Multi Recursive Residual Dense Attention GAN for Perceptual Image Super Resolution

Linlin Yang, Hongying Liu, Yiming Li, Wenhao Zhou, Yuanyuan Liu, Xiaobiao Di, Lei Wang, Chuanwen Li

Multi recursive Residual Dense Attention GAN for Perceptual Image Super Resolution

Linlin Yang[1], Hongying Liu[1], Yiming Li[1], Wenhao Zhou[1], Yuanyuan Liu[1], Xiaobiao Di[2], Lei Wang[2], and Chuanwen Li[2]

[1] Key Laboratory of Intelligent Perception and Image Understanding, School of Artificial Intelligence, Xidian University, China
hyliu@xidian.edu.cn
[2] China Petroleum Pipeline Telecom & Electricity Engineering, Co., Ltd.

Abstract. Single image super-resolution (SISR) has achieved great progress based on convolutional neural networks (CNNs) such as generative adversarial network (GAN). However, most deep learning architectures cannot utilize the hierarchical features in original low-resolution images, which may result in the loss of image details. To recover visually high-quality high-resolution images, we propose a novel Multi-recursive residual dense Attention Generative Adversarial Network (MAGAN). Our MAGAN enjoys the ability to learn more texture details and overcome the weakness of conventional GAN-based models, which easily generate redundant information. In particular, we design a new multi-recursive residual dense network as a module in our generator to take advantage of the information from hierarchical features. We also introduce a multi-attention mechanism to our MAGAN to capture more informative features. Moreover, we present a new convolutional block in our discriminator by utilizing switchable normalization and spectral normalization to stabilize the training and accelerate convergence. Experimental results on benchmark datasets indicate that MAGAN yields finer texture details and does not produce redundant information in comparison with existing methods.

Keywords: Image super-resolution · Generative adversarial networks · Multi-recursive residual dense network · Attention mechanism.

1 Introduction

Single image super-resolution (SR) is a fundamental low-level vision task in computer vision, which aims to recover a high-resolution (HR) image from a single low-resolution (LR) one via SR methods. SR is also a research hotspot in computer vision, and recently attracts increasing attention in image restoration applications. Especially, SR has been popular in various applications [1] such as medical imaging, image generation, security, and surveillance systems. In fact, SR is an ill-posed inverse problem because there are a large number of solutions for restoration from LR images to HR images. To deal with this issue, a great number of SR methods have been proposed, and they mainly can be categorized as the methods based on reconstruction [2], interpolation [3], and

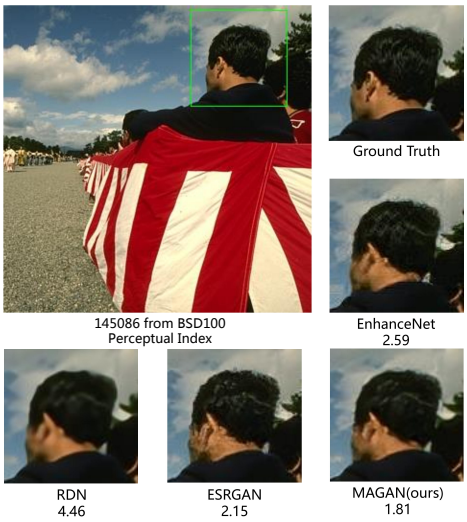


Fig. 1. Comparison of experimental results (i.e., Perceptual Index, PI) of the existing methods and our MAGAN method for image SR tasks with a scale factor of $4\times$. Note that a lower value of PI indicates better perceptual recovery quality.

learning [4,5,6], respectively. Since the pioneering work, SRCNN [7], was proposed, many deep learning based methods such as convolutional deep neural networks (CNNs) have brought about great progress in SR tasks. A variety of architectures and training approaches have continually enhanced the performance of SR in terms of some evaluation metrics.

Deep learning based SR algorithms generally fall into the following two classes: one is based on CNNs by utilizing classical L1- or L2-norm regularization at pixel level as a minimization loss function term, which can usually lead to higher PSNR performance but may over-smooth since it lacks high-frequency details. Some typical methods include EDSR [8], SRResNet [9], RDN [10], and [11,12]. The second category is the perceptual loss based approaches such as SRGAN [13], EnhanceNet [14], ESRGAN [15], and NatSR [16], which aim to make the SR result better accordant with human perception. In these methods, the generative adversarial network (GAN) [17] was introduced for SISR tasks. By using the alternating training between the discriminator and generator, GAN encourages the networks to tend to output results, which are visually more like real images. And they utilized perceptual loss as in [18] to optimize the SR model at a feature level. In [19], the semantic prior knowledge in images is further included to enhance the details of reconstructed texture. With these techniques, the perceptual loss based methods significantly improve the visual quality of the restored SR images, compared with those of the PSNR-oriented methods. However, the objective evaluation is still not satisfied to some extent, and the visual quality can be further improved.

