



Face Super-Resolution with Spatial Attention Guided by Multiscale Receptive-Field Features

Weikang Huang², Shiyong Lan^{1,2}, Wenwu Wang³, Xuedong Yuan¹, Hongyu Yang¹, Piaoyang Li¹, and Wei Ma¹

¹ College of Computer Science, Sichuan University, Chengdu 610065, China

² National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China

lanshiyong@scu.edu.cn

³ University of Surrey, Guildford GU2 7XH, UK

w.wang@surrey.ac.uk

Abstract. Face super-resolution (FSR) is dedicated to the restoration of high-resolution (HR) face images from their low-resolution (LR) counterparts. Many deep FSR methods exploit facial prior knowledge (e.g., facial landmark and parsing map) related to facial structure information to generate HR face images. However, directly training a facial prior estimation network with deep FSR model requires manually labeled data, and is often computationally expensive. In addition, inaccurate facial priors may degrade super-resolution performance. In this paper, we propose a residual FSR method with spatial attention mechanism guided by multiscale receptive-field features (MRF) for converting LR face images (i.e., 16×16) to HR face images (i.e., 128×128). With our spatial attention mechanism, we can recover local details in face images without explicitly learning the prior knowledge. Quantitative and qualitative experiments show that our method outperforms state-of-the-art FSR methods.

Keywords: Face super-resolution · Multiscale receptive-field features · Spatial attention mechanism · Deep learning

1 Introduction

Face super-resolution (FSR), also known as face hallucination, aims to generate high-resolution (HR) face images from corresponding low-resolution (LR) face images. In real world scenarios, there are many low-resolution (LR) face images, generated due to the limitation in an optical imaging system or the program used for image compression. In LR face images, some details may be lost, thus

This work was funded in part by the Key R&D Project of Sichuan Science and Technology Department, China (2021YFG0300), and in part by 2035 Innovation Pilot Program of Sichuan University, China.

leading to performance degradation for tasks such as face recognition and face landmark prediction. Thus, FSR has attracted increasing interest in a wide range of applications (e.g., face tracking, restoration of old face images).

FSR can be considered a special sub-task of single image super-resolution (SISR) [5]. Compared with SISR which takes images in different scenes as input, FSR only considers face images which are of similar structure. Therefore, FSR methods may offer better results than SISR on enhancing LR face images with higher upscaling factors (e.g., $8\times$). In recent years, with the rapid development of deep learning techniques, a number of face super-resolution methods have been proposed [2, 3, 9, 16, 19, 23].

Different from general images, a face image is a highly structured object with facial landmarks and facial parsing maps. Such information has been used by many FSR methods to generate HR face images. For example, Song *et al.* [19] adopted CNNs to learn basic facial components first, and then synthesized fine-grained details from a high resolution training set to enhance these components. Kim *et al.* [9] proposed a progressive FSR model that generated multiscale SR results and applied a distilled face alignment network (FAN) to predict face landmark locations. Chen *et al.* [3] designed an end-to-end FSR network to recover the SR face images using the facial landmarks and parsing maps estimated via the network. Ma *et al.* [16] developed a FSR method using two recurrent networks for image restoration and landmark estimation, respectively. Although joint training with the facial prior information helps recover the key face structures, there are two major limitations. First, it is labour-intensive to manually label the data required for training the network for estimating the prior information. Second, it is difficult to estimate the prior information precisely for each face image, as each person’s face is unique. Inaccurate prior information (e.g. location information) may lead to degraded FSR performance.

In this paper, we propose a multiscale receptive-field residual network (MRR-Net)¹ for face super-resolution, by introducing a spatial attention mechanism within the multiscale receptive-field residual blocks (MRRb). The key idea of our spatial attention mechanism is to obtain multiscale receptive-field features using concurrent convolution operation with different kernel size and then concatenate these features to generate the attention map. The spatial attention mechanism facilitates learning of face components of different size, as well as their outlines, allowing them to be processed at different scales. Our method exploits the advantage of convolution with different receptive-field and the efficiency of a CNN structure.

The main contributions of this paper are summarized as follows:

- We design a deep encoder-decoder residual framework for face super-resolution named MRRNet without explicitly learning facial prior knowledge.
- An improved spatial attention mechanism based on multi-scale receptive fields is used in each embedding layer (i.e., MRRb) of the encoder, for capturing face attributes at different scales for reconstructing the face images.

