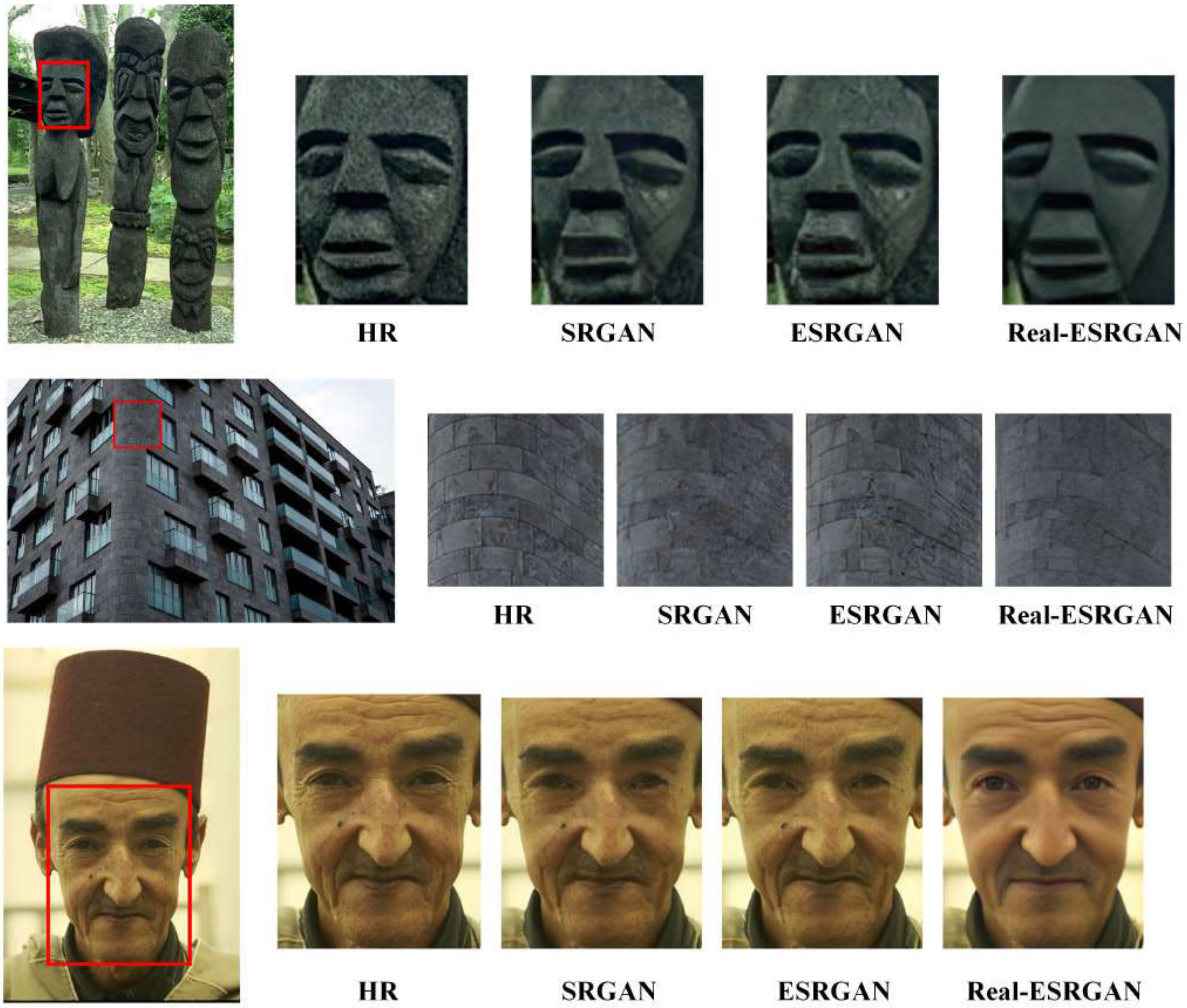


FIGURE 1. The SR results of ×4 for SRGAN, ESRGAN, Real-ESRGAN and the Ground-truth. (Sample 1: 101085 from BSD100 [30]; Sample 2: 001 from Urban100 [31]; Sample 3: 189080 from BSD100 [30]).

kernel interpolation, classical degradation modeling degrades images more complexly, with a higher degree of degradation, making it more suitable for modeling the real-life degradation of image resolution. Using classical degradation modeling, the resulting LR images are closer to the characteristics of real-life images, providing a more practical approach for image SR.

Real-ESRGAN employs the HDM downsampling technique, which models numerous rounds of the degradation process, with each iteration using a classical degradation model [9]. While HDM has been shown to simulate the deterioration of real-world images more accurately, our experiments have revealed that Real-ESRGAN is only occasionally successful in restoring specific images and yields even worse results than traditional models like SRGAN and ESRGAN, as judged by the human eye. The images produced by Real-ESRGAN exhibit two significant shortcomings. The first issue, as shown in Fig.1 for samples 1 and 2, is that the images created with Real-ESRGAN are overly smooth and nearly completely lose features like the texture of the original image. Morever, the image restored using Real-ESRGAN has the best outcomes and the most apparent clarity, as shown in example 3 in Fig.1. A closer examination of the original HR image reveals that they were two entirely separate images, defeating the intended purpose of SR. Therefore, an improved model based on Real-ESRGAN is required to address the issues in Real-ESRGAN.

Overall, the main contributions of this paper is listed as follows:

1) To address these above issues, we made the following improvements to the network structure and loss function of the model:
   - To mitigate the extent of image degradation resulting from the degradation modeling process,

we have eliminated first-order degradation modeling from HDM of Real-ESRGAN and retained only the second-order degradation modeling.

- By incorporating the channel attention mechanism with the dense block, we introduce a novel dense block structure that is deeper and more intricate [10], [11]. This structure enables the entire generator to concentrate more intently on the regions of interest, resulting in generated images endowed with greater texture details.
- We utilized PatchGAN as the fundamental structure of our discriminator. PatchGAN provides a wider perceptual domain than traditional discriminators, enabling our model to focus on more intricate visual details [12].
- We changed the loss function, substituting the SmoothL1 loss function for the original L1 loss function of Real-ESRGAN. The SmoothL1 loss function has demonstrated more excellent stability during training, manifested by markedly fewer gradient fluctuations and heightened resilience to the impact of outliers, compared to its L1 loss function counterpart [13].

2) The feasibility of the proposed model in this paper is based on a large amount of experimental data. In the ablation study, the NIQE, RankIQA, and PI metrics measured by our proposed model on different test datasets are optimal in all cases. Furthermore, the proposed model surpasses five classical SR models, including SRGAN and ESRGAN, with varying degrees of improvement in measuring the NIQE and PI metrics on different test datasets.

The remaining part of the paper proceeds as follows: Section II begins by laying out the pertinent work in image SR approaches. The section III concerns the methodology employed for this study. The experimental section of this paper is covered in section IV, and section V provides a synopsis of the whole paper.

II. RELATED WORK

A. IMAGE SUPER-RESOLUTION BASED ON GAN

Dong et al. introduced deep learning to the field of image SR [14]. They developed the network model SRCNN, which employed a three-layer Convolutional Neural Network (CNN) to learn the mapping connection between LR and HR images. SRCNN plays a fundamental role in the development of image SR since it was the first model that introduced deep learning to the industry. One notable limitation of conventional CNNs is their limited capacity to restore image texture features when faced with substantial upscaling factors. One of the most promising approaches for unsupervised learning on complex distributions in recent years is the deep learning model GAN, described by Yuan et al. [15]. Compared to traditional CNN, GAN is more suitable for image SR due to its unique principle mechanism. Fig.2 illustrates the basic idea behind image SR using GAN.

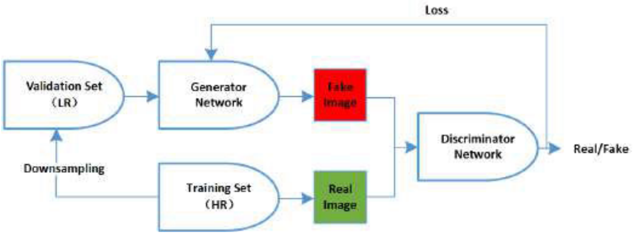


FIGURE 2. Principles of image SR based on GAN.

