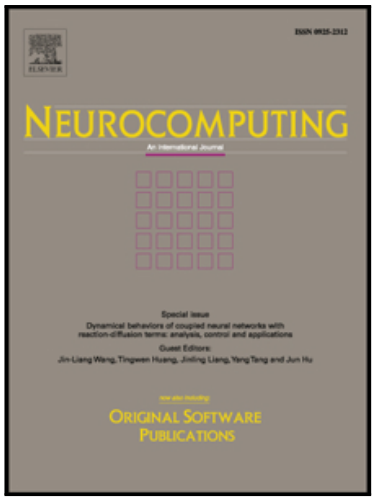
Journal Pre-proof



Facial Image Synthesis and Super-Resolution With Stacked Generative Adversarial Network

Jijun He, Jinjin Zheng, Yuan Shen, Yutang Guo, Hongjun Zhou

PII: S0925-2312(20)30574-9
DOI: https://doi.org/10.1016/j.neucom.2020.03.107
Reference: NEUCOM 22168

To appear in: *Neurocomputing*

Received date: 9 September 2019
Revised date: 21 February 2020
Accepted date: 31 March 2020

Please cite this article as: Jijun He, Jinjin Zheng, Yuan Shen, Yutang Guo, Hongjun Zhou, Facial Image Synthesis and Super-Resolution With Stacked Generative Adversarial Network, *Neurocomputing* (2020), doi: https://doi.org/10.1016/j.neucom.2020.03.107

Highlights

Facial Image Synthesis and Super-Resolution With Stacked Generative Adversarial Network

Jijun He, Jinjin Zheng, Yuan Shen, Yutang Guo, Hongjun Zhou

- End-to-end super-resolution method for facial images which can upscale the input image to a high resolution of up to 256×256
- Training GAN with the in-the-wild facial dataset and generating realistic synthesis results with identifying information kept
- A Flexible network structure which can process wide resolution range of input images with better performance compared to state of art algorithms

Facial Image Synthesis and Super-Resolution With Stacked Generative Adversarial Network

Jijun He[a,b], Jinjin Zheng[a,*], Yuan Shen[b], Yutang Guo[b], Hongjun Zhou[c]

[a] *Department of Precision Machinery and Instrumentations, University of Science and Technology of China, Hefei, P. R. China*
[b] *Department of Computer Science, Hefei Normal University, Hefei, P. R. China*
[c] *NSRL, USTC, Hefei, P. R. China*

Abstract

Image synthesis and super-resolution (SR) have always been a hot spot for computer vision and image processing research. Since the development of Deep Learning, especially after the Deep Convolutional Generative Adversarial Network (DC-GAN) methods, facial image synthesis and SR problem had been solved in many circumstances. But most of the existing works were focused on natural-looking of the synthesized result rather than keeping facial information of the original image. Our paper presented an end-to-end method of getting high-resolution photo-realistic facial images from low-resolution (LR) in-the-wild images without losing the facial identity details. The pipeline used a flexible stacked GAN structure for the SR process with different target image resolutions on different upscaling factors. To avoid getting blur or nonsensical image output and realize the flexibility, "U-Net" architecture and upsampling layers with residual learning blocks were stacked. The stacked network structure makes applying different loss func-

*Corresponding author
Email address: jjzheng@ustc.edu.cn (Jinjin Zheng)

tions in different parts of the network possible, which helps to solve the two problems of keeping identical facial details of the LR input image and generating high-quality output images simultaneously. By using 3 different loss functions in different positions of the stacked network separately, through experimental comparison, we found the best stacked residual block parameters which could get the best output image quality. Experimental results also explicated that the network had a good SR ability compare to state of the art methods in different resolution and upscaling factor.

Keywords:
Generative Adversarial Network, Face hallucination, Super Resolution, Image synthesis

1. Introduction

Facial image super-resolution (SR) problem, also known as face hallucination, had been well observed to fulfill the vast demand of face recognition, while low-resolution (LR) images were easy to capture from webcam or security camera widely positioned, but high-resolution (HR) image that can meet the demand of personal identification was difficult to acquire, so the problem of getting high-quality HR image from corresponding low-quality LR image became a hot spot of image processing research. Since the fast development of recent Convolutional Neural Network (CNN) based image synthesis research, there are many successful methods for image synthesis. But SR methods for facial images are still far from perfect because of its ill-posed nature [1]. After the development of the ground-breaking Generative Adversarial Network (GAN) method [2], many afterward methods achieved

significant results on facial image SR with CNN and GAN based data-driven methods [3], and sparse representation methods [4]. But the main problem of recent research is that normal SR methods usually focus on the synthesized image details like pixel arrangement [5] and local features [6], but not the overall facial identification feature, which would produce the synthesized image unlike the original ones and makes the recognition and identification disputable.