¹ <https://github.com/SYlan2019/MRRNet>.

Our method achieves state-of-the-art performance in terms of several metrics for image quality evaluation.

2 Related Work

Recently, significant progress has been made in face super-resolution using deep learning techniques. Yu *et al.* [24] introduced a generative adversarial network (GAN) to produce HR face images that are similar to real images. Chan *et al.* [1] designed an encoder-bank-decoder architecture for FSR using pre-trained GANs. Facial prior guided FSR methods utilize unique facial information to facilitate face reconstruction. Yu *et al.* [23] developed a convolutional neural network of two branches with one for estimating facial component heatmaps and the other for reconstructing face images aided by the heatmaps. Ma *et al.* [16] introduced a recursive cooperative FSR method with two recurrent networks that focus on image restoration and landmark estimation. More accurate landmark can be predicted based on better SR face image, which in turn can be recovered based on more precise landmarks. Thus the two recurrent networks can benefit from each other. However, such approaches might generate unnatural face images due to the difficulty in accurately estimating the unique facial information. In addition, building an estimation network requires additional effort for labelling data and training the network.

Attention mechanisms have been widely applied in low-level vision tasks, such as image enhancement and face super-resolution. Zhang *et al.* [27] proposed a residual channel attention block (RCAN) which generates different attention for each channel-wise feature to improve the discrimination ability of their network. Liu *et al.* [12] incorporated convolutional block attention module (CBAM) into their UNet-like generator to enhance the representation of regions of interest for anomaly detection. Chen *et al.* [2] proposed a face attention unit (FAU) that generates an attention map to enlarge the weight of the feature map related to face components. Unlike [27] which is based on channel attention, our attention mechanism utilizes spatial attention which contains more location information of face components. Despite being similar to [2] which takes advantages of multiscale features, our attention map is built from the multiscale receptive-field features corresponding to the face components of different size.

Different from the conventional way for building CNN networks in vision tasks, i.e. stacking many small spatial convolutions (e.g., 3×3) to enlarge the receptive field [4], several new ideas have emerged recently. In ConvNeXt [14], a Vision Transformers-like pure CNN is designed which outperformed Swin Transformers [13] on detection and segmentation tasks. One of their modifications in ConvNeXt was to use convolution with large kernel size 7×7 . Ding *et al.* [4] demonstrated that using a few large convolutional kernels instead of using a stack of small kernels could obtain much larger effective receptive field, so as to achieve better performance in low-level vision tasks. Thus, different kernel sizes of convolutions are considered in the design of our spatial attention mechanism.

3 Proposed Method

3.1 Overview

In face super-resolution task, we aim to convert LR face images I_{LR} to SR counterparts I_{SR} which are close to the ground truth face images I_{HR} . In this paper, we propose a residual FSR method with spatial attention mechanism guided by multiscale receptive-field features. As shown in Fig. 1, our proposed method consists of two networks including a multiscale receptive-field residual network (MRRNet) and an average discriminator. MRRNet works as a generator to generate I_{SR} . To recover the face components, MRRNet employs multiscale receptive-field residual (MRR) blocks (which is introduced in Sect. 3.2). In addition, we utilize the average discriminator and other losses (introduced in Sect. 3.3) to recover the face images with additional details.

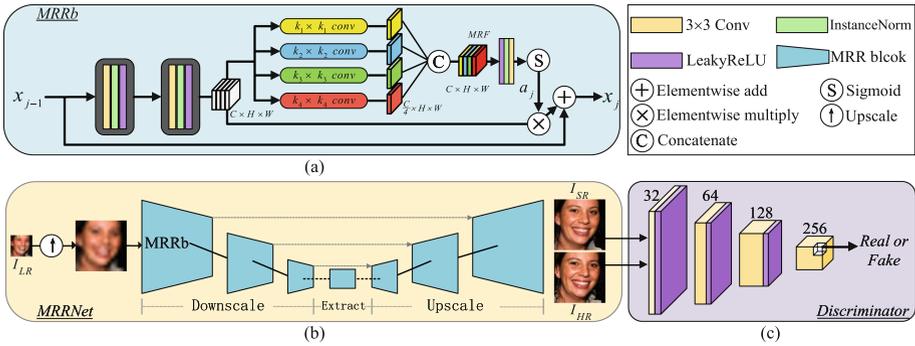


Fig. 1. The framework of the proposed method for face super-resolution. (a) is the detailed structure of our multi-scale receptive-field residual block. (b) is our generative network which is composed of MRR block. Rather than recovering low resolution face images directly, we resized them to 128×128 first through bicubic. (c) is our proposed average discriminator.

