Fine-grained Attention and Feature-sharing Generative Adversarial Networks for Single Image Super-Resolution

Yitong Yan, Chuangchuang Liu, Changyou Chen, Xianfang Sun, Longcun Jin*, *Member, IEEE,* and Xiang Zhou

***Abstract*—The traditional super-resolution methods that aim to minimize the mean square error usually produce the images with over-smoothed and blurry edges, due to the lose of high-frequency details. In this paper, we propose two novel techniques in the generative adversarial networks to produce photo-realistic images for image super-resolution. Firstly, instead of producing a single score to discriminate images between real and fake, we propose a variant, called Fine-grained Attention Generative Adversarial Network for image super-resolution (FASRGAN), to discriminate each pixel between real and fake. FASRGAN adopts a Unet-like network as the discriminator with two outputs: an image score and an image score map. The score map has the same spatial size as the HR/SR images, serving as the fine-grained attention to represent the degree of reconstruction difficulty for each pixel. Secondly, instead of using different networks for the generator and the discriminator in the SR problem, we use a feature-sharing network (Fs-SRGAN) for both the generator and the discriminator. By network sharing, certain information is shared between the generator and the discriminator, which in turn can improve the ability of producing high-quality images. Quantitative and visual comparisons with the state-of-the-art methods on the benchmark datasets demonstrate the superiority of our methods. The application of super-resolution images to object recognition further proves that the proposed methods endow the power to reconstruction capabilities and the excellent super-resolution effects. The code is available at https://github.com/Rainyfish/FASRGAN-and-Fs-SRGAN.**

***Index Terms*—Fine-grained attention, feature-sharing, generative adversarial network, image super-resolution.**

I. Introduction

SINGLE image super-resolution (SISR), which aims to recover a high-resolution (HR) image from its low-solution (LR) version, has been an active research topic in computer graphic and vision for decades. SISR has also attracted increasing attention in both academia and industry, with applications in various fields such as medical imaging, security surveillance, remote sensing, object recognition and so on. However, SISR is a typically ill-posed problem due to the irreversible image degradation process, *i.e.*, multiple HR images can be generated from one single LR image. Learning the mapping between HR and LR images plays an important part in addressing this problem.

Recently, deep convolution neural networks (CNNs) have been shown great success in many vision tasks, such as image classification, object detection, and image restoration. Dong *et al.* [1] firstly proposed a three-layer CNN for single image super-resolution (SRCNN) to directly learn the complex non-linear mapping from LR to HR images. Since then the CNN-based methods have been dominant for the SR problem because they greatly improved the reconstruction performance. Kumar *et al.* [2] tapped into the ability of polynomial neural networks to hierarchically learn refinements of a function that maps LR to HR patches. VDSR [3] obtained the remarkable performance by increasing the depth of the network to 20, proving the importance of the network depth for detecting effective features of images. FSRCNN [4] accelerated the network training by directly extracting features from LR images instead of interpolated images, which greatly reduced the computation cost. Yang *et al.* [5] proposed a deep recurrent fusion network (DRFN) for SR with large-scale factors, which used transposed convolution to jointly extract and upsample raw features from the input and used multi-level fusion for reconstruction. EDSR [6] removed unnecessary batch normalization layer in the ResNet [7] architecture and widened the channels. EDSR significantly improved the performance of SISR and won the first place in the NTIRE 2017 Super-Resolution Challenge [8]. There are more recent methods for SISR based on the work of EDSR. For example, Zhang *et al.* introduced the Residual Dense Network (RDN) [9] to extract hierarchical features, proving the effectiveness of residual dense architecture. RCAN [10] applied residual in residual structure to construct a very deep network and used a channel attention mechanism to adaptively rescale features.

The aforementioned methods use the optimization idea of minimizing the mean squared error (MSE) between the recovered SR image and the corresponding HR image. Such methods are designed to maximize the peak signal-to-noise ratio (PSNR). However, PSNR-oriented methods typically produce over-smoothed edges and lose tiny textures. In order to produce photo-realistic SR images, Ledig *et al.*[11] firstly introduced the residual learning within the generative adversarial network (GAN) [12] framework to decrease the distance between the distributions of real images and SR images. Yan *et al.* [13] proposed a novel full-reference image quality assessment (FR-IQA) approach for SISR, *i.e.*, a loss function called SR-IQA. It was combined with L_2-Norm to guide their proposed SISR network to achieve better results. ESRGAN [14] further extends the network to produce more photo-realistic images. However, as shown in Fig.1, the discriminator in these GAN-based methods only outputs a score of the whole input SR/HR image, which is a coarse

Y. Yan and C. Liu contribute equally in this work and share the first authorship.

*Corresponding author: Longcun Jin.

way to guide the generator. Furthermore, the previous GAN-based methods typically use two independent networks for the generator and the discriminator to generate photo-realistic images and discriminate the HR image and the generated SR image, respectively. However, the shallow parts (first several layers) of the two networks both aim at extracting tiny features such as corners and edges, which we believe should be correlated.

To address these limitations, we propose two novel techniques in the GAN framework for image super-resolution, a fine-grained attention mechanism for the discriminator and a feature-sharing network component for both the generator and the discriminator. Specifically, we use a Unet-like [15] discriminator (Fig.2) to introduce a fine-grained attention in the GAN (FASRGAN). Our discriminator produces two outputs, a score of the whole input image and a fine-grained score map of every pixel in the image. The score map has the same spatial size as the input image, and measures the degree of differences at each pixel between the generated and the true distributions. To produce better high-quality images, we incorporate the score map into the loss function as an attention to make the generator pay more attention on the difficult reconstructing parts of the image, instead of treating all parts equally. In addition, we propose a feature-sharing mechanism (Fig.3) to align shallow feature extraction of both the generator and the discriminator (Fs-SRGAN). This novel structure can significantly reduce the number of parameters and improve the performance.

Overall, our main contributions are three-fold:

- We propose a novel Unet-like discriminator to generate a score of the whole image as well as a pixel-wise score map of the input image. We further incorporate the score map into the loss function as the attention mechanism for the generator. This attention mechanism makes the generator focus on the parts of an image that are difficult to generate.
- We introduce a feature-sharing mechanism to define the shallow feature extraction for the generator and the discriminator. This reduces the number of model parameters and helps the generator and the discriminator extract more useful features, which can also improve the model performance.
- The proposed two components are general, and can be applied to other GAN-based SR models. Extensive experiments on benchmark datasets illustrate the superiority of our proposed methods compared with current state-of-the-art methods.

The remainder of the paper is organized as follows. Section II describes related works. The proposed GAN-based methods are presented in Section III. Experimental results are discussed in Section IV. Finally, the conclusions are drawn in Section V.

II. Related Work

Traditional SISR methods are exemplar or dictionary based. However, these methods are limited by the size of datasets or dictionaries, and are usually time-consuming. These shortcomings can be greatly alleviated by the recent CNN-based methods.