The most important problem in single image SR is to generate the lacking information absent in original images. It's always been a challenge to generate natural and acceptable synthesized facial images for deep learning methods like GAN, whose generator and discriminator balance is hard to fine-tune and difficult to converge. To solve this problem, "U-Net" architecture [7] and skip connections [8] are widely used. For further flexibility, the stacked architecture can help with training complex networks [9], and some flexible stacked network structures are successfully used to solve the problems of text described image synthesis [10], or medical image enhancement [11]. With multiple generators and corresponding discriminators network architecture, the image recovery with complicated pattern can also be accomplished [12]. Although simply stacked networks can synthesize reasonable results, the problem still exists when we need realistic SR facial images with identity information being kept. To guarantee the facial likeness of input and output image, some attempts have been made with the LR image as a guide feature for the SR network training [13]. On relative low-resolution output images, acceptable results could be got, but when the high-resolution situation was encountered, the results could meet multiple failure cases and the whole net-

work training was easy to degrade. We improved the traditional networks by stacked multiple residual blocks [14] after "U-Net" to make the upscaling results have high-quality details, each stacked residual block could double the height and width resolution. In different parts of the stacked network structure, different loss functions were introduced to let them have different optimization goals.

In the experiment, we used selected samples from the CelebA [15] dataset as training and testing datasets and compared different settings of stacked residual blocks, which would affect the upscaling results, then we chose the upscaling factor as 4 with different target resolutions. The SR results were compared with state of art methods and the performance of our work was better. Concerning the traditional peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index value [16, 17] only evaluate the local feature of the target image and can't reflect the likeness of the whole synthesized facial image with the corresponding HR image from the same person, we used the Fréchet Inception Distance (FID) score [18] as the measurement of SR image results, and compared the two datasets of ground truth HR results and synthesized SR results from LR input images, the calculated distance value would reflect the likeness correctly.

The working flow of this paper is demonstrated in Fig. 1. In the training process, the selected HR CelebA training dataset was used for GAN training. The training dataset contained LR facial images and the corresponding HR image as ground truth. The LR images would be resized to the target resolution with Bicubic-interpolation, and combined with their corresponding ground truth HR image into a training pair, then the combined image data

pairs would be fed as input data of GAN. The ground truth part would be used by discriminator only, so there is no need for the ground truth during the testing phase, and the synthesized SR facial image would be got at the generator output. At last, the FID score between synthesized SR images and ground truth HR images would be calculated to evaluate the SR results.

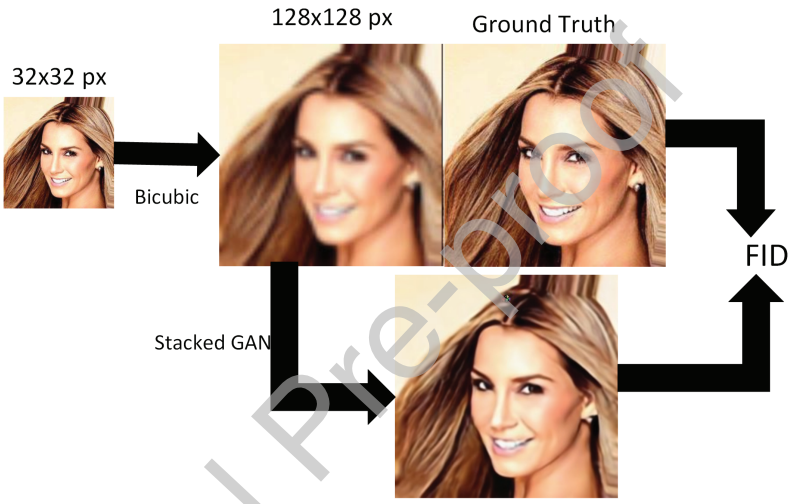


Figure 1: Working flow of this paper. The original LR image is resized to blurry HR image by the Bicubic-interpolation method, then the blurry HR image is fed into our Stacked GAN as input, network output SR image will be compared with the ground truth HR image and FID distance will be calculated.

The main contribution of our work can be summarized into 3 points: (1) We designed a flexible network structure of image synthesis compatible with different scaling factors. The stacked generators can be changed easily to fulfill different optimization purposes. (2) By adjusting residual blocks with different structures, we tested the best parameters for the synthesized image with both facial likeness and realistic detail. (3) Our end-to-end network is

easy to train without any pretrained parameters showing good universality and transferability.

The remaining parts of this paper will be presented as follows: Section 2 will have a review of related works in single facial image SR and facial image synthesis. Section 3 will introduce the architecture of our deep learning network. Section 4 will introduce the experiment in detail and evaluate the experimental results. Finally, some analysis and conclusion will be given in section 5.

2. Related Work

Different GANs have shown outstanding performances in different aspects of the image synthesis issue. In this section, we will discuss 2 related areas our work has involved: image super-resolution, which is also known as image upsampling, and facial image synthesis, especially methods synthesize image with a single input.