To restore high-resolution images with more detailed textures, this paper proposes a novel Multi-recursive residual dense Attention GAN (MAGAN) for image SR tasks. The comparison of the experimental results of existing methods and MAGAN is demon-

strated in Fig. 1. It is clear that our MAGAN generates finer texture details and more realistic visual effects than other methods. We first construct a novel deep GAN, which also includes one discriminator and one generator, as shown in Fig. 2. We also design a new multi-recursive residual dense network (MRDN) for our generator to fully use hierarchical features from the original low-resolution image. Moreover, unlike the conventional GAN-based SR networks, as shown in Table 1, we present a new multi-attention mechanism for our discriminator to further discriminate refined features. Moreover, we present a new convolutional attention block, which consists of convolutional blocks (CBs) and an attention module, i.e., the convolutional block attention module (CBAM). To stabilize training and accelerate convergence, our CBs adopt spectral normalization and switchable normalization. In particular, our CBAM can capture more features by using both channel attention and spatial attention sub-modules. With such an architecture, the discriminator can learn to determine whether the restored image is more actual than the other, and the generator helps to restore more realistic texture details. To the best of our knowledge, it is the first GAN-based network that designs a multi-recursive structure and introduces a multi-attention mechanism to the discriminator.

The main contributions are summarized as follows:

- We propose a new deep GAN-based network (called MAGAN) to recover visually high-quality high-resolution images, which has a novel multi-recursive structure.
- We present the efficiency of the multi-recursive residual dense network, which function as the generator of our MAGAN, by extracting the hierarchical features from original LR images. We also introduce a multi-attention mechanism into our discriminator to extract more refined features.
- Finally, we design the new convolutional blocks for our discriminator, which applies both switchable normalization and spectral normalization to the proposed convolutional blocks. They can help to stabilize the training of the proposed network. Many experimental results show that MAGAN yields finer texture details than state-of-the-art methods.

2 Background

Generative adversarial network (GAN), which was proposed by [17], consists of a generator and a discriminator and has wide applications in a variety of areas, such as image generation, image to image translation, image completion, and image SR. Especially, SRGAN [13] was proposed for image SR, where both a discriminator and a generator were defined. One can optimize between them to solve the adversarial min-max problem in an alternating manner. The purpose of the generator is to yield a realistic image and try its best to fool the discriminator. In contrast, the discriminator aims to differentiate between the ground truth and the super-resolved images. Thus, the discriminator and the generator come into a game. With an alternating training way, the real images and the fake ones can finally follow a similar distribution statistically. [14] proposed the EnhanceNet, which also applied a GAN and introduced an additional local texture matching loss. Thus, EnhanceNet can reduce visually unpleasant artifacts. In the [15], the authors presented a perceptual loss that was posed on features before activation,

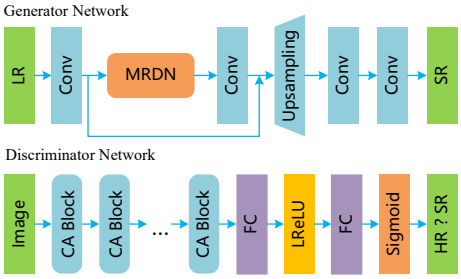


Fig. 2. The architecture of our MAGAN with its generator (top) and discriminator (bottom).

and relativistic GAN [20] was used in this work. As it is known, in GANs, the general discriminator can judge whether the input image is real and natural. In relativistic GAN [20], the discriminator attempts to calculate a probability indicating that a real image is relatively more realistic than a fake one.

The perceptual loss has been proposed by [18] and aims to make the SR result better accordant with human perception. Note that the perceptual loss can be computed by using high-level features extracted from pre-trained networks (e.g., VGG16 and VGG19) for the tasks of style transfer and SR. Previously, the perceptual loss function was defined on the activation layers of a deep pre-trained network, where the distance between two activated features requires to be minimized. Moreover, [21,22] proposed the contextual loss, which is based on natural image statistics and is used in training objective function. The algorithms in [21,22] can achieve better visual performance and perceptual image quality, but they are unable to yield superior results in terms of some objective evaluation criteria. There are other deep neural networks for image SR, for a more comprehensive review of those techniques, please refer to [1,23].