In their pioneer work, Dong *et al.* [1] applied convolutional neural networks with three layers for SISR, namely SRCNN, to learn a mapping from LR to HR images in an end-to-end manner. Kim *et al.* [3] increased the depth of the network and introduced residual learning to the SISR network, called VDSR. VDSR achieved great improvement in accuracy compared to SRCNN. Later, Kim *et al.* [16] used a deeply-recursive convolutional network (DRCN) to reconstruct SR image, which has a very deep recursive layer. DRRN [17] introduced recursive blocks for stabilizing the training. However, the inputs of all these methods are interpolated LR images with the same size as HR images, thus greatly increasing the computation complexity and losing some details. FSRCNN [4] extracted features from the origin LR images and upscaled the spatial size by upsampling layers at the tail of the network. This architecture is widely used in the subsequent image super-resolution methods. Various advanced upsampling structures have been proposed recently, for instance, deconvolutional layer [18, 19], sub-pixel convolution [20], and EUSR [21]. LapSRN [22] and MSLapSRN [23] progressively reconstructed an HR image with increasing scales of an input image by the Laplacian pyramid structure. MRFN [24] employed multi-receptive-filed module to extract different features from different receptive fileds and fused them with a module for learning object/part-depending mappings. Besides, it proposed a two-parameter training loss (Weighted Huber) to adaptively adjust the value of back-propagated derivative according to the residual value. Lim *et al.* [6] proposed a very large network (EDSR) and its multi-scale version (MDSR), which removed the unnecessary batch normalization layer in the ResNet [7] and greatly improved super-resolution performance. D-DBPN [25] introduced an error-correcting feedback mechanism to learn relationships between LR features and SR features. ZSSR [26] uses a unsupervised method to learn the mapping between HR images and LR images. SRMDNF [27] tackled multiple degradation problems in a single network by treating degradation maps and images as inputs. RDN [9] combined dense and residual connections to make full use of information of LR images. Different from RDN, MS-RHDN [28] proposed multi-scale residual hierarchical dense networks to extract multi-scale and hierarchical feature maps. RNAN [29] utilized both local and non-local architectures to bias the most informative feature components. Meta-SR [30] proposed by Hu *et al.* firstly solved the problem of arbitrary scale factor super-resolution within a single model.

The aforementioned methods aim to achieve high PSNR and SSIM [31] values. However, these criteria usually causes heavy over-smoothed edges and artifacts. Images generated by these MSE-based SR methods lose various high-frequency details and have a bad perceptual quality. To generate more photo-realistic images, Ledig et al. firstly introduced generative adversarial network into image super-resolution, called SRGAN [11]. SRGAN combined a perceptual loss and an adversarial loss to improve the reality of generated images. But some visually implausible artifacts still could be found in some generated images. To reduce the artifacts, EnhanceNet [32] combined a pixel-wise loss in the image space, a perceptual

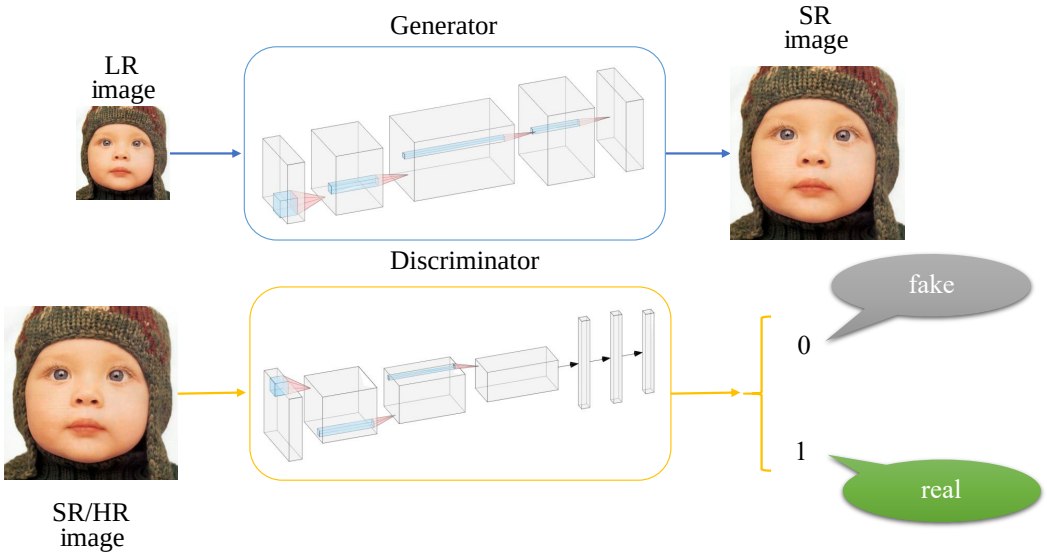

Fig. 1. The architecture of GAN-based Super-Resolution method. The generator aims to reconstruct a SR image similar to the HR image as good as possible, while the discriminator distinguishes the SR image from HR image and conveys a supervisor to the generator.

loss in the feature space, a texture matching loss [33] and an adversarial loss. The texture matching loss helped to generate more realistic textures. Yan *et al.* [13] firstly trained a novel full-reference image quality assessment (FR-IQA) approach for SISR, then employed the proposed loss function (SR-IQA) to train their SR network which contains their proposed highway unit. In addition, they also integrate SR-IQA loss to the GAN-based SR method to achieve better results for both accuracy and perceptual quality. Dahl *et al.* [34] proposed a pixel recursive super resolution model, an extension of PixelCNNs [35, 36], to reconstruct face super-resolution images. The contextual loss [37] was a kind of perceptual loss to make the generated images as similar as possible to ground truth images. Cheon *et al.* [38] creatively utilized DCT transformation to make the generated images closer to the ground truth images in the frequency domain, reducing blurry-edge effects due to pixel loss. Based on SRGAN, ESRGAN [14] *i*) substituted the standard residual block with a residual-in-residual dense block, *ii*) removed batch normalization layers, *iii*) utilized VGG feature before activated as perceptual loss, and *iv*) replaced the standard discriminator with Relativistic Discriminator proposed in RaGAN [39]. In addition, ESRGAN used network interpolation to balance the MSE loss and perceptual quality. Noteworthily, ESRGAN won the first place in the 2018 PIRM Challenge on Perceptual Image Super-Resolution [40], which pursued the high perceptual-quality images.

III. Proposed Methods

A. Overview

Our methods aim to reconstruct a high-resolution image $I^{\text{SR}} \in R^{Wr \times Hr \times C}$ from a low-resolution image $I^{\text{LR}} \in R^{W \times H \times C}$, where W and H are the width and height of the LR image, r is the upscaling factor, and C is the number of channels of the color space. This section details our two strategies within the GAN framework for image super-resolution in order: FASRGAN and Fs-SRGAN. We propose a fine-grained attention in FASRGAN to make the generator focus on the difficult parts of image reconstruction instead of treating every part equally. At the same time, we propose a feature-sharing mechanism in Fs-SRGAN by sharing the shallow feature extraction of the generator and the discriminator. These two strategies contribute to the overall perceptual quality for SR. For simplicity, we use the same network architecture as ESRGAN [14] for the generator.