Ledig et al. developed the SRGAN, which utilized Generative Adversarial Network (GAN) for image SR to make a new advancement in the image SR technology based on deep learning [16]. The GAN primarily consists of a generator and a discriminator. The generator synthesizes the HR image, and the discriminator plays an adversarial game to determine whether the given image is from the generator or the actual sample, allowing the generator to reconstruct the LR image into an HR image eventually. SRGAN is the first model to apply GAN to image SR, which plays a significant role in developing the latest image SR techniques with new research directions.

Initially, the generator receives the LR images from the validation set and generates corresponding HR images through over-reconstruction. The LR images are obtained by downsampling the related HR images in the training set. Subsequently, the discriminator evaluates the integrity of the HR images produced by the generator by comparing them with actual HR images from the training set. The generator optimizes its network based on the loss value computed by the discriminator using these two images, and the process continues iteratively in an adversarial manner. Eventually, the generator learns to convert a given LR image into an HR image while deceiving the discriminator. Following the introduction of SRGAN, numerous GAN-based image SR models have been put forth, with ESRGAN [5] and Real-ESRGAN [9] being among the most well-known. Overall, development on GAN for image SR is ongoing.

B. HIGH-ORDER DEGRADATION MODEL

Several approaches can be used to enhance the performance of image SR. SRGAN introduced a novel generative network structure, SRResNet, as an alternative to the SRCNN-based structure [4]. SRGAN redesigned the loss function by proposing a perceptual loss that circumvents the problem of excessive smoothing in the reconstructed images caused by using the Mean Squared Error (MSE) loss function directly. The perceptual loss of SRGAN combines content loss and adversarial loss with variable weighting factors [16]. ESRGAN, on the other hand, incorporated the network structure of EDSR [17], replacing the residual block [18] of SRGAN with the Residual-in-Residual Dense Block (RRDB) [5], and introduced residual scaling to mitigate the adverse effects of removing the batch normalization (BN) [19] layer on the training stability of the deep network. Unlike SRGAN, the

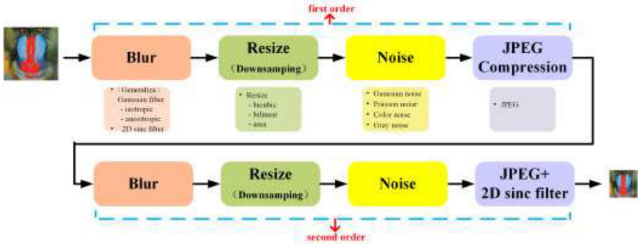


FIGURE 3. Principles of HDM with Real-ESRGAN.

loss function of ESRGAN features Relativistic average GAN (RaGAN), which computes the relative distance between the actual and generated images rather than the absolute distance [20]. This enables better discrimination between the actual and generated images during training. Furthermore, ESRGAN improved the visual quality of the images by employing the feature maps before, rather than after, the activation layer of the VGG [21] network, resulting in sharper edges and a more compelling visual experience.

To address the limitation of SRGAN and ESRGAN, Real-ESRGAN proposes a novel technique called blind SR. As shown in Fig.3, Real-ESRGAN's approach differs from SRGAN and ESRGAN by not using bicubic kernel interpolation to downsample HR images but instead relying on blind SR to simulate image degradation.

As depicted in Fig.3, Real-ESRGAN's proposed HDM constitutes a two-stage degradation modeling technique, utilizing two iterations of classical degradation procedures. Each step involves blurring, downsampling, adding noise, compressing the image, and incorporating the sinc filter to address ringing and overshooting artifacts that often arise in images with excessive segmentation. Overall, the HDM employed by Real-ESRGAN is a more accurate representation of real-life image degradation than conventional bicubic kernel interpolation.

C. CHANNEL ATTENTION MECHANISM

The depth of the CNN is a critical factor in visual SR. However, current SR networks suffer from two limitations. Firstly, training the network becomes more difficult as the network depth increases. Secondly, LR images may contain abundant low-frequency and high-frequency information. The low-frequency region is typically flat, while the high-frequency part includes features such as edges and textures. However, the network treats all data equally, resulting in a reduced expressiveness of CNN. Therefore, In order to address these concerns above, Zhang et al. proposed the channel attention mechanism [10]. The channel attention mechanism can maximize the use of information in the high-frequency part and improve the quality of the reconstructed image. Fig.4 depicts its primary structure.

The channel attention mechanism consists of several key components, including a global pooling layer, two convolutional layers(descending and ascending), and two activation functions(ReLU, Sigmoid).

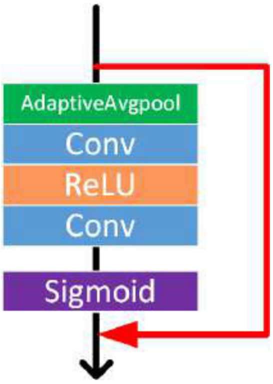


FIGURE 4. Basic structure of the channel attention mechanism.

The global average pooling operation, as denoted in Eq.(1), involves pooling the feature map (size: H*W) by taking its average [10]. The primary role of global average pooling layer(H_{GP}) is to compress and aggregate information so that information from the international sensory field of the network can be fully utilized. Each channel (x_c) in the input feature map is allocated a corresponding value (z_c). Typically, channels with high-frequency information in the image are assigned greater weights, whereas those with low-frequency information are given relatively lower weights.

$$z_c = H_{GP}\left(x_c\right) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

The gate mechanism, as represented by Eq.(2), comprises the channel-downscaling layer (W_D), the ReLU activation function (R(•)), and the channel-upscaling layer (W_U). This mechanism can effectively regulate the complexity of the model and improve its generalization capabilities [10]. Additionally, the sigmoid activation function (S(•)) is employed to facilitate individualized learning for each channel and modulate its excitation level. Each value prior to the sigmoid function determines the appropriate weight (F_C) assigned to the corresponding channel.

$$F_C = S\left(W_U * R\left(W_D * Z_C\right)\right) \tag{2}$$

Eq.(3) denotes the each feature map(x_c) is multiplied by its appropriate weight(F_C) and output [10]. In conclusion, by preserving the high frequencies of the image, the channel attention mechanism enables the reconstructed image to have a more apparent texture detail.

$$I_{\text{output}} = F_C * x_c \tag{3}$$

D. PatchGAN

Isola et al. introduced the PatchGAN, which has a different general structure than a typical GAN discriminator [12]. The input image is processed through several convolution layers and a fully linked layer or activation function before output for an ordinary discriminator. Contrarily, PatchGAN is a completely convolutional network topology. Without going through the fully connected layers or activation functions, the input image is outputted after passing through the multiple convolutional layers. As a result, the output of PatchGAN differs from that of a typical discriminator. A normal discriminator produces one evaluation value (True or False) as its output, which assesses the entire image produced by the generator or the likelihood that the input sample is valid. The result of PatchGAN is an N*N matrix in which each point (True or False) represents a tiny area of the original image, i.e., the likelihood that each area is an actual sample. The final result of the discriminator is the average of these region assessments. In images needing HR and detail, PatchGAN discriminators are more appropriate than standard GAN discriminators.

E. L1 LOSS

The L1 loss is one of the most frequently used losses function in the field of image SR. The mean absolute error function is another name for the L1 loss function. The average of the difference between the projected value of the model f(x) and the actual value y is referred to as the mean absolute error. Eq.(4) and Eq.(5) denotes the function formula and derivative formula for the L1 loss function, respectively, where x represents the discrepancy between the projected value and the actual value of the model.

$$L_1(x) = |x| \tag{4}$$

$$\frac{\mathrm{d}L_1(x)}{dx} = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \end{cases} \tag{5}$$

It has been observed that the penalty is constant for any amount of difference since the L1 loss function is derived as the absolute value of the mistake (y - f(x)). As a result, it does not cause a gradient explosion problem and has a stable gradient for any input value. However, the principal limitation of the L1 loss function is that its derivative is constant, as shown in Eq.(5). So the absolute value of the derivative of the L1 Loss function concerning the predicted value is still 1 when the difference between the predicted value and the ground truth is slight in the later stages of training. This issue severely limits the learning performance of the model, as the loss function oscillates around a steady value, hindering its convergence towards better accuracy.

III. METHOD

As shown in Fig.1, the images reconstructed by Real-ESRGAN were excessively smooth and drastically lacked image texture features. There are two main reasons for this. First off, using HDM results in excessive image deterioration, which has the opposite impact on images that are not as severely deteriorated in real life. Second, due to the unique fundamental mechanism of GAN, training the model is considerably more difficult when dealing with a complicated and high-intensity technique of image degradation. As a result, the quality of the reconstructed image effect is naturally reduced.