2.1. Image upsampling

GAN had been proved to be a powerful tool to generate synthesized facial images since the DC-GAN method [19] was developed. But training GAN that can output high resolution and photo-realistic image, e.g. resolution bigger than 64×64, can be unstable with blurry or nonsensical images generated [20]. Composed by Generator and Discriminator, GAN training must reach a delicate balance between generator loss and discriminator loss for the whole net to work [21]. Deep residual learning [14] can enhance the generated image quality according to results described in [10]. So we adopted the stacked residual block before upsampling layer to improve the quality of the

HR output image and got our output facial image reached the resolution of up to 256×256 with photo-realistic quality.

2.2. Single facial image super resolution

Facial image SR is a special case of general image SR, but there are still differences between them. It is an advantage that facial image SR can have more assumptions than general image SR, and many methods took advantage of that prior knowledge during facial image synthesis [22, 23] and chose high dimensional features to be the training data of GAN networks, but it is also a disadvantage that too many assumptions will make the synthesized facial image can be any suspicious result [24, 25]. On the other hand, GAN based methods had been shown convincible results on LR images transformation like 16×16 to 64×64 pixels [26, 13], but it still had some difficulties in synthesized images with higher target resolutions like 128×128 or 256×256. Meanwhile, the general evaluation method of PSNR and SSIM had shown to be not a good and significant measurement compared to FID score [18, 21] on facial image evaluation, for the traditional PSNR and SSIM focused on evaluating the local texture quality more than the overall likeness of SR image with the original LR source.

3. Theory and Network

In this section, we will demonstrate the process of generating SR facial images from in-the-wild LR images with variant poses. The main procedure includes 3 steps: first, resize the LR image into blur HR image using Bicubic-interpolation; secondly, combine blur HR image and ground truth HR image into a training pair for the training process; thirdly, feed the training pair

image into our stacked GAN, only feed the blur HR image if in testing process, and get the output SR image results.

3.1. Stacked Generative Adversarial Network

After resize of the original LR image, stacked GAN is used to synthesize the SR version of the image according to the provided LR image information. To keep the original details of the input image, we set a "U-Net" architecture [8] for the first part of the network. With the skip connection across the bottleneck vector layer, details of original images can be kept. After "U-Net", upsampling layers are set to get more photo-realistic HR images. Each upsampling layer contains multiple residual learning blocks and a deconvolution upscale layer, doubled the height and width resolutions of the image. The whole network is shown in Fig. 2.

The whole network can be divided into 2 parts: the "U-Net" part and the upsampling part which contains 2 upsampling blocks, each part has a different optimization goal. The whole network takes blur HR images of 256×256 as input whose source is LR images of 64×64 pixels. The first "U-Net" part downsampling the original blur facial image using convolution layers and upsampling it to a 64×64 image with skip connections, which means the synthesized result of this network part will make full usage of original LR image information. Then the "U-Net" output 64×64 image will be fed into upsampling layers one by one, doubled the image size for each time, finally the HR 256×256 output image can be got after two residual blocks. Notice that the stacked residual blocks can have more than two in case of bigger upscaling factors.

The optimization goal of this network can be formulated as the following

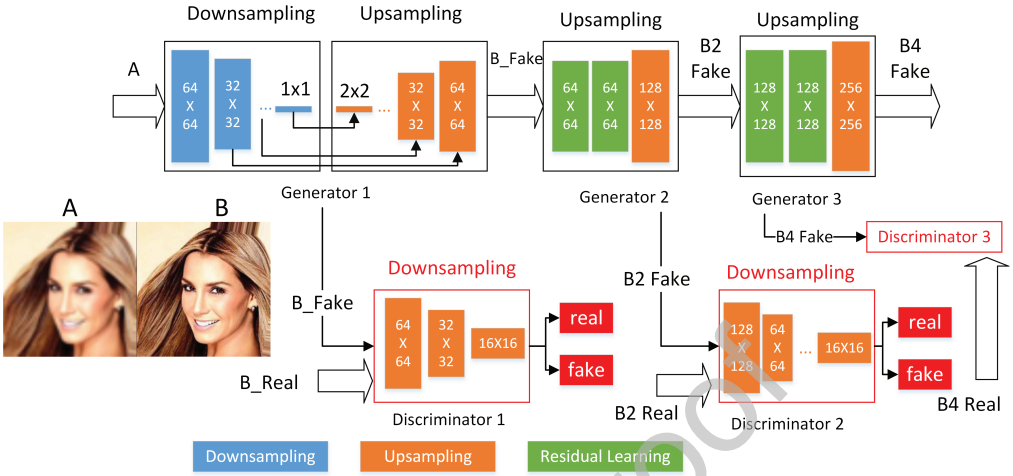


Figure 2: Stacked GAN architecture. There are 3 separate generators and their corresponding discriminators. The blue blocks stand for convolutional downsampling, the orange blocks stand for deconvolution upsampling, and the green blocks stand for residual learning. The input image is combined AB pair, A stands for the resized blur image from LR input, and B is the ground truth for the discriminator to compare with the synthesized fake image. B_Fake is the synthesized image results that have the same resolution with original LR input image, and $B2$, $B4$ Fake stands for 2 and 4 times upscaled SR image.