B. Fine-grained Attention Generator Adversarial Networks

Our proposed fine-grained attention GAN (FASRGAN) designs a specific discriminator for SISR. As discussed above and shown in Fig.1, the discriminator in a standard GAN-based SR model outputs a score of the whole input SR/HR image. This can be considered as a coarse way to judge an input image and cannot discriminate local features of inputs. To tackle this problem, the proposed FASRGAN defines a Unet-like discriminator contained two outputs, corresponding to a score of the whole image and a fine-grained score map, respectively. The score map has the same spatial size as the input image and is used for pixel-wise discrimination. The proposed discriminator is illustrated in Fig. 2.

1) A Unet-like Discriminator: The Unet-like discriminator with two outputs can be divided into two parts: an encoder and a decoder.

Encoder. Similar to the standard discriminator D in SRGAN, the encoder part of the proposed Unet-like discriminator uses a standard max-pooling layer with a stride of 2 to reduce the spatial size of a feature map and increase receptive fields. At the same time, the number of channels is increased for improving representative ability. At the end of the encoder, two fully connected layers are added to output a score, measuring the overall probability of an input image x being real or fake. We further enhance the discriminator based on the Relativistic

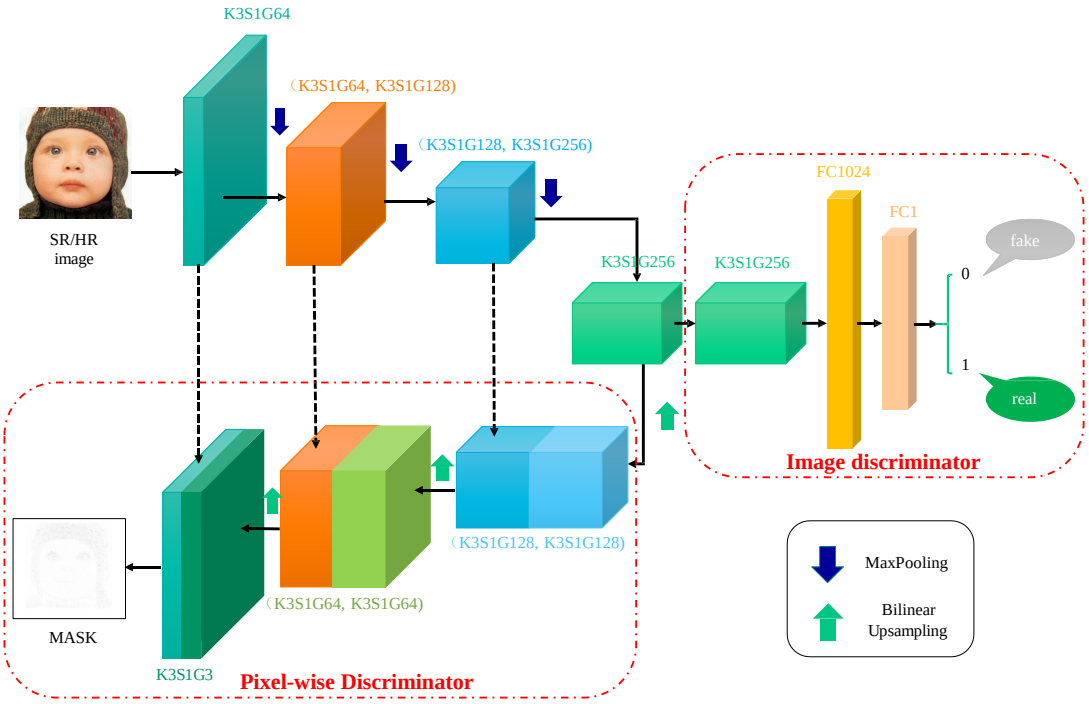

Fig. 2. The discriminator architecture of FASRGAN, where K, S, G represent the kernel size, the stride, and the kernel number of the Conv layer, respectively. FC stands for fully connected layer. The mask is a score map among $[0, 1]$, donating the difficulty of reconstruction of each pixel in the image.

GAN [39], which has also been used in ESRGAN [14]. The loss function is defined as:

$$\begin{aligned} L_{adv}^{D} =& \mathbb{E}_{x_r}[\log(1 - D_{Ra}(x_r, x_f))] \\ &+ \mathbb{E}_{x_f}[\log(D_{Ra}(x_f, x_r))] \\ =& \mathbb{E}_{x_r}[\log(1 - \sigma(C(x_r) - \mathbb{E}_{x_f}[C(x_f)]))] \\ &+ \mathbb{E}_{x_f}[\log(\sigma(C(x_f) - \mathbb{E}_{x_r}[C(x_r)]))], \end{aligned} \tag{1}$$

where x_r and x_f stand for the ground truth image and the generated SR image, respectively. $D_{Ra}(\cdot)$ refers to the function of the relativistic discriminator, which tries to predict the probability that a real image x_r is more realistic than a fake one x_f; $C(x)$ is the discriminator output before sigmoid function and σ is the sigmoid function.

Decoder. We exploit an upsampling layer to extend the spatial size of feature maps as shown in Fig. 2. To make full use of features, we concatenate the previous outputs, which have the same spatial size as current ones. As shown in Fig. 2, the feature maps at the end of the decoder have the same spatial size as input images. Finally, we use the sigmoid function to produce a score map $M \in R^{Wr \times Hr \times C}$ that provides pixel-wise discrimination between real and fake pixels of an input image. The loss function is defined as:

$$\begin{aligned} L_{M}^{D} =& \frac{1}{Wr \times Hr \times C} \\ &\times \sum_{c=1}^{C} \sum_{w=1}^{Wr} \sum_{h=1}^{Hr} \{\log(1 - M_r(w, h, c)) + \log(M_f(w, h, c))\}, \end{aligned} \tag{2}$$

where M_r and M_f represent the score maps of the HR image and the generated SR image, respectively.

2) Fine-grained Attention Loss Function: The score map generated by the Unet-like discriminator is pixel-wise discrimination scores of an input image, with values $M(w, h, c)$ among $[0, 1]$. The higher score the closer to a real image. In this manner, the score map can indicate which parts of an image are more difficult to generate and which parts are easier. For instance, the structure background part of an image is sometimes simpler, and thus it would expect the discriminator reflects this to the generator when updating the generator. In other words, the part with lower scores (more difficult to generate) should receive more attention when updating the generator. As a result, we incorporate the score map as the fine-grained attention mechanism into the L_1 loss function that is defined as:

$$\begin{aligned} L_1 =& \frac{1}{Wr \times Hr \times C} \\ &\times \sum_{w=1}^{Wr} \sum_{h=1}^{Hr} \sum_{c=1}^{C} \parallel F_{\theta}^{G}(I_i^{LR})(w, h, c) - I_i^{HR}(w, h, c) \parallel_1, \end{aligned} \tag{3}$$

where $F_{\theta}^{G}(\cdot)$ represents the function of the generator, θ is the parameters of the generator and I_i means the i-th image. This function treats every position in the image equally. Our proposed fine-grained attention loss function is the following weighted L_1 function:

$$\begin{aligned} L_{attention} =& \frac{1}{Wr \times Hr \times C} \sum_{w=1}^{Wr} \sum_{h=1}^{Hr} \sum_{c=1}^{C} (1 - M_f(w, h, c)) \\ &\times \parallel F_{\theta}^{G}(I_i^{LR})(w, h, c) - I_i^{HR}(w, h, c) \parallel_1, \end{aligned} \tag{4}$$