3 Proposed Methods

In this section, we design MAGAN, which mainly includes a new generator and a new discriminator, as shown in Fig. 2. Our MAGAN is expected to improve the overall perceptual quality for image SR tasks. The goal of SISR is to recover a SR image I^{SR} from a low-resolution input image I^{LR}. Note that I^{HR} denotes the high-resolution counterpart of a low-resolution image I^{LR}.

3.1 Our Multi-recursive Residual Dense Network in the Generator

In our MAGAN, we design a new multi-recursive residual dense network (MRDN) as the main module for the generator, as shown in Fig. 3. Our MRDN module combines the proposed multi-recursive residual network and dense connections. Considering the common observation that more network layers and connections can usually enhance real-world performance, our MRDN module is designed as deeper and more complex network. More specifically, as shown in Fig. 3, our MRDN module has a deep residual

Table 1. Comparison of the architectures in these GAN-based methods. Here, Conv denotes convolution, BN is batch normalization, SpectN is spectral normalization, and SwitN is switchable normalization.

Methods	Generator	Discriminator
SRGAN [13]	Residual	Conv, BN
EnhanceNet [14]	Residual	Conv
ESRGAN [15]	Residual dense	Conv, BN
NatSR [16]	Residual dense	Conv, Maxpool
MAGAN (ours)	Residual dense, Learnable, Multi-recursive	Conv, Multi-attention, SpectN & SwitN

learning structure, where residual structures are used in different layers. To expand the capacity of learning features, we use the dense block as the basic structure in our multi-recursive residual network, which mainly includes the multiple convolutional layers and the LeakyReLU activation function.

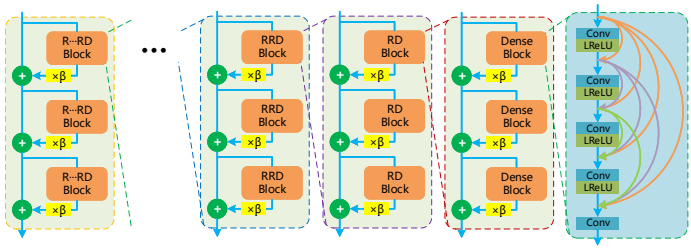


Fig. 3. The architecture of our proposed multi-recursive residual dense network (MRDN) used for the generator of MAGAN, where β is a learnable residual scaling parameter for each block, and $\oplus$ denotes element-wise addition.

Inspired by ReZero [24], we introduce a learnable parameter β into our MRDN module for modulating the non-trivial transformation of its each layer. Here, $F_{i,j}$ is defined as the output of the i-th cell in the j-th level of our MRDN module, where $i = 2, 3, ..., M$, and $j = 1, 2, ..., N$. Let (i, j) represent the i-th cell in the j-th level of our multi-recursive residual dense network, and $F_{(i-1),j}$ is the output of the $(i-1)$-th cell in the j-th level. The multi-recursive residual dense block can be expressed as follows:

$$F_{i,j} = F_{(i-1),j} + \beta \times H_{i,j}(F_{(i-1),j}), \tag{1}$$

where β is a learnable residual scaling parameter. $H_{i,j}()$ is the output of the i-th residual dense (RD) block in j-th level. And we set $\beta = 0$ at the beginning of training, i.e., initializing each layer to perform the identity operation.

In our GAN generator, the basic block is the proposed multi-recursive residual dense network, where most calculation operators were carried out in the feature space of low-resolution images. Our MAGAN with MRDN is shown in Fig. 2. Besides the residual

learning within MRDN, we also utilize global residual learning to obtain the feature-maps before up-sampling between the convolutional layers. Then the up-sampling layers can up-scale I^{LR} to attain an initial I^{SR}. Followed by two layers of convolutional operations, the generator outputs I^{SR}.

Moreover, as shown in Table 1, the MRDN module in the generator of our MAGAN is different from those of ESRGAN [15] and NatSR [16]. Though these three methods all use residual dense networks, our MRDN has a recursive structure that can learn multiple levels of the features.

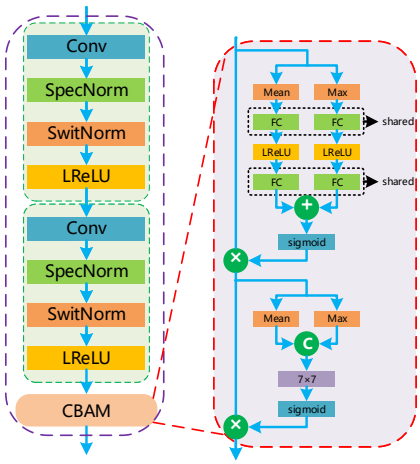


Fig. 4. The structure of the proposed convolutional attention (CA) block includes a sub-module of our convolutional blocks in the green dotted boxes and a convolutional block attention sub-module CBAM.