3 loss functions:

$$G_1 = arg\min_G \max_D (L_{AB'} + L_{BB'} + \lambda L_{L1}) \tag{1}$$

$$G_2 = arg\min_G \max_D (L_{AB'} + \lambda L_{L1}) \tag{2}$$

$$G_4 = arg\min_G \max_D (L_{AB'} + \lambda L_{L1}) \tag{3}$$

Which means the 3 separate generators try to minimize the loss function against the 3 corresponding discriminators try to maximize the loss functions. Their subscript number means scaling factors of every generator.

Loss function Eq. (1) means to evaluate the image synthesis quality compares to original input images. While the other 2 loss functions Eq. (2, 3) mean to optimize the upsampling image quality. G_1 means the optimization target image has the same resolution with the original LR input image, and G_2 and G_4 means their output images have 2 and 4 times of upscaling respectively.

The corresponding discriminator loss functions are expressed in Eq. (4, 5, 6), with D_1, D_2 and D_4 have the same naming regulation with the generator loss functions.

$$D_1 = L_{AB} + L_{AB'} + L_{BB'} \tag{4}$$

$$D_2 = L_{AB} + L_{AB'} \tag{5}$$

$$D_4 = L_{AB} + L_{AB'} \tag{6}$$

The loss functions consist of 4 parts, network cross-entropy L_{AB}, $L_{AB'}$, $L_{BB'}$, and L1 distance loss L_{L1}. With LR input image pixel data x from image A with a distribution of $p_{data(x)}$, synthesized HR image pixel data z from image B' with a distribution of $p_{data(z)}$ and ground truth HR image pixel data y from image B with a distribution of $p_{data(y)}$, following the definition of the GAN [2], the loss functions can be expressed as:

$$L_{AB} = E_{y\sim p_{data}(y)}[logD(y)] + E_{x\sim p_{data}(x),y\sim p_{data}(y)}[log(1 - D(G(x,y)))] \tag{7}$$

$$L_{AB'} = E_{y\sim p_{data}(y)}[logD(y)] + E_{x\sim p_{data}(x),z\sim p_{data}(z)}[log(1 - D(G(x,z)))] \tag{8}$$

$$L_{BB'} = E_{y\sim p_{data}(y)}[logD(y)] + E_{z\sim p_{data}(z),y\sim p_{data}(y)}[log(1 - D(G(z,y)))] \tag{9}$$

L_{AB} means cross-entropy of original input A and ground truth B which is only used in discriminators because only discriminators need the ground

truth of different resolutions. $L_{AB'}$ means cross-entropy of original input A and synthesized image B'. $L_{BB'}$ means cross-entropy of synthesized image B' and LR ground truth B, that is the original input LR image A, which means $L_{BB'}$ can only be got in the "U-Net" part of the network. The front part of these equations are $E[logD(y)]$, means expectation of log-likelihood loss function which the discriminators can divide the synthesized z from the ground truth y, while the latter part of the equations are $E[log(1 - D(G))]$, means the expectation that how well the generators can fool the corresponding discriminators. Notice from the equations that $E[logD(y)]$ is trained by y from the ground truth B and tested by synthesized image B'.

$$L_{L1} = E_{x\sim p_{data}(x),y\sim p_{data}(y),z\sim p_{data}(z)}[|y - G(x,z)|] \tag{10}$$

L1 distance loss L_{L1} is included to make sure the output images will have realistic detail rather than blurry outcomings, which reflects the expectation of the difference between the generated z value and the ground truth y value, according to the conditional x value.

In generator G_1, we can evaluate $L_{AB'}$, $L_{BB'}$ and L_{L1}, because the original input image can be the standard for the synthesized result, and the conditional GAN of G_1 can have proper weight map considering the LR input image as the ground truth. The following G_2 and G_4, although without a standard image as the ground truth, can accept a synthesized image from G_1 as input and upsample it by evaluating $L_{AB'}$ and L_{L1}, which are focus on the synthesis of pattern detail. The discriminators are trained by the product of G_1 and fine-turned by the following upsample generators G_2 and G_4. By the stacked generators and corresponding discriminators together, we can make the output image have the facial likelihood with the input LR image and also

guarantee the SR final effect with realistic details.

3.2. Residual Block

To get a more photo-realistic output image other than an unstable poor-quality image with artifacts, the residual networks are inserted before each upscaling process, they consist of multiple convolutional layers and a skip connection [14]. A residual block consists of multiple residual learning network and a deconvolution upsampling layer. The convolution layers in residual learning network have a 1×1 core and 1 pixel step which keeps the input and output image have the same size, the upsampling deconvolution layer behind upscales the input image to make the height and width pixel doubled. The network structure is shown in Fig. 3.