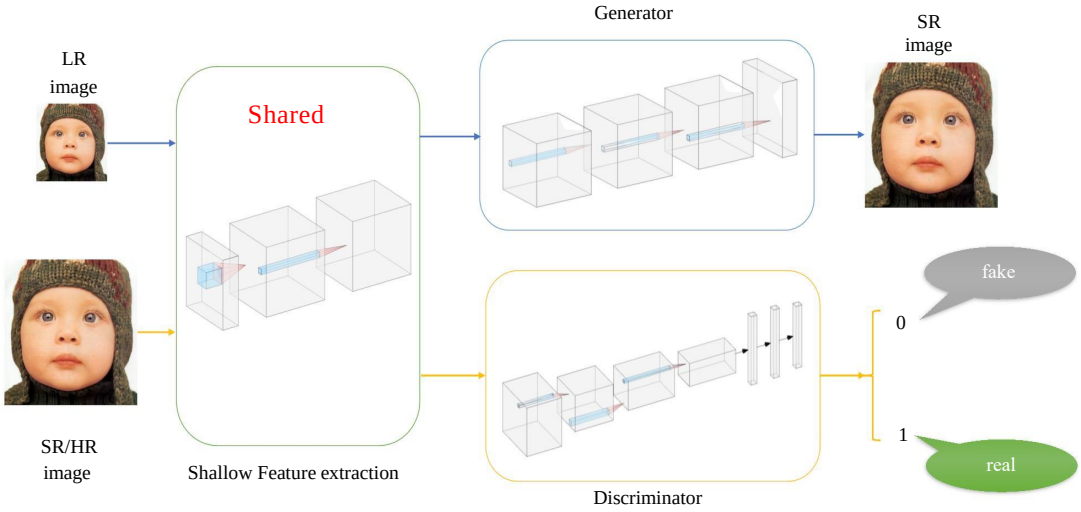

Fig. 3. The illustration of our Feature-sharing Generator Adversarial Networks (Fs-SRGAN). The input sizes of the generator and the discriminator are different. We use a fully Convolutional Neural Network with invariant size of the feature map so that the different input sizes do not matter.

where $M_f(w,h,c)$ is the score map of the generated image given by the discriminator.

3) Loss Function of the Generator: Our fine-grained attention mechanism is general and can be applied to various GAN-based methods. In this paper, we use the stack of Residual-in-Residual Dense Blocks (RRDBs), the basic building block of ESRGAN [14], to define our generator. Our generator consists of several losses, described below:

Content Loss. Following [6, 9, 22, 23], we use an L_1 loss function to constrain the content of a generated SR image to be close to the HR image. The loss is defined in Eq. 3.

Perceptual Loss. The perceptual loss [41] aims to make the SR image close to the corresponding HR image based on high-level features extracted from a pre-trained network. We consider both the SR and HR images as the input to the pre-trained VGG19 and extract the VGG19-54 layer features. The perceptual loss is defined as:

$$L_{percep} = \| F_\theta^{VGG}(G(I_i^{LR})) - F_\theta^{VGG}(I_i^{HR}) \|_1, \quad (5)$$

where $F_\theta^{VGG}(\cdot)$ is the function of VGG and I_i is the i-th image, $G(\cdot)$ is the function of the generator.

Adversarial Loss. The discriminator contains two outputs, a whole estimation of the entire image and the pixel-wise fine-grained estimations of an input image. The total adversarial loss function for the generator is defined as:

$$L_{adv}^G = L_{entire}^G + L_{fine}^G, \quad (6)$$

As shown in Eq. 2, the discriminator tries to distinguish the real and fake image in a fine-grained way, while the generator aims to fool the discriminator. Thus the loss function for the fine-grained attention loss of generator is the symmetrical form of Eq. 2:

$$L_{fine}^G = \frac{1}{Wr \times Hr \times C} \times \sum_{c=1}^{C}\sum_{w=1}^{Wr}\sum_{h=1}^{Hr}\{\log(M_r(w,h,c)) + \log(1 - M_f(w,h,c))\}, \quad (7)$$

L_{entire}^G is also the symmetrical form of Eq.1 and defined as:

$$\begin{aligned} L_{entire}^G =& \mathbb{E}[\log(\sigma(C(x_r) - \mathbb{E}[C(x_f)]))] \\ &+ \mathbb{E}[\log(1 - \sigma(C(x_f) - \mathbb{E}[C(x_r)]))], \end{aligned} \quad (8)$$

Combining the above losses and the attention loss, the total loss of the generator is:

$$L^G = L_1 + \lambda_1 L_{adv}^G + \lambda_2 L_{attention} + \lambda_3 L_{percep}, \quad (9)$$

where λ_1, λ_2, λ_3 are the coefficients to balance different loss terms.

C. Feature-sharing Generator Adversarial Networks

In the standard GANs, the generator and the discriminator are usually defined as two independent networks. In our problem, we observe that the shallow parts of these two networks always extract local textures such as edges and corners. To reflect this, we propose a new network structure (Fs-SRGAN) to share the shallow-feature-exaction parts of the generator and the discriminator. Consequently, our Fs-SRGAN contains three parts: a shared feature extractor, a generator, and a discriminator, as shown in Fig. 3.

1) Shared Feature Extractor: The feature-sharing mechanism allows the generator and the discriminator to jointly optimize the shallow feature extractor. Similar to FASRGAN, we adopt RRDB, the basic block of ESRGAN [14], as the basic structure. The shared feature extractor contains E RRDBs to extract helpful feature maps for both the generator and the

discriminator, described as following:

$$H_{shared} = F_{shared}(x), \tag{10}$$

where H_{shared} is the shallow shared feature maps extracted by the shared part, F_{shared} represents the function of the shared feature extractor, and x is the input. For the generator, the input is an LR image, while for the discriminator it is the SR image or the HR image. The input sizes of the generator and the discriminator are different. We apply a fully Convolutional Neural Network with invariant size of feature map to extract features so that the different input sizes do not matter.

2) The Generator and the Discriminator: The rest parts of the generator and the discriminator are the same as those in standard GAN-based methods, except that the inputs are feature maps instead of images as shown in Fig.3.

Generator. The generator in SR generally contains three parts: shallow feature extraction, deep feature extraction, and reconstruction. Similar to the shared shallow feature extraction, we adopt RRDB as the basic part of deep feature extraction, except that more RRDBs are used to increase the depth of the network with the purpose of extracting more high-frequency feature for reconstruction. The reconstruction part utilizes an upsampling layer to upscale the feature maps and a Conv layer to reconstruct an SR image. The loss function of the generator includes adversarial loss, pixel-based loss, and perceptual loss, similar to Eq.9, except there is no attention loss $L_{attention}$.