3.2 Our CA Network in the Discriminator

To further improve the performance of the discriminator in common GANs, we propose a new CA block, as shown in Fig. 4, which includes a sub-module of convolutional blocks and an attention sub-module. The details of them are given below.

Our attention sub-module aims to capture the fine structures of the images, and we employ the Convolutional Block Attention Module (CBAM), as shown in Fig. 4. The CBAM is a general attention module for feed-forward CNNs, which was proposed in [25] and is widely used in classification and recognition tasks. It consists of both one channel attention sub-module and one spatial attention sub-module. As illustrated in the red dotted box, CBAM utilizes both channel (top) and spatial-wise (bottom) attention. Here, the channel attention sub-module uses both mean-pooling (also called average-pooling) and max-pooling operations with a network with one shared layer (called FC), which is the multi-layer perceptron with a hidden layer. The spatial attention sub-module is composed of mean-pooling, max-pooling, concatenation, and convolutional operations. Note that $\otimes$ and $\oplus$ denote element-wise multiplication and addition

operations, ⓒ is the concatenation operation, and 7×7 is a convolution operation with a filter size of 7×7.

Suppose that the intermediate feature map is $\mathbf{F_{imi}} \in \mathbb{R}^{C\times H\times W}$ as input, CBAM sequentially infers the attention map $\mathbf{A_C} \in \mathbb{R}^{C\times1\times1}$ and the attention map $\mathbf{A_S} \in \mathbb{R}^{1\times H\times W}$, along the two dimensions, channel and spatial. The outputs of the channel attention and spatial attention sub-modules are formulated as follows:

$$\begin{aligned}\mathbf{F_{imt}} &= \mathbf{A_C}(\mathbf{F_{imi}}) \otimes \mathbf{F_{imi}},\\ \mathbf{F_{imo}} &= \mathbf{A_S}(\mathbf{F_{imt}}) \otimes \mathbf{F_{imt}},\end{aligned} \tag{2}$$

where $\otimes$ is element-wise multiplication operation, and $\mathbf{F_{imo}}$ denotes a final refined output. As shown in Fig. 4, our CBAM is a relatively lightweight module, and it is integrated into our attention convolutional module with negligible overheads, which is also end-to-end trainable.

Empirically, we find that adding the attention module to the discriminator of our MAGAN can attain more finer texture, while if the attention module is also added to the generator, this may cause severe texture blending. Since the attention module adopts pooling layers, it easily results in loss of location information. Therefore, we only apply the CBAM to the discriminator of MAGAN to improve SR performance. Thanks to the powerful ability to capture detailed information, relatively shallow neural networks (e.g., a multi-layer perceptron with one hidden layer) also have the ability to recover fine texture details.

Moreover, the proposed convolutional block is a basic one to construct the discriminator of our MAGAN. The detailed structure of our convolutional block is shown in the green dashed box in Fig. 4. Compared with the traditional convolutional structure, which uses batch normalization, our convolutional block is with both switchable normalization [26] and spectral normalization [27].

It is well known that normalization can stabilize the training in each iteration, improve SR performance and reduce computational cost in different PSNR-oriented tasks. However, the computational cost is greatly increased in many experiments, since it enhances the training time of each iteration. Thus, we borrow the idea of switchable normalization and spectral normalization to reduce the computational cost. Unlike the conventional GAN and SRGAN whose normalization is in its generator, we utilize the two types of normalization in our discriminator to stabilize and accelerate the training.

As GANs, our discriminator is also trained to discriminate whether real high-resolution images are more realistic than generated SR samples. Our discriminator contains several convolution layers, which have an increasing number of filters, starting from 64 filters in the first layer and then increasing by a factor of 2. Here strided convolutions are used to reduce the size and computation of each layer, and it doubles the number of map features. The resulting feature maps were followed by two dense layers and a sigmoid function to calculate the probability of image classification: HR or SR.

As it is indicated in the study [26], the performance of the network degrades when the batch size is one for most normalization, such as batch normalization. While the switchable normalization is not sensitive to the batch size. And it does not weaken the performance for small batch sizes compared with other normalizations. Moreover, the spectral normalization can constrain the Lipschitz constant of our discriminator by

restricting the spectral norm of each layer. Compared with other normalization techniques, spectral normalization does not need an additional hyper-parameter tuning.

It is well-known that the conditioning of a generator in various GANs is an important factor affecting the real-world performance of GANs [28]. Therefore, the generator in GANs can benefit from spectral normalization. In fact, we also find empirically that spectral normalization in both our generator and discriminator can make our MAGAN possible to use fewer discriminator iterations to update per generator, which significantly reduces the computational cost during training. With a comprehensive trade-off, we apply normalization in the discriminator of our MAGAN to stabilize and accelerate the training.