3.2 Multiscale Receptive-Field Residual Block

When observing face images, we usually look at the overall outline first, then we pay attention to the key facial components (e.g., eyebrows, eyes, nose, and mouth). This means that a FSR network is expected to not only pay attention to the overall structure, but also the key local details. However, due to the fact that the facial components and overall outline are of different size, it is not trivial to reconstruct the face features at various scales. To address this issue, we propose a spatial attention mechanism guided by multiscale receptive-field features (MRF) embedded in a residual block. Our MRF is motivated by the Inception module in [20] and ConvNeXt [14]. The Inception module applies

convolutions of different kernel size (i.e. 1, 3, 5) for multiscale feature processing, while ConvNeXt uses 7×7 convolution to simulate shifted window in Swin Transformer [13]. Thus, in our system, we design MRF using convolutions of four different kernel sizes (i.e. 1, 3, 5, 7).

By stacking the MRR blocks, our spatial attention mechanism helps the network to focus on facial features in different size. Denote the feature input of the j -th indexed MRR block as $x_{j-1} \in \mathbb{R}^{C_{j-1} \times H_{j-1} \times W_{j-1}}$. The fusion of multiscale receptive-field features can be computed as:

$$f_j^{c2} = \text{Conv}_j^2(\text{Conv}_j^1(x_{j-1})) \quad (1)$$

$$f_j^{fusion} = \text{Concat}[\text{Conv}_j^{k_1}(f_j^{c2}), \text{Conv}_j^{k_2}(f_j^{c2}), \text{Conv}_j^{k_3}(f_j^{c2}), \text{Conv}_j^{k_4}(f_j^{c2})] \quad (2)$$

where $\text{Conv}_j^{i=1,2}$ is the i -th convolutional layer followed by Instance Normalization and Leaky ReLU activation function of the j -th MRR block. Features $f_j^{c2} \in \mathbb{R}^{C_j \times H_j \times W_j}$ extracted by Conv_j^2 are passed to four convolutions (e.g., $\text{Conv}_j^{k_i}$) of different kernel size (i.e., k_1, k_2, k_3, k_4) without using normalization and activation function. Each convolution outputs a feature map where the number of channels is a quarter of f_j^{c2} . We use padding in the four convolutions to ensure the four feature maps to be of same size. Then, we concatenate the four feature maps to get $f_j^{fusion} \in \mathbb{R}^{C_j \times H_j \times W_j}$, where f_j^{fusion} contains multiscale features that can be further utilized for generating the attention map $a_j \in \mathbb{R}^{1 \times H_j \times W_j}$, as follows:

$$a_j = S\left(C_j(f_j^{fusion})\right) \quad (3)$$

where C_j is a convolution operation that only has one kernel. $S(\cdot)$ denotes the sigmoid function. Finally, the output of the j -th MRR block is computed as:

$$x_j = x_{j-1} + a_j \otimes f_j^{c2} \quad (4)$$

where \otimes is element-wise multiplication, and a_j assigns a value between 0 and 1 to f_j^{c2} which is passed through the channels.

We design our MRRNet as a hourglass-like network, due to the potential benefits offered by its downscale and upscale layers in improving feature representation. Specifically, we implement the downscale layer by adding a convolution to the residual branch of the MRR block to downscale the feature size, meanwhile in the upscale layer, the feature map is up-sampled by using the nearest neighbor interpolation. Thus, Eq. (4) is changed in downscale and upscale layers to:

$$x_j = \text{Conv}^d(x_{j-1}) + a_j \otimes f_j^{c2} \quad (5)$$

$$x_j = I_n(x_{j-1}) + a_j \otimes f_j^{c2} \quad (6)$$

where Conv^d denotes downscaling the convolutional layer with stride 2, and I_n is the nearest interpolation where the scale factor is typically chosen as 2 [17].