Discriminator. Because the discriminator is a classification network that distinguishes the SR and HR image, we apply a structure similar to the VGG network [42] as the discriminator. To avoid information loss, we replace the pooling layer for a Conv layer with a stride of 2 to decrease the size of feature map. At the tail of the discriminator, we use a Conv layer to transform the feature map into a one-dimensional vector, then use two fully connected layers to output the classification score s among $[0, 1]$. The value of s closer to 1 means more real, otherwise more fake. The loss function of the discriminator is defined as follow:

$$L_D = [\log(1 - D(I^{HR}))] + \log(D(G(I^{LR}))), \tag{11}$$

where $D(\cdot)$ is the discriminator function and $G(\cdot)$ is the function of the generator.

IV. Experimental Results

In this section, we first describe our model training details, then provide quantitative and visual comparisons with several state-of-the-art methods on benchmark datasets for our two proposed novel methods, FASRGAN and Fs-SRGAN. We further combine the fine-grained attention and the feature-sharing mechanisms into one single model, termed FA+Fs-SRGAN.

A. Training Details

The DIV2K dataset was proposed in NTIRE 2017 Challenge on Single Image Super-Resolution [8] and widely used in previous SR methods, which contains 800, 100 and 100 images of 2K-resolution for training, validation, and testing, respectively. We use the training set from DIV2K dataset for training. The LR images are obtained by bicubic downsampling (BI) from the source high-resolution images. At testing, we also use five standard benchmark datasets: Set5 [43], Set14 [44], BSD100 [45], Urban100 [46], and Manga109 [47]. Blau *et al.*. [48] proved perceptual quality is not always improved with the increase of PSNR value and revealed the trade-off between the average distortion and perceptual quality. We adopt perceptual index (PI) and root mean square error (RMSE) as our quantitative measurements, where PI measures the perceptual quality of the SR image and RMSE measures the reconstruction loss between HR image and SR image. Both PI and RMSE with lower values mean better results.

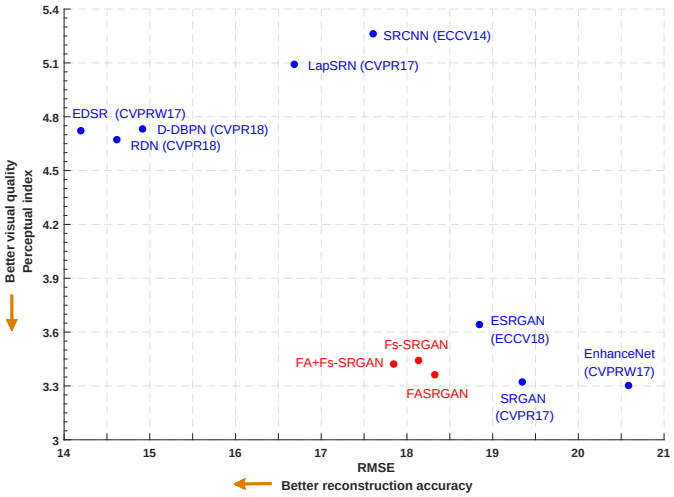

Fig. 4. The trade-off of RMSE and PI on Urban100 of our methods and the state-of-the-art methods for $4\times$ super-resolution.

In training, images are augmented by rotating and flipping. The batch size is set to 16. Our methods are trained based on image patches and optimized with the ADAM optimizer [49]. The hyperparameters β_1 and β_2 in the ADAM optimizer are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We randomly crop 48×48 patches from LR images as the input of the network. The generator is pre-trained by the L_1 loss function. Following [6, 9, 20, 22, 29], the initial learning rate is set to 1×10^{-4}, and then decays to half every 2×10^5 iterations. In Fs-SRGAN, we set the number of RRDBs in the shared feature extractor as $E = 1$. We implement our models with the PyTorch [50] framework on a Titan Xp GPU.

B. Quantitative Comparisons

We first present the quantitative comparisons between our methods and the state-of-the-art methods. The results are shown in Fig. 4. These methods can be roughly divided into two categories: the top-left and the bottom-right. Methods in the top-left part are almost MSE-based with low RMSE loss and high PI scores due to the over-smoothed edges and lack of high-frequency details. The bottom-right category includes the GAN-based methods, such as SRGAN, EnhanceNet, ESRGAN, and our methods. These methods usually gain high-

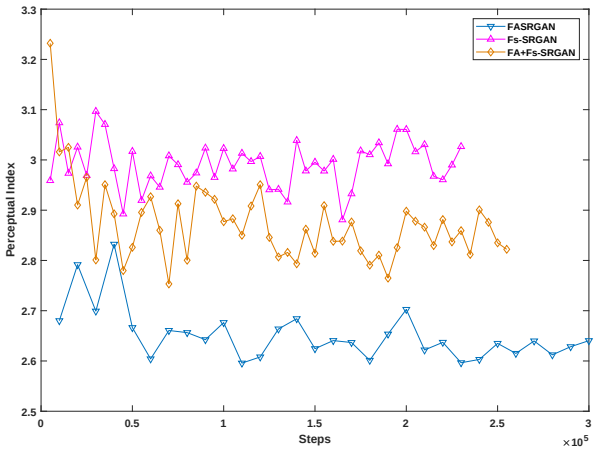

Fig. 5. The changes of average PI on Set14 during the training process for $4\times$ super-resolution.

visual quality images even if their RMSE losses are larger than those of the MSE-based methods. Among these methods, our FASRGAN and Fs-SRGAN get better visual quality and less reduction error. As shown in Fig. 4, the FA+Fs-SRGAN attains the best reconstruction accuracy among all the GAN-based methods. FA+Fs-SRGAN also achieves a lower PI value. Fig. 5 plots the curves of PI values in the training process of our proposed methods on Set14. We observe that the training process of FASRGAN is more stable and obtains better perceptual quality. The average PI value of Fs-SRGAN is higher than FASRGAN. As mentioned above, Fs-SRGAN contains fewer RRDBs than FASRGAN. We speculate that less RRDBs caused higher PI values. FA+Fs-SRGAN, which combines the fine-grained attention mechanism into Fs-SRGAN, obtains the lower PI values than Fs-SRGAN.

C. Qualitative Results

We compare our final models on several public benchmark datasets with the state-of-the-art MSE-based methods: SRCNN [1], FSRCNN [4], EDSR [6], SRMDNF [27], RDN [9], and GAN-based approaches: SRGAN [11], EnhanceNet [32], ESRGAN [14]. We conduct comparisons with our two methods respectively.

1) Visual Comparisons of FASRGAN: Some representative quality results are presented in Fig. 6. PSNR (evaluated on the luminance channel in YCbCr color space) and the perceptual index used in the 2018 PIRM-SR Challenge [40] are also provided for reference.