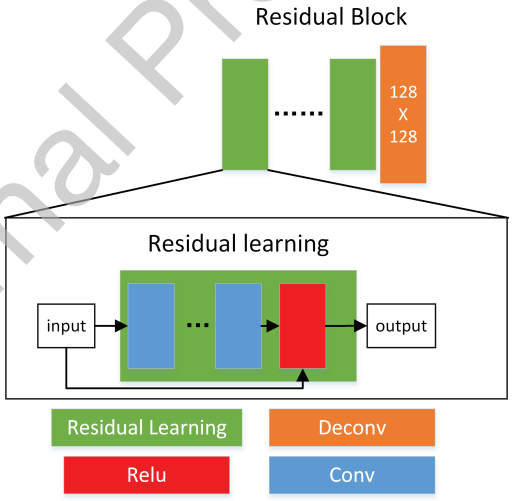


Figure 3: Residual block, a flexible structure that can have multiple convolution layers in one residual learning structure and each block can contain multiple residual learning structures.

The residual learning network can be expressed as:

$$z_o = z_i + F(z_i, W) \tag{11}$$

Where z_i is input pixel value which is synthesized by the former network, and z_o is the output value of the residual learning network. $F(z_i, W)$ stands for the residual part consists of one or multiple convolutional layers and one Relu activation function [27]. W is the network weight. With different residual block structures, different optimization results can be got, the details will be discussed in the following experimental parts.

4. Experiment and Evaluation

In this section, we will describe the training process, we have chosen the dataset of CelebA [15] which has 202599 image samples including 10177 different identities of celebrities, from which we selected 6000 random images as training data and another 2000 random images as testing data with the resolution of 256×256 pixels. We use the FID score as a metric to evaluate the synthesis results, as while as the traditional PSNR and SSIM value, to check if our synthesized facial images have a better performance than other state-of-art methods.

4.1. GAN Training

CelebA dataset provided images with different resolutions, but to generate 256×256 high-resolution images, the training process needs HR facial images of the same size as ground truth. So before training, we have chosen 5 points facial landmarks and cropped the CelebA images into 256×256 pixels

with aligned facial landmarks, the face area was centered and had a ratio of 0.7 of the whole image on average.

In the training process, we used Bicubic-interpolation method to resize the HR images into LR images, and then resize it again to get blur HR images. At last, we combined the HR ground truth with blur HR images to generate training image pairs. To evaluate the SR ability in different target sizes, we generated training and testing samples with different target resolutions, including 64×64, 128×128 and 256×256, which stands for "16 to 64", "32 to 128" and "64 to 256" transformation.

4.2. Residual Blocks Parameters

The experimental results had shown that different residual block network settings have different influences on synthesized results.

We conducted 2 sets of experiments, the "16 to 64" and "32 to 128" transformation. Generator networks with different numbers of residual learning structures are tested as well as different numbers of convolution layers in each residual learning structure. The labels in the form of "$m \times n$" in Fig. 4 means the training set has a residual block structure including m residual learning structures and each structure has n convolution layers. According to the experimental results shown in Fig. 4, we chose the "4×4" residual block setting in the following compare experiments, which has the best performance in FID score in both resolutions.

After introduced the "4×4" residual blocks structure into the generator network, the training and testing calculation consumption were not increased apparently for the simple operation of residual training layers. The training time increment was less than 20%, which is about 20 hours, and the test-

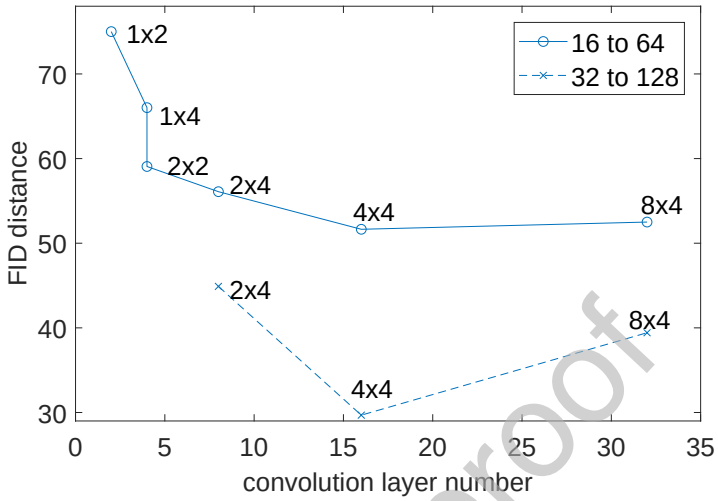


Figure 4: Different residual block setting has different FID results. "4×4" label on top of the line means the corresponding residual block is consists of 4 residual learning structure and each structure contains 4 convolution layers.

ing time increment was unnoticeable for around 73ms in our TITAN X $\times 2$ platform.

4.3. Result Evaluation

Some of the SR results are shown in Table 1, Table 2 and Fig. 5. We calculated the FID score, PSRN and SSIM values of transformations in different target resolution, and marked the best value in bold with the FID score is the smaller the better and the other 2 values are on the contrary. The target resolutions were 64×64, 128×128 and 256×256 pixels, and the scale factors were 4. The results have shown that our method had the best average performance on all targets, even though our result was not the best in the first and third row, but still had acceptable results and very close to

the best ones.