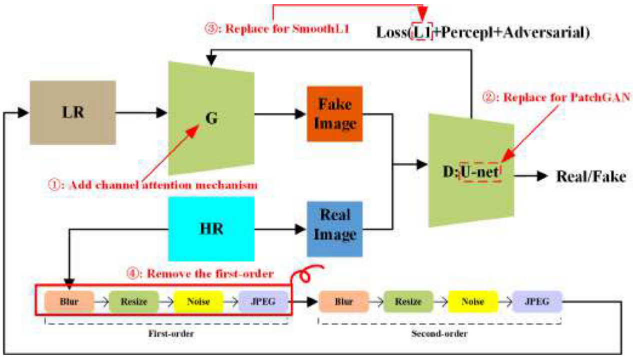


FIGURE 5. Overall framework diagram.

In order to address these above concerns, we have improved Real-ESRGAN in three aspects: HDM, network structure, and loss function. The overall framework diagram is shown in Fig.5.

A. REMOVE THE FIRST-ORDER IN HDM

In order to create a similar LR image with a significant amount of deterioration, HDM in Real-ESRGAN involves blurring, downsampling, adding noise, and compressing the image in two rounds of an HR image. Unlike SRGAN and ESRGAN, which solely employ bicubic kernel interpolation, Real-ESRGAN uses area and bilinear interpolation for the downsampling step alone. Additionally, HDM employs a randomly chosen mix of four different forms of noise — Gaussian, Poisson, Color, and Gray noise — applied to the image, together with image blurring and compression, which severely and irreversibly damage the image.

Although removing the first-order degradation process may not improve the simulation as intended, as it does not match the actual degradation process. This also reflects a problem with the current blind SR: the degree of image degradation is not well balanced with the final performance of the model. Although the higher degradation of the image can better simulate the actual degradation of the image in reality, the model cannot reconstruct it perfectly with the existing technology, and the results are uneven. Therefore, balancing the degradation level of images and the model's performance is a hot and challenging issue in the field of blind SR in the future, which is why the model proposed in this paper discards the first-order degradation modeling and keeps only the second-order degradation modeling. The schematic illustration of our adjustments is shown in Fig.6.

B. NETWORK ARCHITECTURE

1) GENERATOR WITH CHANNEL ATTENTION

The underlying design of SRResNet, split into a residual depth module and a sub-pixel convolution module, serves as the foundation for the overall structure of the Real-ESRGAN generator, as shown in Fig.7. The sub-pixel convolution module magnifies the image, while the residual depth module extracts features from the input image. The feature extraction

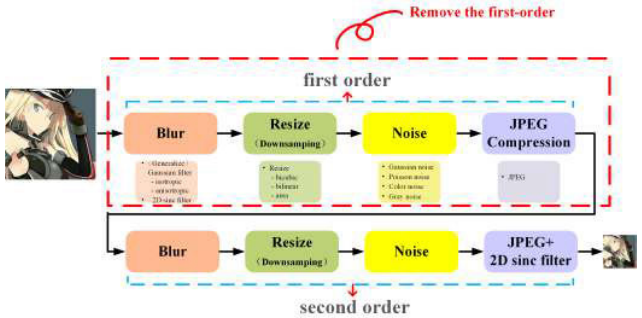


FIGURE 6. Remove the first-order and retain the second-order schematic.

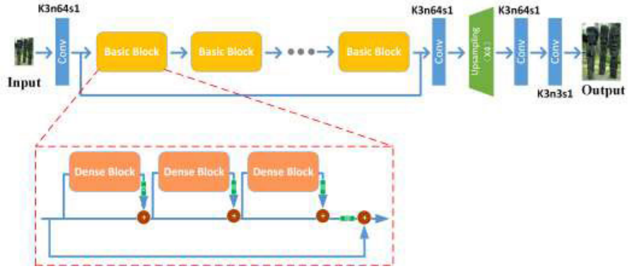


FIGURE 7. Generative network structure of Real-ESRGAN.

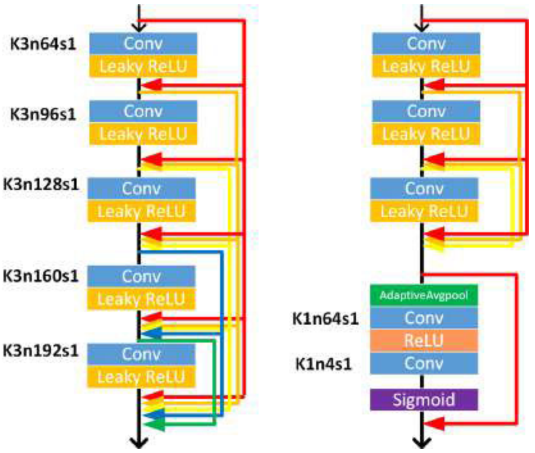


FIGURE 8. Original dense block structure (left) and improved dense block structure (right). Conv means the convolutional layer and AdaptiveAvgpool means the global average pooling layer.

block comprises three dense blocks with dense connections, as shown in Fig.8 [22]. By incorporating the dense block of Real-ESRGAN with the channel attention mechanism of RCAN, we propose a novel dense block structure based on the generative network topology of Real-ESRGAN.

The original dense block has been altered in the ways listed below. The first three convolutional layers are still present and densely connected in the dense block. A channel attention method takes the role of the final two convolutional layers. In Fig.8, the updated structure is displayed.

2) DISCRIMINATOR

To substitute the original discriminator of U-net, we developed a new discriminator illustrated in Fig.9, which employs PatchGAN as its underlying architecture. The discriminator network structure comprises of an interleaved combination of convolutional layers, maximum pooling layers, BN layers, and Leaky ReLU activation functions. The maximum pooling layers have two main functions. Firstly, the information retrieved by the convolution layer can be further reduced in dimension, which lowers the computational load. Secondly, it increases resilience in offset, rotation, and other factors of images. It also improves the invariance of image attributes. The training and convergence of the network are accelerated using the BN layer.

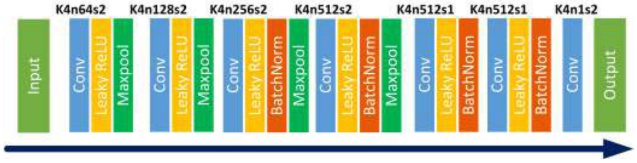


FIGURE 9. A discriminator structure designed with PatchGAN as the basic structure. Conv means convolutional layer, Maxpool means maximum pooling layer, BatchNorm means BN layer.

The PatchGAN technique enables the integration of local and global image properties by distinguishing various small patches of the original image and characterizing local image components. This approach facilitates the creation of HR images. Furthermore, averaging the resulting categorized feature map can accurately distinguish between real and fake images.

C. LOSS FUNCTIONS

L1 is the loss function in most classical models, including SRGAN, ESRGAN, and Real-ESRGAN. We employ SmoothL1 as our loss function in the model put out in this research. To make up for the limitations of the L1 loss function, we have enhanced the loss function of Real-ESRGAN by utilizing the SmoothL1 loss function instead of the L1 loss function, which is more ‘‘smooth’’ than the L1 loss function. Eq.(6) denotes the updated total loss function formula, where $L_{VGG/i,j}$ denotes the perceptual loss [23] and L_{GAN} denotes the adversarial loss [16].

$$L_{IRE} = L_{SmoothL1} + L_{VGG/i,j} + 0.1 * L_{GAN} \tag{6}$$

1) SMOOTHL1 LOSS

It has been observed from the name of the function that the SmoothL1 loss function is essentially an L1 loss function that has been smoothed, as shown in Fig.10. Eq.(7) and Eq.(8) denotes the functional and derivative formulae for the SmoothL1 loss function, respectively [13].

$$\mathrm{Smooth}_{L_1}(x) = \begin{cases} 0.5\,x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{7}$$

$$\frac{d\mathrm{Smooth}_{\mathrm{L}_1}(x)}{dx} = \begin{cases} 1, & \text{if } x \geq 1 \\ x, & \text{if } |x| < 1 \\ -1, & \text{if } x \leq -1 \end{cases} \tag{8}$$

As shown in Eq.(8), the SmoothL1 loss function also imposes an upper limit of 1 on the absolute value of the gradient of x for large x values and a modest gradient for small x values. This feature prevents the gradient from being too large and disrupting the network parameters. Researchers have suggested that the SmoothL1 loss function has advantages over the L1 loss function, as it is better suited for function convergence and model learning. Notably, Fig.10 highlights the distinct functional plots of the L1 and SmoothL1 loss functions.