3.3 Loss Functions

Similar to other GAN-based networks, we alternately optimize the generator and discriminator of the proposed MAGAN until our model converges. First, the adversarial loss function of our generator is

$$L_{ad} = -\mathbb{E}_{I^{HR}}[1 - \log(D(I^{HR}, I^{SR}))] - \mathbb{E}_{I^{SR}}[\log(D(I^{SR}, I^{HR}))], \tag{3}$$

where $D(I^{HR}, I^{SR}) = \phi(H(I^{HR}) - \mathbb{E}_{I^{SR}}[H(I^{SR})])$, $\mathbb{E}_{I^{SR}}[\cdot]$ denotes average computation in the mini-batch for all fake images, $H(\cdot)$ is the output of discriminator, and ϕ denotes a sigmoid function.

The total loss function of the generator of our MAGAN is formulated as follows:

$$L = L_p + \gamma_1 L_{ad} + \gamma_2 L_1, \tag{4}$$

where L_p denotes the perceptual loss used in [29], $L_1 = \|I^{SR} - I^{HR}\|_1$ represents the content loss, γ_1 and γ_2 are two parameters to balance these loss function terms.

For our discriminator, its loss function can be formulated as follows:

$$L_D = -\mathbb{E}_{I^{HR}}[\log(D(I^{HR}, I^{SR}))] - \mathbb{E}_{I^{SR}}[1 - \log(D(I^{SR}, I^{HR}))]. \tag{5}$$

Fig. 5. The image SR visual results of our MAGAN without (middle, Non_MRDN) and with (right) MRDN on the Baboon image.

4 Experiments and Analysis

In this section, we conduct many experiments for image SR tasks to verify the effectiveness of our MAGAN method.

4.1 Implementation Details

In all the experiments, the scaling factor is $\times 4$ for the SR. In addition, we can downsample all the HR images to obtain LR images following [14,15], and the mini-batch size is set to 16. It is known that a larger receptive field can obtain more informative features from the image, and as in the work [15], we set the size of cropped high-resolution patches to 128×128.

Our training process of MAGAN includes the following two steps: pre-training and fine-tuning. The first step is pre-training, i.e., we train our model with the L1-norm regularized term. Specifically, an initial learning rate is set to 2×10^{-4}. And the pre-trained model was used as initialization to the generator for fine-tuning of our MAGAN. The generator is trained by using the weighted sum of the loss function from the generator and the perceptual loss with $\gamma_1 = 5 \times 10^{-3}$ and $\gamma_2 = 10^{-2}$. The learning rate here is 10^{-4}. Moreover, the discriminator is trained, where the LeakyReLU activation function [30] is used. We adopt eight convolutional layers with an increasing number of 3×3 filters. The size of the resulted feature maps is 512. We use the optimization algorithm, ADAM, to train our MAGAN. In addition, the generator and discriminator are alternately updated until they converge. Here, in the generator, we set $M = 4$ and $N = 3$. That is the basic recursive residual connection, which consists of 3 dense blocks, and we use 3 layers of recursive RD blocks considering the computational cost and effectiveness. Our MAGAN is implemented with the Pytorch framework (version 1.0) on a GPU with NVIDIA Titan Xp (12GB memory).

4.2 Experimental Data

In the experiments, the images from the DIV2K dataset [31] are utilized for training our network, and this dataset mainly contains 800 RGB high-quality (2K resolution) images training for image restoration. Our MAGAN method was trained in the RGB channels, and the training dataset is augmented by using the widely used techniques, such as random flips. Some popular benchmark datasets including Set14 [32], BSD100 [33] and Urban100 [34] were used to evaluate the SR performance of our MAGAN and existing state-of-the-art (SOTA) methods. Note that the former two benchmark datasets consist of natural RGB images, and the Urban100 dataset contains RGB images about building structures.

4.3 Ablation Studies

In order to study the contributions of some components (e.g., MRDN, CBAM and pre-training) in our MAGAN, we conduct the following ablation studies.

MRDN The MRDN structure is used to extract more detailed features for our network. We compare the results of MAGAN with and without MRDN, as shown in Fig. 5. Note that MAGAN without MRDN uses the residual-in-residual dense block (RRDB) proposed in [15] as the generator, as the RRDB structure is similar to our MRDN. The result shows that texture features become more complete and natural by using MRDN, and no obvious cracks appear at the corners of the eyes. This is because the multi-recursive residual learning structure can better retain the information of original images.