According to Table 2, the PSNR and SSIM values were not suitable for evaluating the synthesis results because of the unregular distribution of the best values for every column. The reason is that the PSNR and SSIM values are focus on the details of the image pattern rather than the similarity of facial likelihood, while the FID is focusing on the latter value with more convincing evaluation. The FID score calculated the distance of 2048 dimensional features for the synthesized image and the original ground truth, and we defined the distance as the similarity metric for the facial images.

Table 1: FID score compared in transformations of different resolution

	16 to 64	32 to 128	64 to 256
Wavelet SRNet [26]	**45.952**	54.533	31.929
Cycle GAN [28]	63.837	32.228	42.218
SRN-Deblur [29]	125.576	41.463	**12.729**
ESRGAN [30]	85.668	30.531	17.709
Multi CinCGAN [31]	228.822	162.034	51.214
ours	51.646	**29.685**	13.431

According to the results shown in Fig. 5, the reason for wavelet-SRNet had a better FID score in "16 to 64" column of Table 1, was because the synthesized image had more wavelet details, which is the main advantage of this method, but the likeness with the ground truth is not so convincible. Our result has shown more stable performance and likeness in all the transformations. SRN, ESRGAN and our method had the best performance, with unnoticeable differences in the high resolution of 256×256, but our method

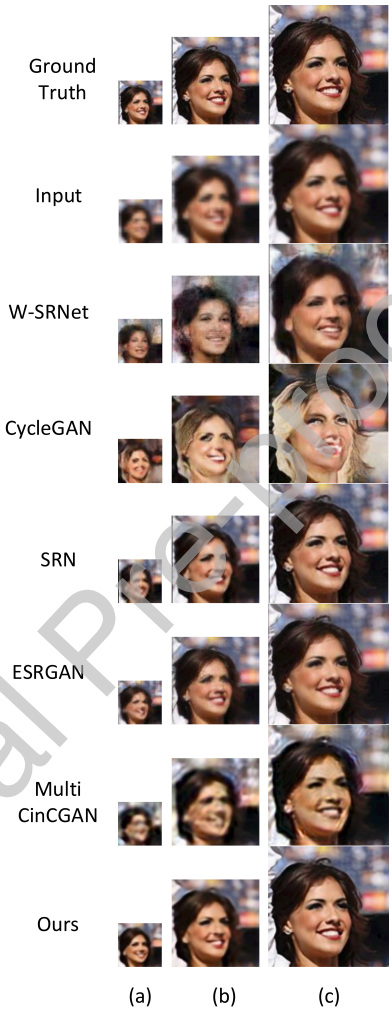


Figure 5: SR results of different target resolutions, the input row is 4 times upscaled of the original LR image by Bicubic-interpolation, column (a) is "16 to 64", column (b) is "32 to 128" and column (c) is "64 to 256" transformation respectively, images in column (c) have been resized to fit the page.

Table 2: PSNR and SSIM values in transformations of different resolution

	16 to 64		32 to 128		64 to 256	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Wavelet SRNet [26]	18.881	0.541	19.011	0.456	22.048	0.587
Cycle GAN [28]	7.464	0.156	7.431	0.159	6.630	0.120
SRN-Deblur [29]	**24.084**	0.755	28.235	0.852	**31.365**	0.885
ESRGAN [30]	22.821	**0.778**	25.854	0.840	29.079	**0.886**
Multi CinCGAN [31]	19.068	0.586	**28.591**	0.663	24.095	0.748
ours	18.487	0.670	21.435	**0.864**	24.326	0.812

had better results in LR situations.

Furthermore, we conducted "16 to 128" transformation with 8 times of upscaling on different methods with the results shown in Table 3 and Fig. 6.

Table 3: FID score compare in 16 to 128 transformation

method	FID
Cycle GAN [28]	59.682
SRN-Deblur [29]	53.800
ours	54.136

According to Table 3 and Fig. 6, the SRN-Deblur method and ours had equivalent upscaling results and the CycleGAN result was degraded in this large upscaling situation. The ESRGAN and MultiCinCGAN methods had unacceptable long training times to conduct with 8 times of scaling factor, so the results were not listed.