2) PERCEPTUAL LOSS WITH SMOOTHL1

Following are the definitions of the SmoothL1 loss function and perceptual loss function in Eq.(7) [13]. The Eq.(9) and Eq.(10), as shown at the bottom of the page, denotes the pixel-wise L1 loss function and pixel-wise SmoothL1 loss function, respectively [4]. Eq.(7) and Eq.(10) may be compared to demonstrate that $I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y}$ in Eq.(10) is equivalent to x in Eq.(7). The difference between the anticipated and actual values is represented by x in Eq.(7). For a portion of the LR feature map with position point (x, y) in the whole image, it is defined in Eq.(10) as the difference between the reconstructed fake HR feature map and the original HR feature map. Each position of the feature map in the network is represented by (x, y), and its width and height are each represented by W and H, respectively. I^{LR} and I^{HR}, respectively, stand for the input LR image and HR image. The $G_{\theta_G}(I^{LR})_{x,y}$ shows the fabricated HR image.

Eq.(11) and Eq.(12), as shown at the bottom of the page, refers to the perceptual loss based on L1 and SmoothL1 that is obtained prior to acquiring the VGG-based activation layer, respectively [21]. This is achieved by calculating the distance between the abstracted feature maps of SR and HR, which is based on the average absolute distance using the L1 loss function in Real-ESRAGN. However, in our proposed model, we have modified the perceptual loss function to use the SmoothL1 loss function instead of the original L1 loss function for calculating the distance between the SR and HR feature maps. Thus, x in Eq.(7) is $\emptyset_{i,j}(I^{HR})_{x,y} - \emptyset_{i,j}(G_{\theta_G}(I^{LR})_{x,y})$ in Eq.(12). Here, $\emptyset_{i,j}$ denotes the feature map acquired by the j-th convolutional layer in the VGG19 network before the i-th maximum pooling layer. Additionally, $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the individual feature maps in the VGG network.

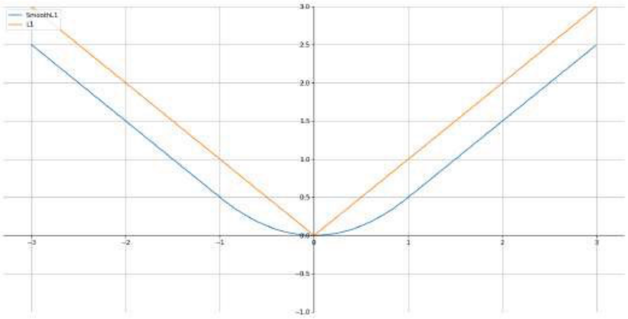


FIGURE 10. L1 loss function and SmoothL1 loss function plots.

IV. EXPERIMENTS

A. DATASETS DETAILS AND IMPLEMENTATION

The DIV2K dataset, which includes 800 HR and 800 LR images each, served as our training dataset [24]. We use MATLAB bicubic kernel interpolation to downscale the HR images into the LR images, with the downsampling factor set to 4. Since the images of the DIV2K dataset have a 2K resolution, a larger size is unnecessary for the training procedure. Therefore, before training, we first used the pre-crop script function to divide each image into sub-image blocks with overlapping parts; the size of the sub-image blocks is 480*480; this resulted in 32592 HR and LR images, respectively. Fig.11 shows a diagram of the image pre-crop.

In this experiment, the input size of the images fed into the model is 256*256, and the kernel size is 3. In the ablation study, we use HDM to degrade the images. And in the final comparison experiments, our proposed model(IRE and IRE+) uses second-order degenerate modeling. The model presented in this paper uses several different types of

$$L_{L1}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left\{\left|I_{x,y}^{HR} - G_{\theta_G}\left(I^{LR}\right)_{x,y}\right|\right. \tag{9}$$

$$L_{\text{SmoothL1}}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\begin{cases}0.5 * \left(I_{x,y}^{HR} - G_{\theta_G}\left(I^{LR}\right)_{x,y}\right)^2, & \text{if } \left|I_{x,y}^{HR} - G_{\theta_G}\left(I^{LR}\right)_{x,y}\right| < 1\\ \left|I_{x,y}^{HR} - G_{\theta_G}\left(I^{LR}\right)_{x,y}\right| - 0.5, & \text{otherwise}\end{cases} \tag{10}$$

$$L_{VGG/i,j}^{L1} = \frac{1}{\mathrm{W_{i,j}H_{i,j}}}\sum_{\mathrm{x=1}}^{\mathrm{W_{i,j}}}\sum_{\mathrm{y=1}}^{\mathrm{H_{i,j}}}\left\{\left|\emptyset_{i,j}\left(I^{HR}\right)_{x,y} - \emptyset_{i,j}\left(G_{\theta_G}\left(I^{LR}\right)_{x,y}\right)\right|\right. \tag{11}$$

$$L_{VGG/i,j}^{SmoothL1} = \frac{1}{\mathrm{W_{i,j}H_{i,j}}}\sum_{\mathrm{x=1}}^{\mathrm{W_{i,j}}}\sum_{\mathrm{y=1}}^{\mathrm{H_{i,j}}}\begin{cases}0.5 * \left(\left(\emptyset_{i,j}\left(I^{HR}\right)_{x,y} - \emptyset_{i,j}\left(G_{\theta_G}\left(I^{LR}\right)\right)_{x,y}\right)^2\right), & \text{if } \left|\left(I^{HR}\right)_{x,y} - G_{\theta_G}\left(I^{LR}\right)_{x,y}\right| < 1\\ \left|\emptyset_{i,j}\left(I^{HR}\right)_{x,y} - \emptyset_{i,j}\left(G_{\theta_G}\left(I^{LR}\right)_{x,y}\right)\right| - 0.5, & \text{otherwise}\end{cases} \tag{12}$$

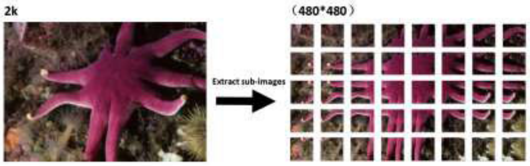


FIGURE 11. Image pre-cropping diagram(Sample: 001 from DIV2K [24]).

activation functions. In the dense block of the original generator, the activation function used is Leaky ReLU. In contrast, the generator is improved by adding a channel attention block with both ReLU and Sigmoid activation functions. For the PatchGAN discriminator used in this model, the activation function used is Leaky ReLU.

Similar to Real-ESRGAN, our implementation is built on BasicSR. Our experiments were conducted using the PyTorch framework, and training was performed on a GPU with NVIDIA Tesla P100 and GeForce GTX 1080. The tests were conducted on a CPU with an Intel(R) Core(TM) i3-8130U CPU clocked at 2.2GHz.

B. ABLATION STUDY

In order to show the effectiveness of our suggested improvement strategy, we have compiled and tested eight distinct situations based on the improvement approach. The ablation study was divided into eight distinct scenarios based on the type of loss function employed (L1 loss function or SmoothL1 loss function), the type of discriminator architecture used (U-net [25] or PatchGAN), and whether a channel attention mechanism was added to the dense block of the generator. To ensure the impartiality of the experiment, each scenario was conducted in the same experimental environment.

1) TRAINING SETTINGS

We carried out ablation study using ESRGAN because Real-ESRGAN is challenging to train. We set the batch size to 16 and the training HR patch size to 128. The pre-trained ESRGAN model is used during training and is optimized using the Adam function [26]. The number of fundamental building blocks in the generator is set to 23 by Real-ESRGAN, and we alternate the generator and discriminator until the model converges.

Eq.(13) denotes the loss function training the generators for the first to fourth scenarios of the ablation study.

$$\mathrm{L}_{1\sim4} = \mathrm{L}^{L1}_{\mathrm{VGG/i,j}} + \alpha * \mathrm{L}^{SR}_{\mathrm{L1}} + \beta * \mathrm{L}_{\mathrm{GAN}} \tag{13}$$

Eq.(14) denotes the loss function training the generators for the fifth to eighth scenarios of the ablation study.

$$\mathrm{L}_{5\sim8} = \mathrm{L}^{SmoothL1}_{\mathrm{VGG/i,j}} + \alpha * \mathrm{L}^{SR}_{\mathrm{SmoothL1}} + \beta * \mathrm{L}_{\mathrm{GAN}} \tag{14}$$

We used $\alpha = 10^{-2}$ and $\beta = 5 * 10^{-3}$ as the parameters for each set of trials. We set the number of iterations to 100k and cut the learning rate(10^{-4}) in half when the number of iterations hit 50k.