CBAM In the experiment, we add the attention module (i.e., CBAM) to the generator and discriminator of MAGAN, respectively. In Fig. 6 (center), we can clearly observe that CBAM is used in the generator, there is more sharper and conspicuous feather edges than that CBAM is used in the discriminator (right), but its generated images are unreal. It is probably because the generator does not has more constraints on the images when it is with the CBAM module as the discriminator is not subjected to the attention module and can not determine whether the generated image is fake. The CBAM module is used in the discriminator, which enhances the discriminative ability of the discriminator and can improve the quality of the generated images.

Pre-training It can be seen clearly from Fig. 7 (right) that the restored image of our pre-trained network is better than that of our network without pre-training in terms of detail texture discrimination. With a large number of experiments, we find that pre-training is helpful for the performance of most GAN-based models, especially in the case that the model is more complex.

Fig. 6. Visual results of our MAGAN without (left, Non_Attention) and with CBAM applied in our generator (middle, Attention_G) or discriminator (right, Attention_D) on the Baboon image.

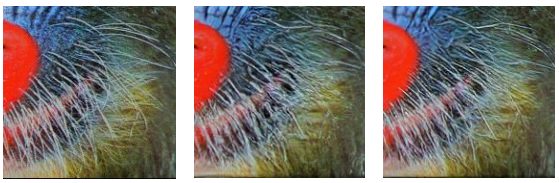


Fig. 7. Visual results of our MAGAN without (middle, Non_Pre) and with (right, Pre) the use of pre-training on the Baboon image.

4.4 Experimental Results

The image SR results of our MAGAN are compared with both of the PSNR-oriented algorithms such as SRResNet [7], RDN [10], and GAN-based methods which are EnhanceNet [14] and ESRGAN [15]. The quantitative results of the SR images recovered

Table 2. Average perceptual index results of SR images with scaling factor $\times 4$ on the three benchmark data sets. Note that the best results are shown in bold and the second-best results in italics.

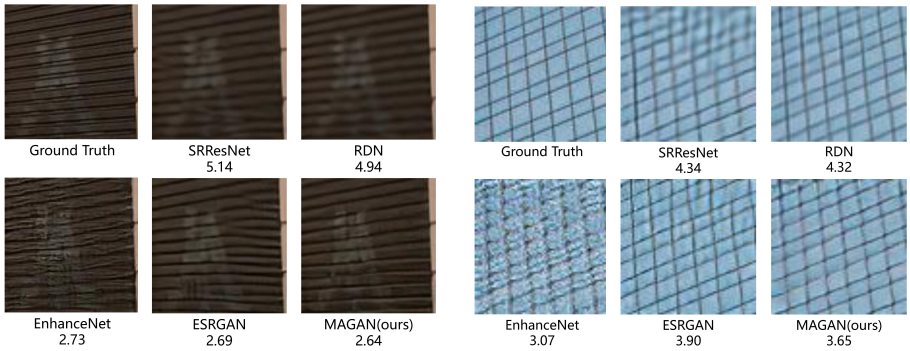
Methods	SRResNet [7]	RDN [10]	EnhanceNet [14]	SRGAN [13]	ESRGAN [15]	NatSR [16]	MAGAN (ours)
Set14	4.96	5.25	3.01	3.09	2.93	3.11	**2.89**
Urban100	5.15	5.05	**3.47**	3.70	3.77	3.65	3.59
BSD100	5.34	5.24	2.92	2.55	2.48	2.78	**2.43**

by these methods are reported in Table 2. We adopt the Perceptual Index (PI) [29] as a measurement metric for comparison. PI is a relatively effective indicator of visual quality than the others. It can be seen from the table that our MAGAN has gained a relatively lower average perceptual index 2.89 than the popular PSNR-oriented methods and the other GAN-based methods on Set14. The PIs of SRResNet and the RDN are much larger than GAN-based methods. It is because these methods are not trained with perceptual index. All the results show that our MAGAN usually outperforms other methods in terms of perceptual index, especially on the Set14 and BSD100 datasets.

The image SR results of a selected region of the image_089 from the Urban 100 dataset are shown in Fig. 8 (a). It can be seen that our MAGAN achieves the best visual result and its perceptual index are the lowest, i.e., 2.64. Compared with the traditional deep learning algorithms, SRResNet and RDN, which are optimized with classic L1- or L2-norm regularization, the GAN-based methods with perceptual loss can restore more detailed textures and gain a lower perceptual index. Moreover, our MAGAN yields much better results than the GAN-based methods such as EnhanceNet and ESRGAN. The result of EnhanceNet appears mixed textures and that of the ESRGAN represents blurry. The reason is probably that our MAGAN learns more detailed features with the proposed techniques: MRDN and CBAM, and therefore it generates perceptually superior results.