3.3 Objective Functions

Pixel Loss: We first train the MRRNet by optimizing the L1 loss with N pairs of LR-HR images as follows

$$\mathcal{L}_{pixel} = \frac{1}{N} \sum_{i=1}^N \|I_{HR}^i - \mathcal{F}_{MRRNet}(I_{LR}^i, \Theta)\| \quad (7)$$

where \mathcal{F}_{MRRNet} and Θ are the MRRNet and its parameters, respectively. I_{HR}^i is the i -th HR image, I_{LR}^i is the i -th LR image. We then apply other losses including adversarial loss and perceptual loss to train MRRGAN (generative adversarial version of MRRNet) for reconstructing high-fidelity face image.

Adversarial Loss: Recently, GAN [6] has been shown to be a powerful method for generating high-fidelity images. Therefore, we introduce Relativistic Average HingeGAN (RAHingeGAN) [8] to generate photo-realistic face images. RAHingeGAN D outputs a matrix as shown in the discriminator in Fig. 1. Each element in this matrix reflects the confidence level on how similar a certain area of the SR images is to that of the HR images. The discriminator D differentiates the ground-truth and I_{SR} by minimizing:

$$\begin{aligned} \mathcal{L}_D = & \mathbb{E}[\max(0, 1 - (D(x_h) - \mathbb{E}[D(x_g)]))] \\ & + \mathbb{E}[\max(0, 1 + (D(x_g) - \mathbb{E}[D(x_h)]))] \end{aligned} \quad (8)$$

Meantime, the generator G tries to deceive D by minimizing:

$$\begin{aligned} \mathcal{L}_G = & \mathbb{E}[\max(0, 1 + (D(x_h) - \mathbb{E}[D(x_g)]))] \\ & + \mathbb{E}[\max(0, 1 - (D(x_g) - \mathbb{E}[D(x_h)]))] \end{aligned} \quad (9)$$

where x_h and x_g are ground-truth and super-resolved face image, respectively.

Perceptual Loss: Perceptual loss [7] encourages G to generate natural results in perception. Perceptual loss here is defined as the l_1 norm between the feature maps of the ground-truth and I_{SR} extracted by a pre-trained VGG19 network [18], as follows:

$$\mathcal{L}_{per} = \mathbb{E} \left[\sum_{i=1}^I \|\phi_i(x_g) - \phi_i(x_h)\|_1 \right] \quad (10)$$

where ϕ_i extracts feature maps by the i -th layer of the VGG network and I is the number of the layers used. Finally, by combining the loss functions with different weights, we get the total loss defined as

$$\mathcal{L}_{total} = \lambda_G \mathcal{L}_G + \lambda_p \mathcal{L}_{per} + \lambda_{pix} \mathcal{L}_{pixel} \quad (11)$$

where λ_G , λ_p and λ_{pix} are weights, which are used to adjust the relative importance of \mathcal{L}_G , \mathcal{L}_{per} , and \mathcal{L}_{pixel} , respectively.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on two widely used face datasets: CelebA [15] and Helen [11]. For both datasets, we first crop face images roughly according to landmark. Then, we further remove the excess background in the cropped face images by setting a threshold. Finally, we resize the cropped images to 128×128 as HR (i.e., ground-truth) face images and downsample the HR images to 16×16 as LR face images. For the CelebA dataset, we use 193K images for training and 1K images for testing. For the Helen dataset, we use 2000 images for training and 50 images for testing. Our test sets are the same as in [16]. We evaluate SR results using performance metrics PSNR, SSIM [21] and LPIPS [26], respectively. PSNR and SSIM are most commonly used evaluation metrics in super-resolution task, calculated on the Y channel of the transformed YCbCr space in our experiments. Learned Perceptual Image Patch Similarity (LPIPS) is a deep features based metric, evaluating the perceptual similarity between two images.

4.2 Implementation Details

We set the number of MRRb blocks in the encoder, extractor, decoder to 3, 10, 3, respectively. For training the MRRNet model, we set $\lambda_{pixel} = 1$ and when training the MRRGAN, the parameters λ_G , λ_{per} , and λ_{pixel} are set as 0.01, 0.01 and 1, respectively. The kernel sizes k_1, k_2, k_3, k_4 are set to 1, 3, 5, 7, respectively. The entire network is optimized using Adam [10] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$ and a learning rate 0.0001. For data augmentation, we use random horizontal flipping, and image rescaling. We built our network in Pytorch and trained it on an NVIDIA RTX 3080 GPU.