L_{GAN} denotes the adversarial loss, which consists of the generator loss(L^{G}_{Ra}) as well as the discriminator loss(L^{D}_{Ra}) as shown in Eq.(15).

$$L_{GAN} = L^{G}_{Ra} + L^{D}_{Ra} \tag{15}$$

$$\begin{aligned} L^{G}_{Ra} = &-\mathbb{E}_{I_{HR}}\left[\log\left(1 - D_{Ra}\left(I_{HR}, I_{SR}\right)\right)\right] \\ &- \mathbb{E}_{I_{SR}}\left[\log\left(D_{Ra}\left(I_{SR}, I_{HR}\right)\right)\right] \end{aligned} \tag{16}$$

$$\begin{aligned} L^{D}_{Ra} = &-\mathbb{E}_{I_{HR}}\left[\log\left(D_{Ra}\left(I_{HR}, I_{SR}\right)\right)\right] \\ &- \mathbb{E}_{I_{SR}}\left[\log\left(1 - D_{Ra}\left(I_{SR}, I_{HR}\right)\right)\right] \end{aligned} \tag{17}$$

The Eq.(16) and Eq.(17) denotes the generator (G) loss and discriminator (D) loss, respectively. I^{HR} and I^{SR}, respectively, stand for the HR image and the image reconstructed by the generator. The Ra denotes the RaGAN and $E_{I_{(\bullet)}}(\bullet)$ represents the operation of taking average for all data (super-resolved) in the mini-batch [5].

2) RESULTS

Table 1 presents the experimental results for eight distinct scenarios. The NIQE (Natural Image Quality Evaluator) [27] metric was used to evaluate the performance on five test datasets, namely Set5 [28], Set14 [29], BSD100 [30], Urban100 [31], and DIV2K100 [24]. Lower NIQE scores indicate better performance. The training time for each scenario is also reported in Table 1. Among the eight scenarios that deviated from our suggested model (scenario 8), scenario 7 achieved the best performance on all testing datasets except Set5. The highest-performing scenarios used the SmoothL1 loss function, lacked a channel attention mechanism in the generator, and used PatchGAN as a discriminator. It is worth noting that scenario 1 is essentially Real-ESRGAN without the HDM for image degradation. Thus, the NIQE values obtained by scenario 8 on Set5, Set14, BSD100, Urban100, and DIV2K100 are respectively 16.09%, 12.03%, 8.87%, 9.89%, and 13.48% lower than those of Real-ESRGAN (without HDM). Moreover, 4 hours and 43 minutes were reduced in the training period. Notably, the models were subjected to an equivalent number of iterations in the eight unique scenarios examined. In such instances, it was observed that the employment of the SmoothL1 loss function leads to an overall enhancement of model accuracy under identical experimental conditions with a faster convergence rate.

To better understand the influence of the various approaches on the performance of the model, we compiled the data in Table 1. We report the mean NIQE and training time for each testing dataset using the various approaches in Table 2. As seen from Table 2, we observed that when SmoothL1 was employed as the loss function, the training time and average NIQE metric in four of the five testing datasets were notably lower than those obtained with L1. This suggests that the SmoothL1 loss function leads to faster convergence and enables the model to achieve higher accuracy. It is common for a dense block to become more intricate and consequently take longer to train with the addition of a channel attention mechanism. Furthermore,

TABLE 1. Results of ablation study [4×upscaling](h:hour m:minute).

Number	Loss	CA	Discriminator	Testing Datasets(NIQE ↓)					Training time
				Set5	Set14	BSD100	Urban100	DIV2K100	
1	L1	N	U-net	5.8766	4.2482	3.9318	4.2674	3.5216	18h7m
2	L1	Y	U-net	5.3997	4.0058	3.7894	4.0814	3.3276	20h11m
3	L1	N	PatchGan	4.7521	3.6560	3.5242	3.8466	2.9880	13h23m
4	L1	Y	PatchGan	4.8642	3.8241	3.4930	3.8731	3.0130	13h19m
5	SmoothL1	N	U-net	5.8507	4.1260	3.7939	4.1441	3.3539	18h16m
6	SmoothL1	Y	U-net	6.0216	4.2337	3.8514	4.2144	3.4197	18h12m
7	SmoothL1	N	PatchGan	5.0286	3.6298	3.3466	3.7688	2.9875	13h31m
8	SmoothL1	Y	PatchGan	4.9309	3.7370	3.5829	3.8455	3.0469	13h24m

the average training time was reduced by approximately 28% (13.40/18.69 × 100%). The average NIQE metric was consistently lower in all scenarios when PatchGAN was used as the discriminator instead of U-net on all five test sets. Thus, it is evident that utilizing PatchGAN as the discriminator as the fundamental structure not only results in a lower NIQE measure but also significantly reduces the training time, establishing the superiority of PatchGAN.

Additionally, to further illustrate the superior performance of our suggested model. Perceptual Index (PI) [32] and RankIQA [33], both of which exhibit superior performance when their values are lower. Fig.12 and Fig.13 present the results, respectively. Fig.12 displays the PI of the model, which was determined using five test sets for eight distinct scenarios. The lower the PI, the better the model performed. Given that scenario 4(L1+CA+PatchGAN) has the lowest PI and the most excellent performance on Set5 and DIV2K100, as can be seen from the figure, the height of the column symbolized by red is the lowest. The scenario with the lowest PI on Set14 is scenario 3(L1+PatchGAN). The lowest PI was obtained in scenario 7(SmoothL1+PatchGAN) in two of the five test sets(BSD100, Urban100). In conclusion, the PI values determined by scenario 8(IRE+) on BSD100 and Urban100 are, respectively, 15.76% and 8.67% lower than those of scenario 1(Real-ESRGAN without HDM).

The measured RankIQA and proportionate size of the model for each of the eight distinct situations for each of the five test sets are shown in Fig.13. We emphasize the sector with the lowest percentage of model values on each test set, as lower RankIQA values indicate better model performance. In three of the five test sets(Set5, Set14, and DIV2K100), as seen from the figure, our suggested model(SmoothL1+CA+PatchGAN) has the lowest percentage of RankIQA measured with 11.3%, 10.9%, and 11.3%, respectively. It shows that our proposed model has the most excellent performance among the eight models. Additionally, on BSD100 and Urban100, respectively, the lowest percentages of RankIQA for scenarios 3(L1+PatchGAN) and 7(SmoothL1+PatchGAN) were found. In summary, the RankIQA values determined by scenario 7(IRE) on Set5, Set14, and DIV2K100 are, respectively, 15.20%, 20.46%, and 10.27% lower than those of scenario 1(Real-ESRGAN without HDM).

3) ANALYSIS

Our suggested model(SmoothL1+CA+PatchGAN) only provides lower metric data on RankIQA in the ablation research. Conversely, the model represented by scenario 7(SmoothL1+PatchGAN) has lower measured values in NIQE and PI. The results in ablation study show that it does not show the best performance of our suggested model. Therefore, to ensure the validity of our suggested model, we designate it as IRE(SmoothL1+CA+PatchGAN) and the model that performed the best in the ablation study as IRE+(SmoothL1+PatchGAN).

C. FINAL COMPARISON EXPERIMENT

1) TRAINING SETTINGS

Our two suggested models (IRE and IRE+) are trained using Real-ESRGAN with batch size set to 1 and training HR patch size set to 256. The Adam function parameters and the number of fundamental building blocks in the generator are the same as in the ablation study. We employ the pre-trained model of Real-ESRGAN for the training procedure. We eliminated the first-order degenerate modeling for the HDM and did not alter any other settings. We set the number of iterations to 100k and trained the generator using the loss function of Eq.(6).

2) QUANTITATIVE COMPARISON

On seven test sets(Set5, Set14, BSD100, Urban100, DIV2K100, RealSR-Canon [34], and RealSR-Nikon [34])—we utilized NIQE to evaluate two of our suggested models (IRE, IRE+) with those traditional models, such as SRResNet, BSRGAN [35], SDSR [35], Real-ESRGAN, and SwinIR [36]. In each row, we highlight the top data findings in red and the second-best results in blue. To enable comparable and fair experimental findings, we evaluated and recorded the data for each model using the same NQIE metric test script function because the training sets utilized by various models and the network topology vary significantly.