Furthermore, the image SR results of a selected region of the image_099 from the Urban 100 dataset are shown in Fig. 8 (b). All the results also indicate that our MAGAN achieves the best visual result and its perceptual index is 3.65. The original image has many grids as details. The images recovered by SRResNet and RDN, which are PSNR-oriented models, are blurred. Although the perceptual index of EnhanceNet is the lowest for this image, the visual result is inferior to those of MAGAN and ESRGAN. Since this image is a selected small region from image_099, the local result is probably not good. That may also result from the weakness of the metric of the perceptual index, which measures the global result of a recovered image but not locally. As indicated by the result of ESRGAN, for this image with fine texture details, the GAN based model easily generates redundant and nonexistent information resulting in a sharper effect so that the resulting images are different from the real images. On the contrary, our MAGAN performs well without generating redundant information.

Moreover, we present more representative results of all the methods for image SR tasks, as shown in Fig. 9. As there is no standard and effective metric for perceptual quality evaluation, we show the three common measurements: PSNR, SSIM, and PI (perceptual index). It is clear that our MAGAN usually recovers better images than the



(a) Image_089 (b) Image_099

Fig. 8. Comparison of the image SR results of SRResNet [7], RDN [10], EnhanceNet [14], ESRGAN [15] and MAGAN (ours). Among them, (a) and (b) are the SR results of Image_089 and Image_099, respectively.

other methods in terms of PI, and is much better than the perceptual-driven methods including EnhanceNet [14] and ESRGAN [15] in terms of PSNR, SSIM and PI.

It can be observed from these experimental results that our MAGAN method consistently outperforms other approaches in terms of both details and sharpness. For example, MAGAN produces better restored images (e.g., sharper, more natural baboon's whiskers and fur) than the PSNR-oriented methods (e.g., SRResNet), which tend to produce blurry results, and the GAN-based approaches, whose textures contain unpleasing noise and are unnatural. MAGAN can be capable of generating more detailed structures in buildings (see image_089), while other methods (including EnhanceNet and ESRGAN) either fail to add undesired textures or produce enough details. Moreover, existing GAN-based methods usually introduce unpleasant artifacts in their results. For instance, ESRGAN produces superfluous whiskers that do not exist as shown by the image baboon. Our MAGAN method can get rid of artifacts and also produces natural restored results.

5 Conclusion and Further Work

In this paper, we proposed MAGAN that performs consistently better in terms of perceptual quality than existing image SR methods. We also designed a novel architecture of a multi-level residual dense network for the generator in our MAGAN. Moreover, we introduced a multi-attention mechanism to our MAGAN by the CBAM, which can capture more detailed textures. In addition, we also improved our discriminator to stabilize the training by utilizing a new convolutional block, which applies both switchable normalization and spectral normalization. Experimental results confirmed the effectiveness of our MAGAN, and indicated the advantage of the proposed method: 1) MAGAN can restore visually high-quality images compared with existing state-of-the-art methods. 2) It can retain sharpness and yield fine textures for images as it utilizes the hierarchical features. 3) It does not generate unpleasant artifacts. In the future, more techniques such

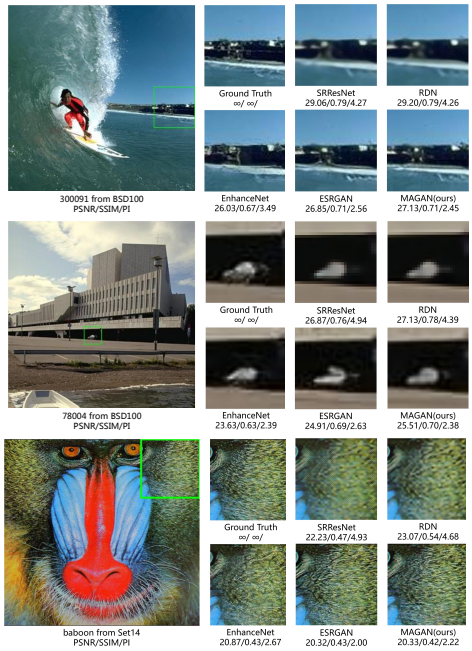


Fig. 9. Image SR qualitative results of SRResNet [7], RDN [10], EnhanceNet [14], ESRGAN [15], and MAGAN (ours) for a scale factor of $4\times$.

as the improved perceptual loss [35] and deformable convolution [36] will be investigated to generate more realistic images, and we will also apply the proposed network for video SR tasks [37,38,39].