4.3 Results and Analysis

Comparison with the State-of-the-Art Methods: We compare our method with state-of-the-art FSR and general SR methods qualitatively and quantitatively on Helen and CelebA test sets provided by [16]. For methods that provide training codes, we retrain these models on our training set. Table 1 shows the PSNR and SSIM results on Helen and CelebA datasets. Different from our method, DIC [16], PFSR [9] and FSRNet [3] applied facial prior knowledge to improve super-resolution performance. It can be observed that our MRRNet outperforms other methods in terms of the PSNR and SSIM metrics. Our MRRGAN has a lower PSNR and SSIM than MRRNet, but gives comparable performance with other methods. This shows that our method offers better performance, achieving balance in perceptual quality and pixel accuracy of super-resolved face image, thanks to the special network design with the attention mechanism and integrated loss function. Apart from PSNR and SSIM, we evaluate our methods and other methods with LPIPS [26] which reflect perceptual similarity based on deep features. Our MRRGAN has achieved best scores on both datasets, which

demonstrates that the super-resolved face images generated by our MRRGAN are perceptually more similar to the HR face images.

Table 1. Quantitative comparison on CelebA (rows 2 to 4) and Helen (rows 5 to 7) with state-of-the-art FSR methods. The best and second best performance are highlighted in red and blue, respectively.

Method	Bicubic	PFSR [9]	URDGN [24]	FSRNET [23]	DIC [16]	DICGAN [16]	MRRNET	MRRGAN
PSNR	23.73	23.97	24.73	26.29	27.22	26.38	27.27	26.64
SSIM	0.6261	0.6787	0.6871	0.7518	0.7845	0.7517	0.7873	0.7632
LPIPS	0.5329	0.2716	0.2427	0.2315	0.1974	0.0976	0.1921	0.0930
PSNR	24.04	23.61	24.22	25.20	26.75	26.02	26.86	26.15
SSIM	0.6743	0.6486	0.6909	0.7091	0.7894	0.7518	0.7912	0.7614
LPIPS	0.5253	0.2636	0.2449	0.2356	0.2050	0.0966	0.2038	0.0941

In Fig. 2, we visualize some super-resolution results of different methods. We can see that even without face prior information, MRRNet can still correctly generate face key components including eye, nose, and mouth. This is because our spatial attention mechanism exploits feature maps obtained through different receptive fields. Furthermore, compared with DICGAN where a discriminator D is used to differentiate the ground-truth and the super-resolved images. MRRGAN achieved better visual results, i.e. giving clearer textures and more realistic details in eyebrow, teeth and other facial components. The qualitative comparisons demonstrate the powerful ability of our methods for generating human face images.

Table 2. Average Euclidean distance between the estimated landmarks and ground truth landmarks.

Method	DICGAN [16]	MRRGAN
AED	1.4564	1.4126

As [3], we conduct facial landmark estimation comparison on Helen between DICGAN and MRRGAN that have the best visual effect in Fig. 2. The more consistent the predictions between super-resolved face images and GT are, the better the generated face images. We use OpenFace [25] to detect 68 facial landmarks of each face image. Then, we calculate the average Euclidean distance (AED) between the 68 landmarks of DICGAN, MRRGAN and GT respectively. We tabulate the results in Table 2. Figure 3 shows a landmark estimation example of DICGAN, MRRGAN and GT. From Fig. 3 we can see the lower AED value, the closer the image is to GT. The comparison demonstrates that MRRGAN has powerful generation ability in recovering facial components of different sizes precisely.