Table 3 shows that while the remaining three test sets have the lowest NIQE indicators on Real-ESRGAN, our proposed model (IRE, IRE+) achieves optimality on four of the seven test sets (Set5, Urban100, RealSR-Canon, and RealSR-Nikon). Specifically, the NIQE value measured by IRE on Set5 is 12.31% lower than Real-ESRGAN, but the NIQE value recorded by IRE+ on Urban100 was just 0.32% lower. Additionally, the NIQE values determined by IRE+ on RealSR-Canon and RealSR-Nikon are 3% and 2.41% lower than Real-ESRGAN, respectively. Except for the Urban100 and RealSR-Nikon test sets, all test sets attained optimality on our model for the sub-optimal data outcomes. In conclusion,

TABLE 2. Results of the collated ablation study.

	Testing Datasets(Average NIQE ↓)					
	Set5	Set14	BSD100	Urban100	DIV2K100	Average Training time
L1	5.2232	3.9335	3.6846	4.0171	3.2126	16.25h
SmoothL1	5.4580	3.9316	3.6437	3.9932	3.2020	15.80h
CA	5.3041	3.9502	3.6791	4.0036	3.2018	16.28h
Without CA	5.3770	3.9150	3.6491	4.0067	3.2128	15.82h
U-net	5.7872	4.1534	3.8416	4.1768	3.5057	18.69h
PatchGAN	4.8940	3.7117	3.4866	3.8335	3.0089	13.40h

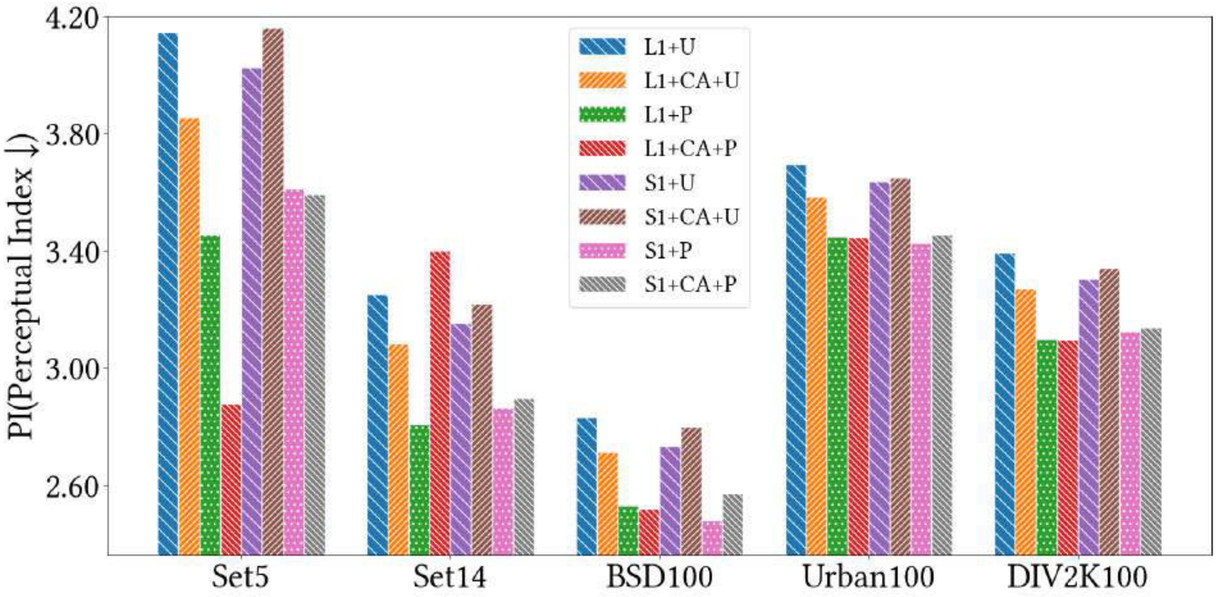


FIGURE 12. Visual bar chart: PI(Perceptual Index) measured over 5 test sets(Set5 [28], Set14 [29], BSD100 [30], Urban100 [31], DIV2K100 [24]) for eight scenarios of the model[4×upscaling]. (L1 represents L1 loss function, S1 represents SmoothL1 loss function, CA represents channel attention, U represents U-net discriminator, P represents PatchGAN discriminator.)

the quantitative comparison shows that our model incorporates 9 of the 14 data metrics (both optimal and sub-optimal), which is sufficient to demonstrate the superiority of our suggested model.

Moreover, to make our proposed model more convincing. We tested different models using PI and visualized them using line plots, the results of which are shown in Fig.14. It can be seen that our proposed model (IRE, IRE+) achieves the lowest PI on four test sets (Set5, Urban100, RealSR-Nikon, RealSR-Canon) with values of 4.403, 3.851, 3.705, 4.162, respectively, out of seven test sets, which proves the better performance of our proposed model.

Finally, to make the paperwork more convincing, we did comparison experiments based on whether the image degradation process of Real-ESRGAN uses the HDM or only second-order degradation modeling. We tested on seven test sets using the NIQE metric. And the comparison results are shown in Table 4, which shows that five test sets measured lower NIQE values when only using second-order degradation modeling, proving that Real-ESRGAN can achieve better results when only using second-order degradation modeling.

3) QUALITATIVE ANALYSIS

In order to more clearly demonstrate the viability of our suggested model, we chose a few example images from the test set and compared them qualitatively using various models. The results of the qualitative comparison are presented in Fig.15.

It has been observed that the reconstructed images using our suggested model(IRE+) exhibit crisper texture features and better human-eye outcomes than the other traditional models from the three samples in Fig.15. Additionally, the veggies in sample 2 demonstrate that even the original high-quality images cannot match the resolution of the reconstructed images using IRE+.

4) DISCUSSION AND FUTURE WORK

Real-ESRGAN is a relatively new model in the field of SR. If we do not focus on the previously reported methods but only on Real-ESRGAN, the four ways proposed in this paper are innovative for the improved model based on Real-ESRGAN. First, the change from HDM to second-order degenerate modeling is the first of its kind. Second, most SR models, including SRGAN and ESRGAN, use L1 loss

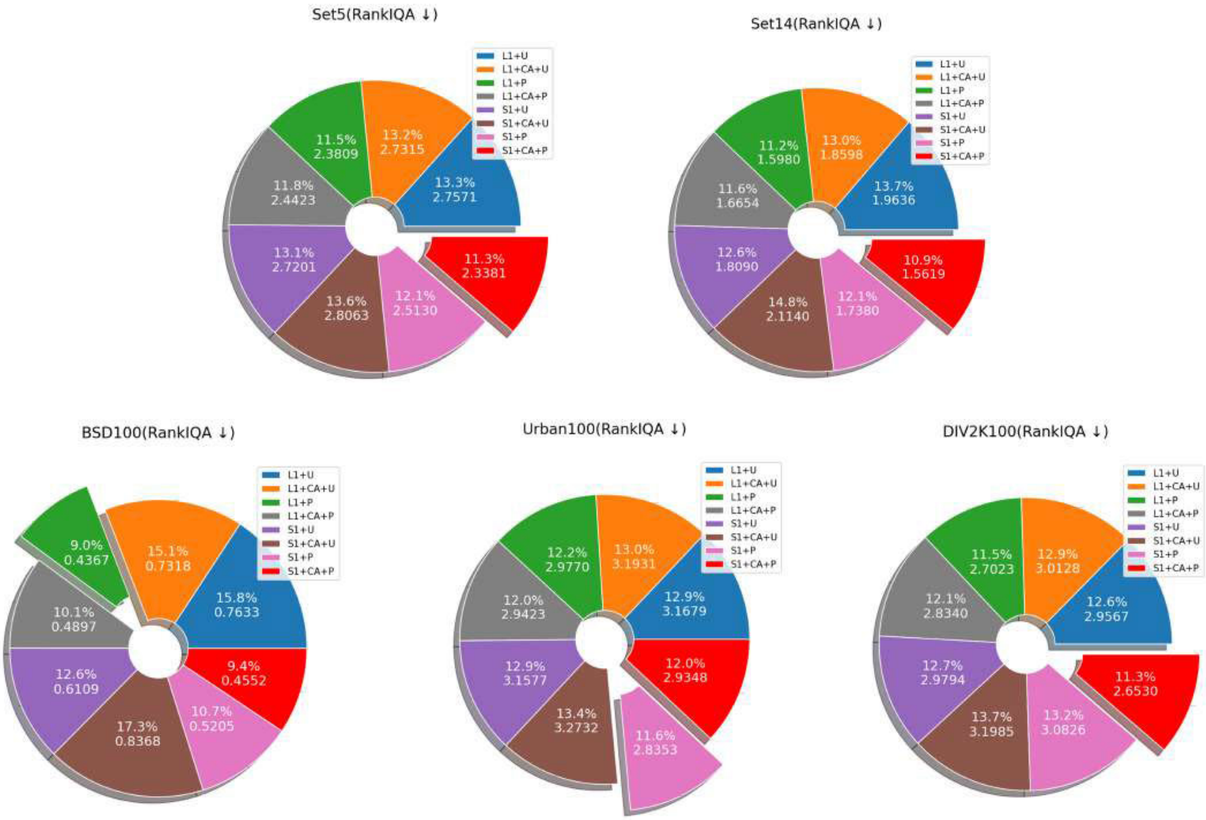


FIGURE 13. Visual pie chart: RankIQA measured over 5 test sets(Set5 [28], Set14 [29], BSD100 [30], Urban100 [31], DIV2K100 [24]) for eight scenarios of the model[4×upscaling]. (L1 represents L1 loss function, S1 represents SmoothL1 loss function, CA represents channel attention, U represents U-net discriminator, P represents PatchGAN discriminator.)