As shown in Fig. 6, our proposed FASRGAN outperforms previous methods by a large margin. Images generated by FASRGAN contain more fine-grained textures and details. For example, for image '0801' of DIV2K, MSE-based methods tend to generate blurry results, while results from SRGAN and EnhanceNet tend to be noisy; results from ESRGAN have blurry and over-smoothed edges. FASRGAN can produce sharper and more natural textures of the penguin beak. The cropped parts of image '0828' and 'YumeirCooking' are full of textures. As we can see, all the compared MSE-based methods suffer from heavy blurry artifacts, failing to recover the structure and the gap of the stripes. SRGAN, EnhanceNet, and ESRGAN generate high-frequency noise and wrong textures; while our FASRGAN can recover them more correctly, producing more faithful results and being closer to the HR images. For image 'img_093' in Urban100, the cropped part of the image generated by the compared methods contains heavily blurry artifacts and lines with wrong directions. By contrast, our FASRGAN can alleviate the artifacts better and recover zebra crossing with correct structures. These comparisons demonstrate the strong ability of FASRGAN for producing more photo-realistic and high-quality super-resolution images.

2) Visual Comparisons of Fs-SRGAN: We further compare our Fs-SRGAN with the state-of-the-art methods in Fig. 7. Obviously, our Fs-SRGAN obtains better performance than other methods in producing SR images, in terms of sharpness and details. For image 'baboon', the cropped parts of the image generated by the MSE-based methods are over-smoothed. Previous GAN-based methods not only fail to produce clear whiskers but also introduce lots of unpleasing noise. ESRGAN generates too many whiskers, which have not appeared in the original HR image. Our Fs-SRGAN produces more correct whiskers. For image '0812' and 'img_069', MSE-based methods still suffer from heavy blurry artifacts and generate unnatural results. GAN-based methods cannot maintain the structures of the stairs or the train tracks. Our proposed Fs-SRGAN outperforms the compared methods and produces closer images to the original HR images. For image '0879', our Fs-SRGAN can recover tiny textures of windows that look more natural, while previous methods still have difficulties to produce high-quality SR images. This also indicates that the shared shallow feature extractor of the generator and the discriminator is beneficial.

TABLE I
THE RESULT OF OBJECT RECOGNITION BETWEEN OUR METHODS AND THE STATE-OF-THE-ART METHODS. THE BASELINE USES THE ORIGINAL HR IMAGE AS THE INPUT OF RESNET-50 MODEL.

Evaluation	Top-1 error	Top-5 error
Bicubic	0.526	0.277
SRCNN [1]	0.464	0.230
FSRCNN [4]	0.488	0.252
SRGAN [11]	0.410	0.191
EnhanceNet [32]	0.454	0.224
Fs-SRGAN (ours)	0.338	0.136
FA+Fs-SRGAN (ours)	0.337	0.134
FASRGAN (ours)	**0.323**	**0.124**
Baseline	0.241	0.071

D. Object Recognition Performance

To further demonstrate the quality of our generated SR images, we treat them as a pre-processing step for other high-level computer vision tasks such as object recognition, image classification and so on. In this section, we use the same setting as EnhanceNet and evaluate the object recognition performance with the generated images by our methods and

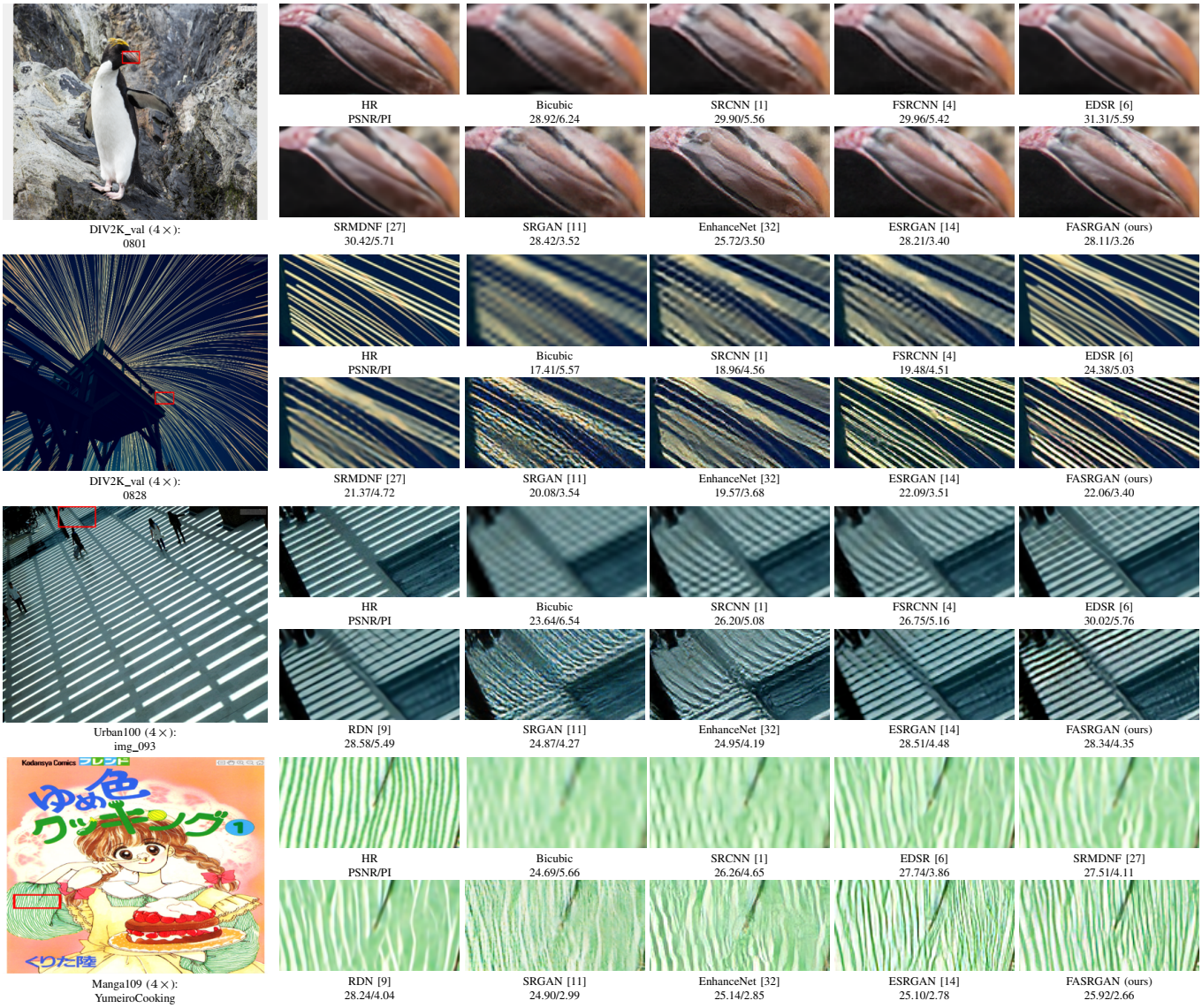

Fig. 6. The visual comparisons between FASRGAN and the state-of-the-art SR methods for $4\times$ super-resolution.

other state-of-the-art methods: SRCNN [1], FSRCNN [4], SRGAN [11], EnhanceNet [32].