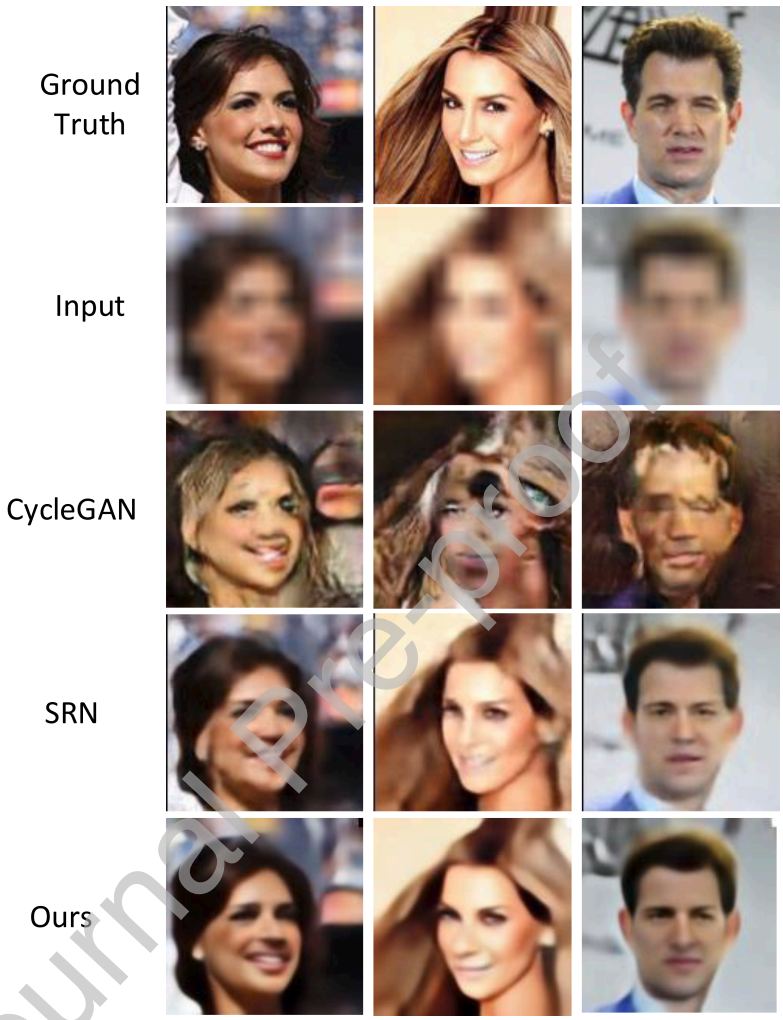


Figure 6: SR results of "16 to 128" transformation with 8 times of upscaling.

To test the robustness of our model, we used the pretrained network weights for the testing of the LFW dataset [32]. We compared 2 sets of transformation: from 32×32 to 128×128 and from 64×64 to 256×256. The FID scores are shown in Table 4.

Trained by the CelebA dataset and tested by the LFW dataset, our net-

Table 4: FID score for LFW datasets in different pixel resolution

method	32 to 128	64 to 256
Bicubic-resize	143.59	43.99
Wavelet SRNet [26]	113.78	54.37
Cycle GAN [28]	75.28	49.34
SRN-Deblur [29]	25.81	**3.82**
ESRGAN [30]	20.63	8.57
Multi CinCGAN [31]	88.24	43.32
ours	**16.95**	4.68

work has shown better SR results and has a smaller FID score compare to the other state of art methods, which means better performance in the transferable and universal ability for facial image SR.

5. Conclusion

This work demonstrated a flexible end-to-end pipeline of getting SR facial images from in-the-wild images that contain different poses and variant resolutions. It can get a high-resolution image from a relatively low resolution with different scale factors. The network had shown good generalization ability and can be used in the transformation of various resolution targets. The stacked upsampling backend of the network had shown a potential possibility of getting bigger upscaling factors.

The stacked architecture of the network has shown good flexibility and expansibility for future application, loss functions and upsampling layers can be varied to fit more circumstances. From the comparison of experimental

results, we got the conclusion that the FID score is a better standard for evaluating facial image synthesized results than the traditional PSNR and SSIM values. From the appearance of synthesized images, we can tell that our results have the best average SR performance in different resolutions and scale factors than the other methods, and the generated images are more realistic.

6. Acknowledgments

This work was supported by National Natural Science Foundation of China Joint Fund (Grant nos. GG2090090072, U1332130, U1713206), 111 Projects (Grant no. B07033), Key research and development plan of Anhui Province (Grant nos. 1704a0902051, 18030901033), Natural Science Foundation of Anhui Province (Grant nos. 1908085ME135, KJ2018A0487).

References

[1] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, M. Nixon, Super-resolution for biometrics: A comprehensive survey, Pattern Recognition 78 (2018) 23–42. doi:10.1016/j.patcog.2018.01.002.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.

[3] A. Bulat, J. Yang, G. Tzimiropoulos, To learn image super-resolution, use a gan to learn how to do image degradation first, in: Proceedings

of the European Conference on Computer Vision (ECCV), 2018, pp. 185–200. doi:10.1007/978-3-030-01231-1_12.

[4] L. Ye, B. Zhang, M. Yang, W. Lian, Triple-translation gan with multi-layer sparse representation for face image synthesis, Neurocomputing 358 (2019) 294–308. doi:10.1016/j.neucom.2019.04.074.

[5] G. Lin, Q. Wu, L. Qiu, X. Huang, Image super-resolution using a dilated convolutional neural network, Neurocomputing 275 (2018) 1219–1230. doi:10.1016/j.neucom.2017.09.062.

[6] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with deep convolutional sufficient statistics, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL: http://arxiv.org/abs/1511.05666.

[7] O. Ronneberger, P. Fischer, 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. doi:10.1007/978-3-319-24574-4_28.

[8] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134. doi:10.1109/CVPR.2017.632.