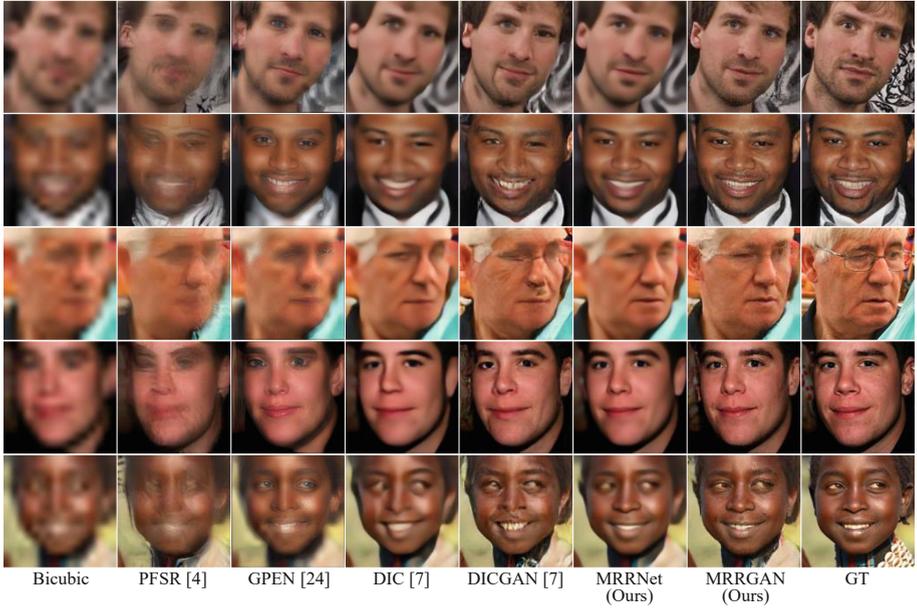


Fig. 2. Visual comparison with state-of-the-art methods. The size of low resolution face image is 16×16 and magnified by a factor of 8. More details could be observed by zooming into the figures.

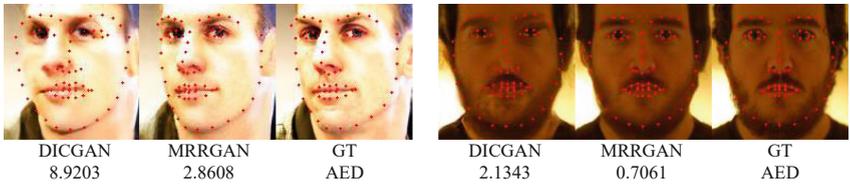


Fig. 3. Landmark detection comparison between DICGAN and our MRRGAN. For example, the landmark detection accuracy at the nose by the proposed method is significantly higher than that of the baseline.

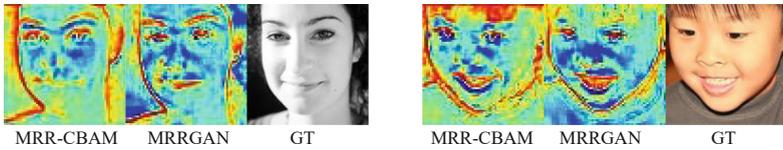


Fig. 4. Visualization of attention map in MRRGAN and MRR-CBAM.

Table 3. Quantitative comparison of different kernel sizes combinations used in MRRGAN on Helen.

Method	MRRGAN- $k_{1,3}$	MRRGAN- $k_{5,7}$	MRRGAN- $k_{1,7}$	MRRGAN
PSNR	26.23	26.49	26.31	26.64
SSIM	0.7515	0.7586	0.7530	0.7632

Effects of Spatial Attention Mechanism: In order to demonstrate the effectiveness of the proposed spatial attention mechanism. We conduct an experiment between our spatial attention mechanism and CBAM [22] for extracting key facial components. Specifically, we replace our spatial attention mechanism with CBAM in MRRGAN (denoted as MRR-CBAM) and keep other settings the same. In Fig. 4, we visualize some attention maps, generated in the upscale process of MRRGAN and MRR-CBAM. We can see that: 1) Our spatial attention mechanism can effectively learn to focus on facial components of different sizes (such as eyes, eyebrows, mouths, and facial contours). However, MRR-CBAM lacks attention to the small-scale lip lines in the left image. 2) CBAM pays more attention to the outer contours of the face, and even to the areas outside the face. However, our attention mechanism mainly focuses on the key parts of the face, maintaining low attention outside the face. We believe this is due to the use of convolution in different kernel size in the spatial attention mechanism. This experiment demonstrates that our spatial attention mechanism can guide the generation of HR face images.