We use the pre-trained ResNet-50 on imageNet as an evaluation model and fetch the first 1000 images in ImageNet CLS-LOC validation dataset for evaluation. The test images are first down-sampled by bicubic and then upscaled by our methods and the compared methods. These SR images are then used as inputs to the ResNet-50 model to calculate their Top-1 and Top-5 errors for evaluation. As shown in Table I, both two methods we proposed and the variant FA+Fs-SRGAN achieve better accuracy compared to the state-of-the-art methods. Among these three methods, FASRGAN achieves the lowest Top-1 and Top-5 errors, demonstrating the effectiveness of both the fine-grained attention and the feature-sharing mechanisms.

To further illustrate the effectiveness of our models, we use another pre-trained model on imageNet, called vgg-19. The experiment setting is the same as that of ResNet-50, and the results are shown in Table II. Among the SR methods, our models still obtain better results, and Fs-SRGAN obtains the lowest Top-1 and Top-5 errors.

TABLE II
THE RESULT OF OBJECT RECOGNITION BETWEEN OUR METHODS AND THE STATE-OF-THE-ART METHODS. THE BASELINE USES THE ORIGINAL HR IMAGE AS THE INPUT OF VGG-19 MODEL.

Evaluation	Top-1 error	Top-5 error
Bicubic	0.617	0.351
SRCNN [1]	0.549	0.324
FSRCNN [4]	0.597	0.361
SRGAN [11]	0.539	0.29.7
EnhanceNet [32]	0.546	0.286
Fs-SRGAN (ours)	**0.407**	**0.168**
FA+Fs-SRGAN (ours)	0.434	0.184
FASRGAN (ours)	0.416	0.173
Baseline	0.285	0.093

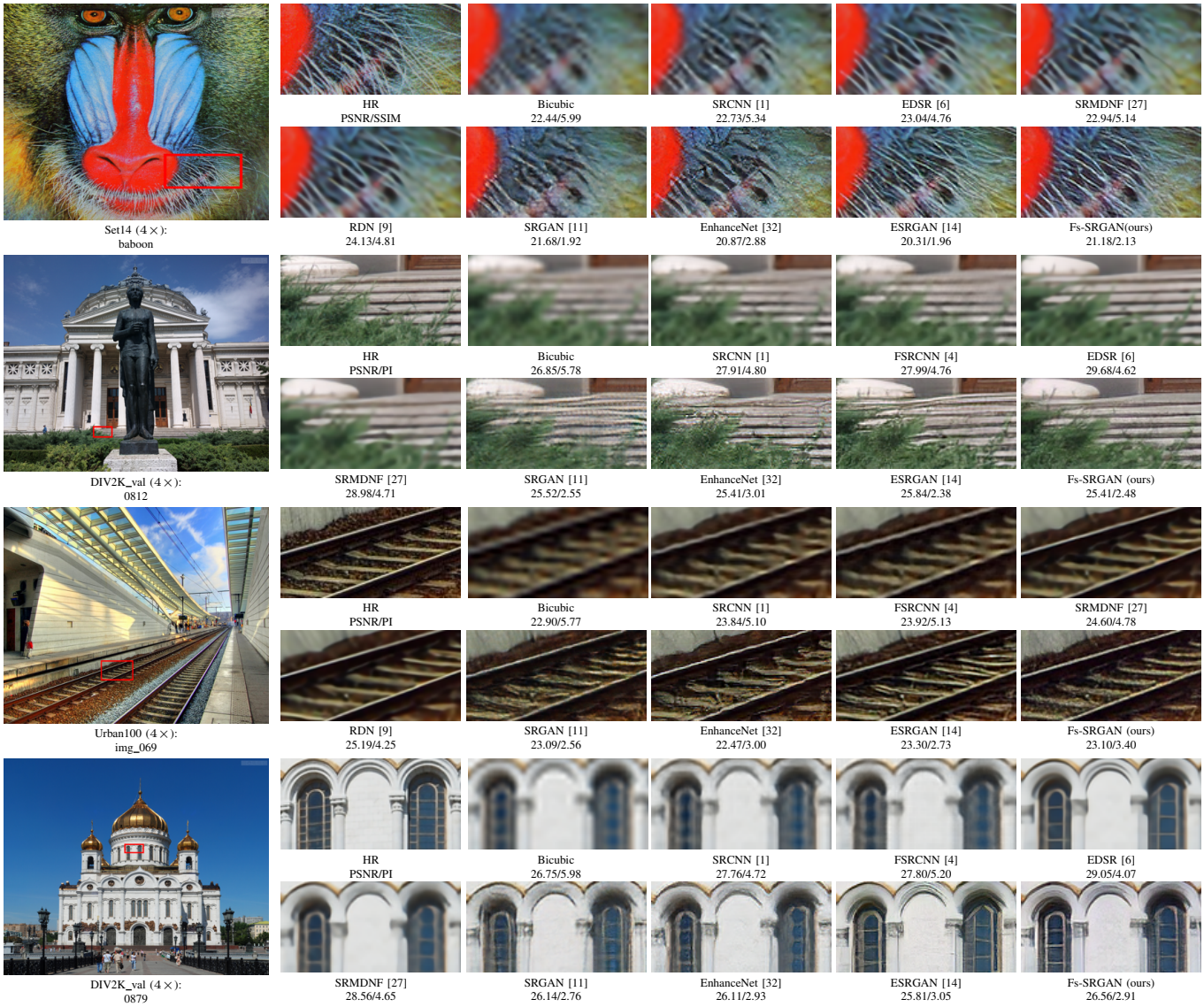

Fig. 7. The visual comparisons between Fs-SRGAN and the state-of-the-art SR methods for $4\times$.

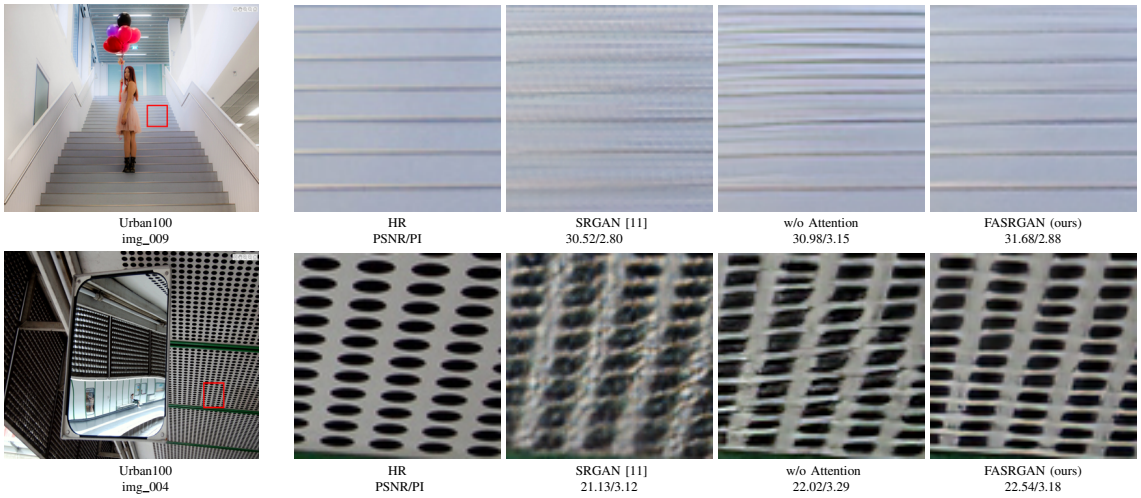

Fig. 8. The visual result of ablation study of FASRGAN.