[9] G. Liu, L. Li, L. Jiao, Y. Dong, X. Li, Stacked fisher autoencoder for sar

change detection, Pattern Recognition 96 (2019) 106971. doi:10.1016/j.patcog.2019.106971.

[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE transactions on pattern analysis and machine intelligence 41 (2019) 1947–1962. doi:10.1109/TPAMI.2018.2856256.

[11] D. Mahapatra, B. Bozorgtabar, R. Garnavi, Image super-resolution using progressive generative adversarial networks for medical image analysis, Computerized Medical Imaging and Graphics 71 (2019) 30–39. doi:10.1016/j.compmedimag.2018.10.005.

[12] Y. Chen, J. Sun, W. Jiao, G. Zhong, Recovering super-resolution generative adversarial network for underwater images, in: International Conference on Neural Information Processing, Springer, 2019, pp. 75–83. doi:https://doi.org/10.1007/978-3-030-36808-1_9.

[13] S. Lian, H. Zhou, Y. Sun, Fg-srgan: A feature-guided super-resolution generative adversarial network for unpaired image super-resolution, in: International Symposium on Neural Networks, Springer, 2019, pp. 151–161. doi:10.1007/978-3-030-22796-8_17.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[15] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738. doi:10.1109/ICCV.2015.425.

[16] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, Fsrnet: End-to-end learning face super-resolution with facial priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2492–2501. doi:10.1109/CVPR.2018.00264.

[17] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2016) 295–307. doi:10.1109/TPAMI.2015.2439281.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, pp. 6626–6637. URL: http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-e

[19] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL: http://arxiv.org/abs/1511.06434.

[20] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: 6th International

Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018. URL: https://openreview.net/forum?id=Hk99zCeAb.

[21] M. Lucic, K. Kurach, M. Michalski, S. Gelly, O. Bousquet, Are gans created equal? a large-scale study, in: Advances in neural information processing systems, 2018, pp. 700–709. URL: http://papers.nips.cc/paper/7350-are-gans-created-equal-a-large-scale-study.

[22] S. Zhu, S. Liu, C. C. Loy, X. Tang, Deep cascaded bi-network for face hallucination, in: European conference on computer vision, Springer, 2016, pp. 614–630. doi:10.1007/978-3-319-46454-1_37.

[23] B. Huang, W. Chen, X. Wu, C.-L. Lin, P. N. Suganthan, High-quality face image generated with conditional boundary equilibrium generative adversarial networks, Pattern Recognition Letters 111 (2018) 72–79.

[24] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797. doi:10.1016/j.patrec.2018.04.028.

[25] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410. doi:10.1109/CVPR.2019.00453.

[26] H. Huang, R. He, Z. Sun, T. Tan, Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1689–1697. doi:10.1109/ICCV.2017.187.

[27] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

[28] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232. doi:10.1109/ICCV.2017.244.

[29] X. Tao, H. Gao, X. Shen, J. Wang, J. Jia, Scale-recurrent network for deep image deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8174–8182. doi:10.1109/CVPR.2018.00853.

[30] 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: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 63–79. doi:10.1007/978-3-030-11021-5_5.

[31] Y. Zhang, S. Liu, C. Dong, X. Zhang, Y. Yuan, Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution,

IEEE transactions on Image Processing 29 (2019) 1101–1112. doi:10.1109/TIP.2019.2938347.

[32] G. Learned-Miller, Labeled faces in the wild: Updates and new reporting procedures, University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003 (2014). URL: http://vis-www.cs.umass.edu/lfw/lfw_update.pdf.

Jijun He now is a Ph.D. candidate in University of Science and Technology of China. He received his M.S. and B.S. degree from the same university in 2010 and 2007 respectively. His research is focus on computer vision and computer graphic.

Jinjin Zheng is a professor in the Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, China. He received his Ph.D. in computer aided geometric modeling from the University of Birmingham, UK, in 1998. His research interests include computer-aided geometric design, computer-aided engineering design, micro electro mechanical systems and computer simulation.

Yuan Shen is a vice professor in the Department of Computer Science, Hefei Normal University, Hefei, China. He received his Ph.D. degree from University of Science and Technology of China, Hefei, China. His research interests include image processing and data mining.

Yutang Guo is a professor in the Department of Computer Science, Hefei Normal University, Hefei, China. He received his Ph.D. degree from Anhui University, Hefei, China. His research interests include image processing and machine learning.

Hongjun Zhou is a senior engineer at the National Synchrotron Radiation Labortory, Hefei, China. She received her M.S. from the University of Central England, Birmingham, UK. Her research interests include mechanical design, micro-electro-mechanical systems and vacuum technology

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name	Affiliation

The authors declare that they have no conflicts of interest.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jijun He: Writing-Original Draft, Methodology, Software. **Jinjin Zheng**: Project administration, Writing-Review & Editing. **Shen Yuan**: Validation, Visualization. ***Yutang Guo****:* Supervision.**:** **Hongjun Zhou**: Conceptualization, Funding acquisition.