Study of Kernel Size: To investigate the impact of the convolution in different kernel size used in our spatial attention mechanism. We conducted an experiment by combining convolutions with different kernel sizes pairwise in spatial attention within MRRGAN. In particular, we implement 1×1 with 3×3 , 5×5 with 7×7 and 1×1 with 7×7 (denoted as MRRGAN- $k_{1,3}$, MRRGAN- $k_{5,7}$ and MRRGAN- $k_{1,7}$) three combinations, which correspond to small-small, large-large and small-large convolution kernel size paired. We evaluate PSNR and SSIM on CelebA for these three combinations and show the results in Table 3. From this table, it can be observed that MRRGAN which has convolution with four kernels in spatial attention achieves the best performance, while MRRGAN- $k_{5,7}$ is better than other combinations.

Table 4. Quantitative comparison of different models, where w/o SA means without the spatial attention module.

Method	MRRNet w/o SA	MRRGAN w/o SA	MRRNet	MRRGAN
PSNR	26.83	25.47	27.27	26.64
SSIM	0.7585	0.7469	0.7873	0.7632

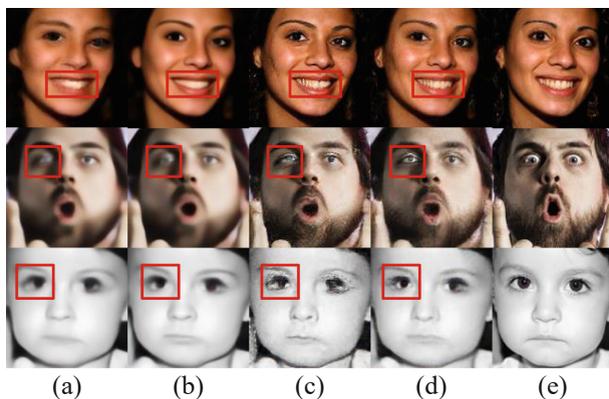


Fig. 5. Visual comparison between (a) MRRNet w/o SA. (b) MRRNet. (c) MRRGAN w/o SA. (d) MRRGAN. (e) GT. Better zoom in to see the detail

4.4 Ablation Study

We further perform an ablation study to demonstrate the effectiveness of our spatial attention mechanism. In the ablation experiment, we remove the spatial attention mechanism in MRRNet and MRRGAN which are called MRRNet w/o SA and MRRGAN w/o SA. The quantitative comparison results of PSNR and SSIM on Helen are presented in Table 4. It can be observed that, compared with MRRNet w/o SA and MRRGAN w/o SA, MRRNet and MRRGAN achieve better performances in all metrics. In Fig. 5, we visualize SR images generated with and without spatial attention mechanism and GAN. We can see that MRRNet w/o SA and MRRGAN w/o SA produced artifacts in key face components such as the eyes, while MRRGAN generates the best quality face images with the guidance of spatial attention.

5 Conclusion

We have presented a multiscale receptive-field residual network for face super-resolution. Specifically, a spatial attention mechanism guided by multiscale receptive field features embedded in a vanilla residual block helps recover the facial components of different size. The qualitative and quantitative experimental results on the CelebA and Helen datasets show the effectiveness of our method, as compared with other state-of-the-art FSR methods.

References

1. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: GLEAN: generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14245–14254 (2021)

2. Chen, C., Gong, D., Wang, H., Li, Z., Wong, K.Y.K.: Learning spatial attention for face super-resolution. *IEEE Trans. Image Process.* **30**, 1219–1231 (2020)
3. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: end-to-end learning face super-resolution with facial priors. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2492–2501 (2018)
4. Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J.: Scaling up your kernels to 31×31 : revisiting large kernel design in CNNs. *arXiv preprint arXiv:2203.06717* (2022)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014*. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
6. Goodfellow, I., et al.: Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27** (2014)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
8. Jolicœur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734* (2018)
9. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239* (2019)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
11. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
12. Liu, G., Lan, S., Zhang, T., Huang, W., Wang, W.: SAGAN: skip-attention GAN for anomaly detection. In: *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2468–2472. IEEE (2021)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. *arXiv preprint arXiv:2201.03545* (2022)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738 (2015)
16. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5569–5578 (2020)
17. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. *Distill* **1**(10), e3 (2016)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
19. Song, Y., Zhang, J., He, S., Bao, L., Yang, Q.: Learning to hallucinate face images via component generation and enhancement. *arXiv preprint arXiv:1708.00223* (2017)

20. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
22. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
23. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 217–233 (2018)
24. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 318–333. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_20
25. Zadeh, A., Chong Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3D facial landmark detection. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2519–2528 (2017)
26. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
27. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301 (2018)