TABLE 3. Quantitative comparison results of our proposed model with other classical models, the lower the NIQE the better[4×upscaling].

	Model						
Testing Datasets(NIQE ↓)	SRResNet	BSRGAN	SDSR	Real-ESRGAN	SwinIR	IRE	IRE+
Set5	6.9816	6.2738	8.3941	6.2516	6.4591	5.4819	5.7673
Set14	6.1476	4.6630	6.2458	4.5860	5.7136	4.7679	4.5866
BSD100	6.3510	4.4497	6.3642	4.1340	5.9457	4.8760	4.3985
Urban100	5.6286	4.4481	6.0668	4.1217	5.0756	4.5410	4.1081
DIV2K100	5.6511	3.7330	5.6020	3.6391	5.1465	3.9494	3.7012
RealSR-Canon	6.2742	4.9673	7.9814	4.7195	6.0278	4.6346	4.5742
RealSR-Nikon	5.7982	4.2605	6.7343	4.1785	5.4089	4.4463	4.0779

as the loss function, while the model proposed in this paper uses SmoothL1 loss. Thirdly, the channel attention used in this paper is not considered a content innovation but a formal innovation. Unlike most models that directly put the channel attention block into the network structure, this paper replaces the channel attention block with the last two convolutional layers in the dense block in the generator. Finally, this paper uses PatchGAN as a discriminator, a first for Real-ESRAGN and other models based on its improvement.

Of course, there are several limitations in the study of this paper listed as follows:

- The disadvantage of the PSNR and SSIM metrics commonly used in the SR domain is that the metric values measured in the images reconstructed by Real-ESRGAN are often not the best to demonstrate their performance [37]. As a result, NIQE, RankIQA and PI was employed as the ultimate reference indication, and this argument was made throughout the experiments in this paper. However, some of the images produced by the model suggested in this study do not provide the best illustrations regarding how they affect the human eye while having the lowest NIQE, RankIQA, and PI metrics, among other classical models.
- The NIQE values in Table 1 obtained with the channel attention mechanism included were higher than those tested without it. It is because we employed the original generator to train directly in this research rather than pre-training the generator with the channel attention mechanism. Due to this distinction, the formally trained

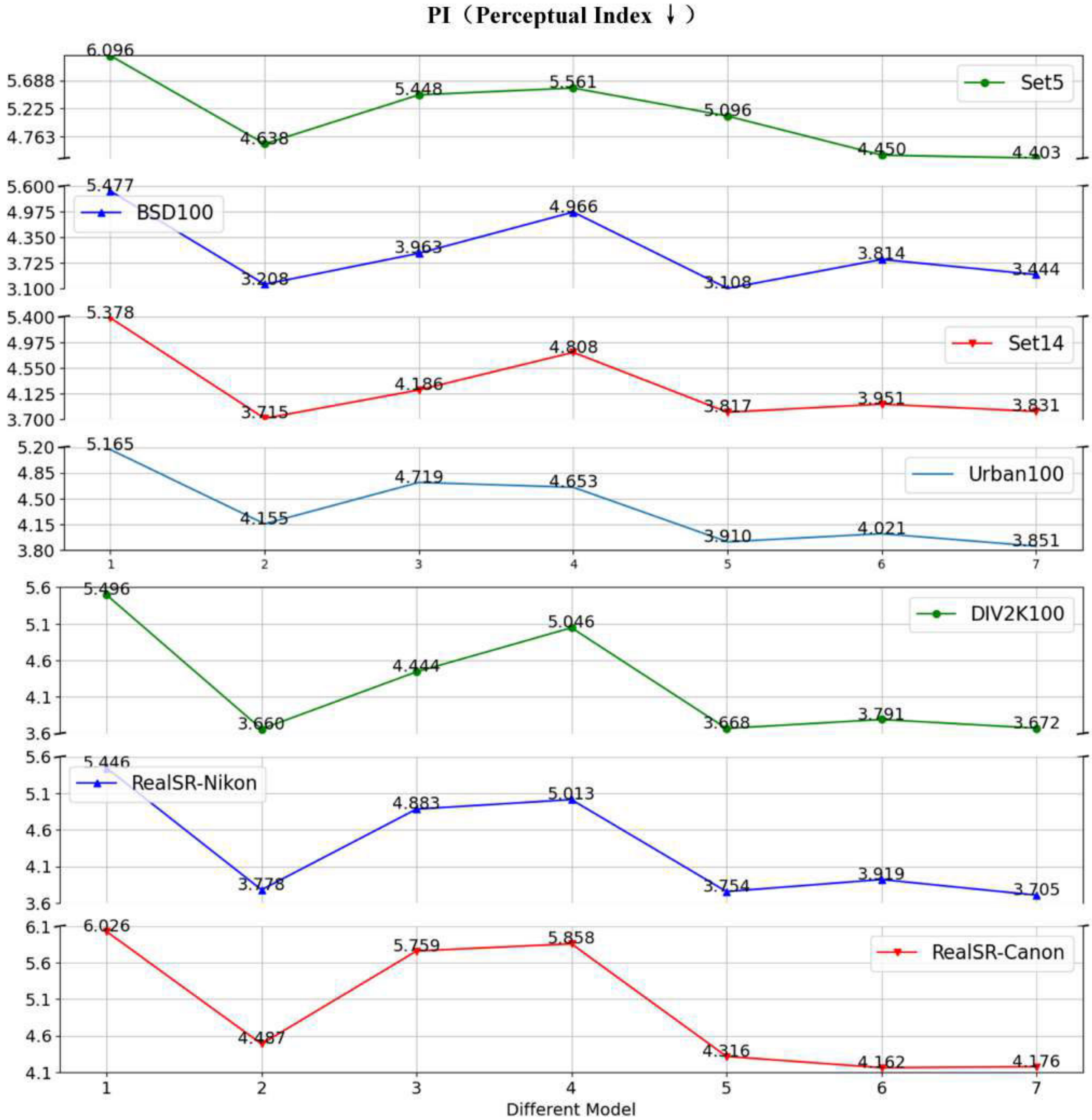


FIGURE 14. Quantitative comparison results of our proposed model with other classical models, the lower the PI the better[4×upscaling]. (In the horizontal coordinate, 1 represents SRResNet [4], 2 represents BSRGAN [35], 3 represents SDSR [35], 4 represents SwinIR [36], 5 represents Real-ESRGAN [9], 6 represents IRE, 7 represents IRE+.)

TABLE 4. Quantitative comparison results of Real-ESRGAN with HDM or the only second-order degradation modeling, the lower the NIQE the better[4×upscaling].

	Model	
Testing Datasets(NIQE ↓)	Real-ESRGAN(HDM)	Real-ESRGAN(only the second-order degradation modeling)
Set5	6.0880	6.2863
Set14	4.8355	4.7248
BSD100	4.3649	4.1962
Urban100	4.3790	4.3663
DIV2K100	3.7754	3.6697
RealSR-Canon	4.7147	4.8033
RealSR-Nikon	4.0873	4.0733

model fails to deliver the expected outcomes when put to the test. We accept that this is a flaw in the experiments presented in this paper, and we will take what we've learned from it to improve our future trials.

- The major limitation of image SR is that most existing models only target upscaling factors of 4× and less. Therefore, more study is needed to develop a model targeting 8× or even more upscaling factors.