Fig. 9. The visual result of ablation study of Fs-SRGAN.

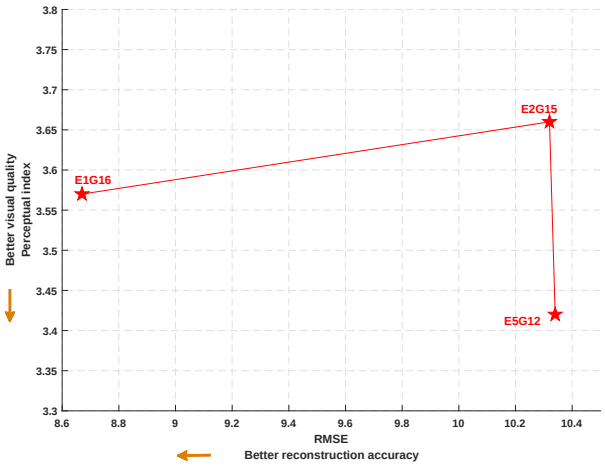

Fig. 10. The change of the number of RRDBs in shared shallow feature extractor (E) in Fs-SRGAN. G represents the number of RRDBs in deep feature extraction part.

E. Ablation Study

In order to study the effects of the two mechanisms in the proposed methods, we conduct ablation experiments by removing the mechanisms and test the differences, respectively. The overall visual comparisons are illustrated in Fig. 8 and Fig. 9. A detailed discussion is provided as follows.

a) Removing the Fine-grained Attention Mechanism: We first remove the fine-grained attention (FA) mechanism in the FASRGAN. An obvious performance decrease can be observed in Fig. 8. For image 'img_009', the model without FA mechanism introduces some unnatural noise and undesired edges, while FASRGAN can maintain the structure and produce high-quality SR images. For image 'img_004', the FA mechanism can improve the deformation appeared in the image from the model that removes the FA mechanism. The images generated by FASRGAN are closer to the original HR images. The visual analysis indicates the effectiveness and benefit of the FA mechanism in removing unpleasant and unnatural artifacts.

b) Removing the Feature-sharing Mechanism: Fig. 9 shows the results of removing the feature-sharing (Fs) mechanism and use two independent networks as the generator and the discriminator. We can observe that Fs-SRGAN outperforms SRGAN and the model without Fs mechanism by a large margin. The removal of feature-sharing mechanism tends to introduce unpleasant artifacts. For image 'zebra', by employing the Fs mechanism, Fs-SRGAN can alleviate heavy artifacts and noises. For image 'img_083', characters in the cropped image generated by Fs-SRGAN are clearer and more recognizable due to the benefit of the Fs mechanism.

To further study the effect of the depth of the shared feature extractor in Fs-SRGAN, we vary the number of RRDBs in both the shared shallow feature extractor and the deep feature extractor. As shown in Fig. 10, increasing the number of the shared part from 1 to 2 decreases the performance, manifested in the reconstruction accuracy. However, when the number is increased to 5, the model E5G12 performs better in visual qualities.

V. Conclusion

We propose two GAN-based models, FASRGAN and Fs-SRGAN, for SISR to overcome the limitations of existing methods. FASRAGN introduces a fine-grained attention mechanism into the GAN framework, where the discriminator has two outputs to measure quality of the overall input, as well as a fine-grained attention estimation for the input. The fine-grained attention delivers a fine-grained supervisor to the generator to ensure generation of pixel-wise photo-realistic images. The Fs-SRGAN shares the shallow feature extractor of the generator and the discriminator, reducing the number of parameters and improving the reconstruction performance. These two mechanisms are general and could be applied to other GAN-based SR models. Comparisons with other state-of-the-art methods on benchmark datasets demonstrate the effectiveness of our proposed methods.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.

[2] N. Kumar and A. Sethi, "Fast learning-based single image super-resolution," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1504–1515, Aug 2016.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1646–1654.

[4] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*. Springer, 2016, pp. 391–407.

[5] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. Wei, "Drfn: Deep recurrent fusion network for single-image super-resolution with large factors," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 328–337, Feb 2019.

[6] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 114–125.

[9] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[13] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, "Deep objective quality assessment driven single image super-resolution," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[14] 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 *European Conference on Computer Vision*. Springer, 2018, pp. 63–79.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[17] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2017, pp. 3147–3155.

[18] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks." in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 10, 2010, p. 7.

[19] M. D. Zeiler, G. W. Taylor, R. Fergus *et al.*, "Adaptive deconvolutional networks for mid and high level feature learning." in *ICCV*, vol. 1, no. 2, 2011, p. 6.

[20] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[21] J.-H. Kim and J.-S. Lee, "Deep residual network with enhanced upscaling module for super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 800–808.

[22] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5835–5843.

[23] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, Aug 2018.

[24] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[25] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.

[26] A. Shocher, N. Cohen, and M. Irani, "'Zero-shot' super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.

[27] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.

[28] C. Liu, X. Sun, C. Chen, P. L. Rosin, Y. Yan, L. Jin,

and X. Peng, "Multi-scale residual hierarchical dense networks for single image super-resolution," *IEEE Access*, vol. 7, pp. 60 572–60 583, 2019.

[29] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *ICLR*, 2019.

[30] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1575–1584.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.

[33] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[34] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5439–5448.

[35] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[36] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[37] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor, "Maintaining natural image statistics with the contextual loss," *arXiv preprint arXiv:1803.04626*, pp. 1–16, 2018.

[38] M. Cheon, J.-H. Kim, J.-H. Choi, and J.-S. Lee, "Generative adversarial network-based image super-resolution using perceptual content losses," in *European Conference on Computer Vision*. Springer, 2018, pp. 51–62.

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

[40] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *European Conference on Computer Vision*. Springer, 2018, pp. 334–355.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[43] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, pp. 135.1–135.10.

[44] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proceedings of the 7th International Conference on Curves and Surfaces*. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 711–730.

[45] D. Martin, C. Fowlkes, D. Tal, J. Malik *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[46] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[47] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21 811–21 838, 2017.

[48] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. Adv. Neural Inf. Process. Syst. Workshop Autodiff, Dec.*, 2017, pp. 1–4.