Acknowledgements This work was supported by the National Natural Science Foundation of China (Nos. 61976164, 61876221, 61876220), and Natural Science Basic Research Program of Shaanxi (Program No. 2022GY-061).

References

1. W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, Q. Liao, Deep learning for single image super-resolution: A brief review, IEEE Transactions on Multimedia 21 (12) (2019) 3106–3121.
2. K. Zhang, X. Gao, D. Tao, X. Li, Single image super-resolution with non-local means and steering kernel regression, IEEE Transactions on Image Processing 21 (11) (2012) 4544–4556.
3. L. Zhang, X. Wu, An edge-guided image interpolation algorithm via directional filtering and data fusion, IEEE Transactions on Image Processing 15 (8) (2006) 2226–2238.
4. J. Hsu, C. Kuo, D. Chen, Image super-resolution using capsule neural networks, IEEE Access 8 (2020) 9751–9759.
5. Y. Shi, S. Li, W. Li, A. Liu, Fast and lightweight image super-resolution based on dense residuals two-channel network, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 2826–2830.

6. M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1664–1673.
7. C. Dong, C. C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Proceedings of European Conference on Computer Vision (ECCV), 2014, pp. 184–199.
8. B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 1132–1140.
9. W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig, Z. Wang, Is the deconvolution layer the same as a convolutional layer? (2016). arXiv:1609.07009.
10. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2472–2481.
11. Y. Hu, X. Gao, J. Li, Y. Huang, H. Wang, Single image super-resolution with multi-scale information cross-fusion network, Signal Processing 179 (2021) 107831.
12. K. Chang, M. Li, P. L. K. Ding, B. Li, Accurate single image super-resolution using multi-path wide-activated residual network, Signal Processing 172 (2020) 107567.
13. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.
14. M. S. M. Sajjadi, B. Scholkopf, M. Hirsch, Enhancenet: Single image super-resolution through automated texture synthesis, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4501–4510.
15. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: The European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 63–79.
16. J. W. Soh, G. Y. Park, J. Jo, N. I. Cho, Natural and realistic single image super-resolution with explicit natural manifold discrimination, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
17. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.
18. J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: The European Conference on Computer Vision (ECCV), 2016, pp. 694–711.
19. X. Wang, K. Yu, C. Dong, C. Change Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 606–615.
20. A. Jolicoeur-Martineau, The relativistic discriminator: a key element missing from standard GAN, in: International Conference on Learning Representations, 2019.
21. R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: The European Conference on Computer Vision (ECCV), 2018, pp. 800–815.
22. R. Mechrez, I. Talmi, F. Shama, L. Zelnik-Manor, Learning to maintain natural image statistics, arXiv: 1803.04626.
23. Z. Wang, J. Chen, S. C. H. Hoi, Deep learning for image super-resolution: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1.
24. T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, J. McAuley, ReZero is all you need: Fast convergence at large depth (2020). arXiv:2003.04887.

25. S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: The European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
26. P. Luo, J. Ren, Z. Peng, R. Zhang, J. Li, Differentiable learning-to-normalize via switchable normalization, in: International Conference on Learning Representations (ICLR), 2019.
27. T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations (ICLR), 2018.
28. A. Odena, J. Buckman, C. Olsson, T. Brown, C. Olah, C. Raffel, I. Goodfellow, Is generator conditioning causally related to GAN performance?, in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 3849–3858.
29. Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, L. Zelnik-Manor, The 2018 pirm challenge on perceptual image super-resolution, in: The European Conference on Computer Vision (ECCV) Workshops, 2018.
30. A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the 30th International Conference on Machine Learning (ICML), 2013.
31. E. Agustsson, R. Timofte, Ntire 2017 challenge on single image super-resolution: Dataset and study, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1122–1131.
32. R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, 2012, pp. 711–730.
33. D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 2001, pp. 416–423.
34. J. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197–5206.
35. M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, J.-P. Thiran, SROBB: Targeted perceptual loss for single image super-resolution, in: The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 2710–2719.
36. X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets V2: More deformable, better results, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9300–9308.
37. H. Liu, P. Zhao, Z. Ruan, F. Shang, Y. Liu, Large motion video super-resolution with dual subnet and multi-stage communicated upsampling, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), 2021.
38. H. Liu, Z. Ruan, C. Fang, P. Zhao, F. Shang, Y. Liu, L. Wang, A single frame and multi-frame joint network for 360-degree panorama video super-resolution, arXiv Preprint arXiv:2008.10320.
39. H. Liu, Z. Ruan, P. Zhao, F. Shang, L. Yang, Y. Liu, Video super resolution based on deep learning: A comprehensive survey, arXiv Preprint arXiv:2007.12928.