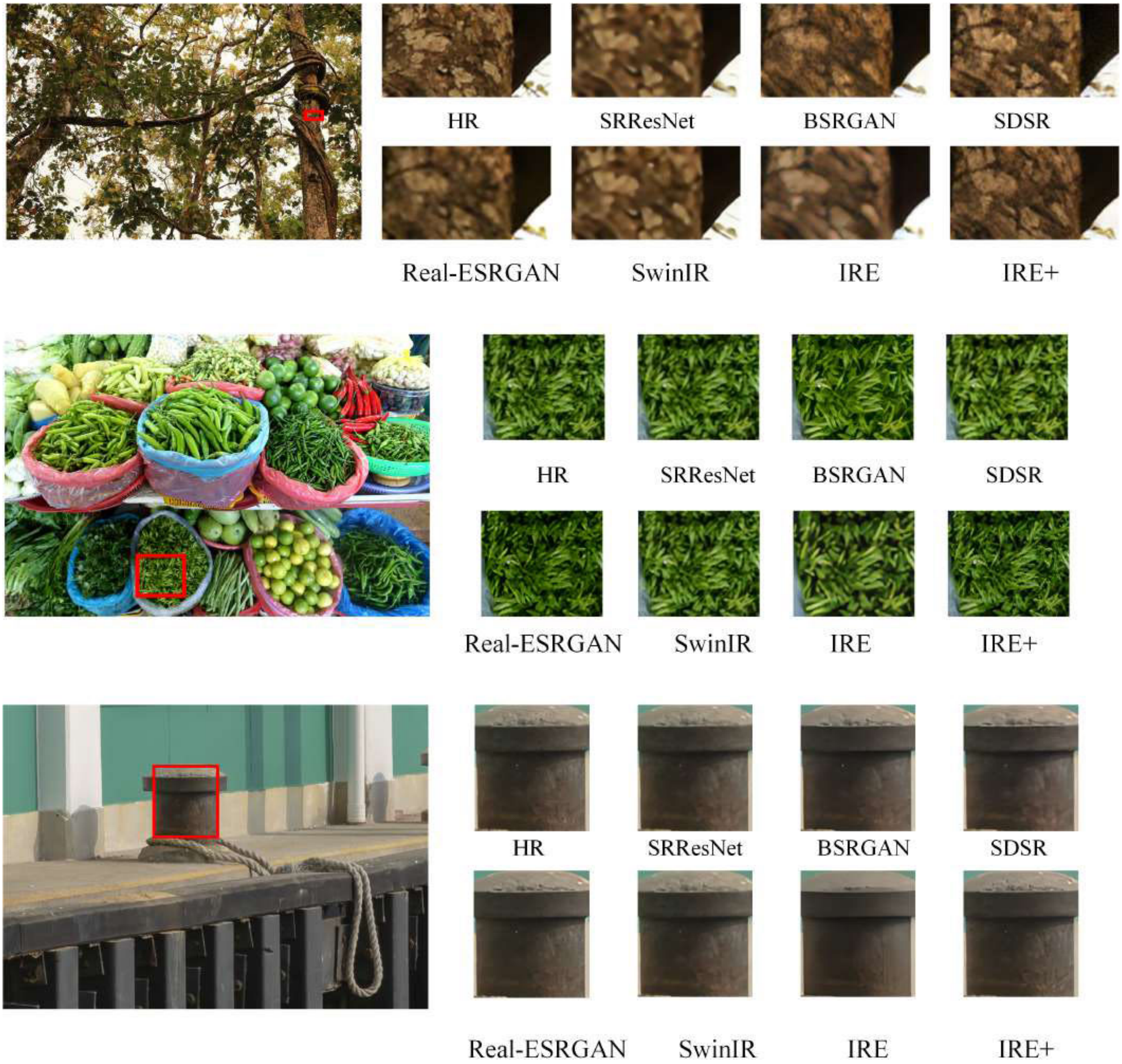


FIGURE 15. Qualitative comparison results of our proposed model with other classical models[4×upscaling]. (Sample 1: 0849 from DIV2K100 [24]; Sample 2: 0889 from DIV2K100 [24]; Sample 3: 08 from RealSR-Canon [34].

In comparison with previous models, the HDM employed in Real-ESRGAN can more accurately simulate the process of image degradation in real-world scenarios. However, as the degree of image degradation increases, the difficulty of model training also increases, and image restoration becomes more challenging. Thus, achieving a balance between image degradation and restoration presents a significant challenge for future research, necessitating the development of new network structures and advanced methods.

V. CONCLUSION

The main goal of the study was to develop an improved model based on Real-ESRGAN. In the final comparison experiments, we compare the model proposed in this paper with five authoritative models, such as BSRGAN and Real-ESRGAN. The experimental results show that the two traditional metrics(NIQE and PI) measured on the five test datasets(Set5, Set14, BSD100, Urban100, and DIV2K100) commonly used in the SR domain are optimal, which is sufficient to prove the superiority of our proposed model. In summary, these experiments confirm that the model(IRE and IRE+) proposed in this paper can generate images with more texture detail and lower metrics than the previous classical model.

The proposed method based on the improved model by Real-ESRGAN in this paper achieves better results in both qualitative and quantitative aspects. First, the model can attain higher accuracy because the SmoothL1 loss function has better convergence than the L1 loss function. Second, compared to the original discriminator, the discriminator created using PatchGAN as the fundamental structure reconstructs the image with more distinct texture features. Finally, most of the RankIQA and PI metrics measured by our proposed model on five different test datasets outperformed Real-ESRGAN.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers who provided helpful suggestions to improve the quality of the paper.

REFERENCES

[1] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 349–356.

[2] P. N. Srinivasu, T. B. Krishna, S. Ahmed, N. Almusallam, F. K. Alarfaj, and N. Allheeib, "Variational autoencoders-BasedSelf-learning model for tumor identification and impact analysis from 2-D MRI images," *J. Healthcare Eng.*, vol. 2023, pp. 1–17, Jan. 2023.

[3] Q. Wang, H. Zhou, G. Li, and J. Guo, "Single image super-resolution method based on an improved adversarial generation network," *Appl. Sci.*, vol. 12, no. 12, p. 6067, Jun. 2022.

[4] C. Ledig, L. Theis, F. Huszar, 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[5] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018.

[6] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 945–952.

[7] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1646–1658, Dec. 1997.

[8] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[9] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.

[10] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[15] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 701–710.

[16] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

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

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[20] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

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

[24] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2012.

[28] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. BMVC*, 2012, pp. 1–135.

[29] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.

[30] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[31] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[32] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.

[33] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

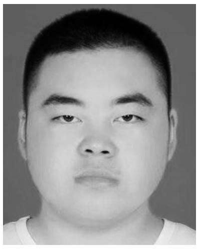
[34] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3086–3095.

[35] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.

[36] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

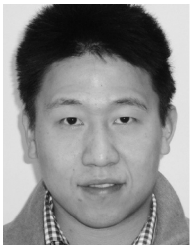
[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

ZHENGWEI ZHU received the B.S. degree in electrical technology from Southeast University, Jiangsu, China, in 1984, the M.S. degree in automation instrumentation from the East China University of Science and Technology, Shanghai, China, in 1997, and the Ph.D. degree in measurement technology and instrumentation from the Nanjing University of Science and Technology, Jiangsu, in 2006. He is currently a Professor with the School of Microelectronics and Control Engineering, Changzhou University.



YUSHI LEI is currently pursuing the degree in circuits and systems with Changzhou University, Jiangsu, China. His current research interests include deep learning and computer vision.

YILIN QIN is currently the Principal with the Changzhou Technical Institute of Tourism and Commerce, Changzhou, China.



CHENYANG ZHU received the B.S. degree in communication engineering from the Huazhong University of Science and Technology, Hubei, China, in 2012, the M.S. degree in embedded engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2014, and the Ph.D. degree in computer science from the University of Southampton, Southampton, U.K., in 2020. Since 2020, he has been a Lecturer with the School of Computer and Artificial Intelligence, Changzhou University.

YANPING ZHU received the B.S. degree in electronics and information technology, the M.S. degree in communication and information systems, and the Ph.D. degree in signal and information processing from the Nanjing University of Aeronautics and Astronautics, Jiangsu, China, in 2001, 2004, and 2010, respectively. She is currently an Associate Professor with the School of Computer and Artificial Intelligence, Changzhou University.