Deep Face Super-Resolution with Iterative Collaboration between Attentive Recovery and Landmark Estimation
Supplementary Material

1. More Details on Network Architecture

Here we describe more details of our recurrent networks. Table 1 shows the detailed architecture of the SR branch. Given input LR images, LR features are extracted by G_1 and are subsequently concatenated with the feedback features. Then through G_R, which consists of a convolutional layer, an attentive fusion module and a recurrent SR module, the obtained features are used as both the feedback signals and the features for the following generation. Finally, SR images are recovered by the generation layers G_2 and the addition operation. G_2 is comprised of a deconvolutional layer with a kernel size of 8 and a convolutional layer.

Besides, Table 2 presents the details of our recurrent alignment branch. A_1 and A_2 are the pre-processing and post-processing blocks, which have the same architecture as those in [4] except that the batch normalization layers are removed. The recurrent hourglass module has similar architecture to the single hourglass module in [4]. Differently, the input and output of A_R both include two components. The input is obtained by concatenating the pre-processing feature with the feedback feature while the output is split into two parts, a feedback feature and a feature for the final landmark estimation.

Table 1. Detailed architecture of the recurrent SR branch.

Layer	Output size
Input I^{LR}	$16 \times 16 \times 3$
Conv (G_1)	$16 \times 16 \times 192$
PixelShuffle (G_1)	$32 \times 32 \times 48$
Concatenation	$32 \times 32 \times 96$
Conv (G_R)	$32 \times 32 \times 48$
Attentive Fusion (G_R)	$32 \times 32 \times 48$
Recurrent SR Module (G_R)	$32 \times 32 \times 48$
Deconv (G_2)	$128 \times 128 \times 48$
Conv (G_2)	$128 \times 128 \times 3$
Addition	$128 \times 128 \times 3$
Output I^{SR}	$128 \times 128 \times 3$

Table 2. Detailed architecture of the recurrent alignment branch.

Layer	Output size
Input I^{SR}	$128 \times 128 \times 3$
A_1	$32 \times 32 \times 256$
Concatenation	$32 \times 32 \times 512$
Conv (A_R)	$32 \times 32 \times 512$
Recurrent HourGlass (A_R)	$32 \times 32 \times 512$
Split	$32 \times 32 \times 256$
	$32 \times 32 \times 256$
A_2	$32 \times 32 \times 68$
Output L	$32 \times 32 \times 68$

2. User Study

We conduct a user study to further evaluate the visual quality of the super-resolved images. We randomly select 30 images from the testing set of CelebA [3] and display the corresponding SR results of our DICGAN, FSRGAN [1], PFSR [2] and the HR images in a random order. 39 human raters are asked to rank these four versions of images in terms of perceptual satisfaction. The results are shown in Figure 1. As expected, most of the HR images are regarded as the best among the four versions. Moreover, our DICGAN obtains much more votes of rank-1 and rank-2 than FSRGAN and PFSR, which means the proposed method outperforms the state-of-the-art face SR methods by a large margin. We observe that PFSR scores the worst among three FSR methods. We think the reason is that PFSR mainly focuses on well-aligned face images. Hence when the input faces are with large variations of pose and rotation, PFSR fails to present satisfactory SR results.

3. Visual Results

In Figure 2 and Figure 3 (**the next pages**), we present more qualitative comparison with state-of-the-art FSR methods including RDN [5], FSRNet [1], FSRGAN [1] and

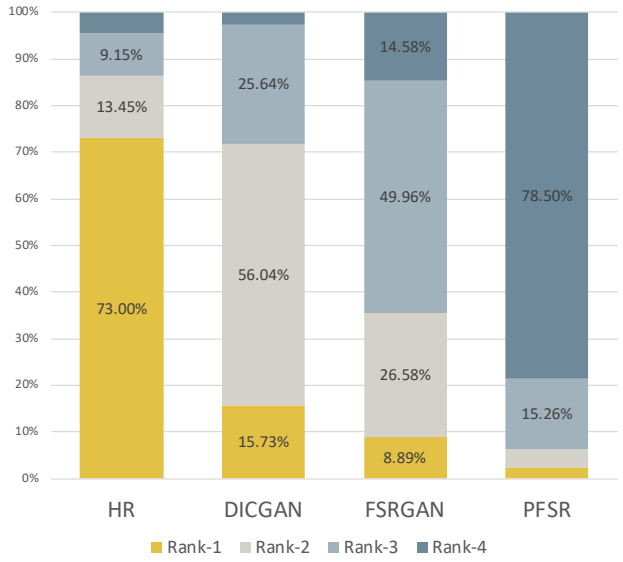

Figure 1. Results of the user study. Our method performs better than state-of-the-art FSR methods in recovering perceptual-pleasant face images.

PFSR [2]. The results demonstrate the effectiveness of our proposed method.

References

[1] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, pages 2492–2501, 2018. 1

[2] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019. 1, 2

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 1

[5] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 1

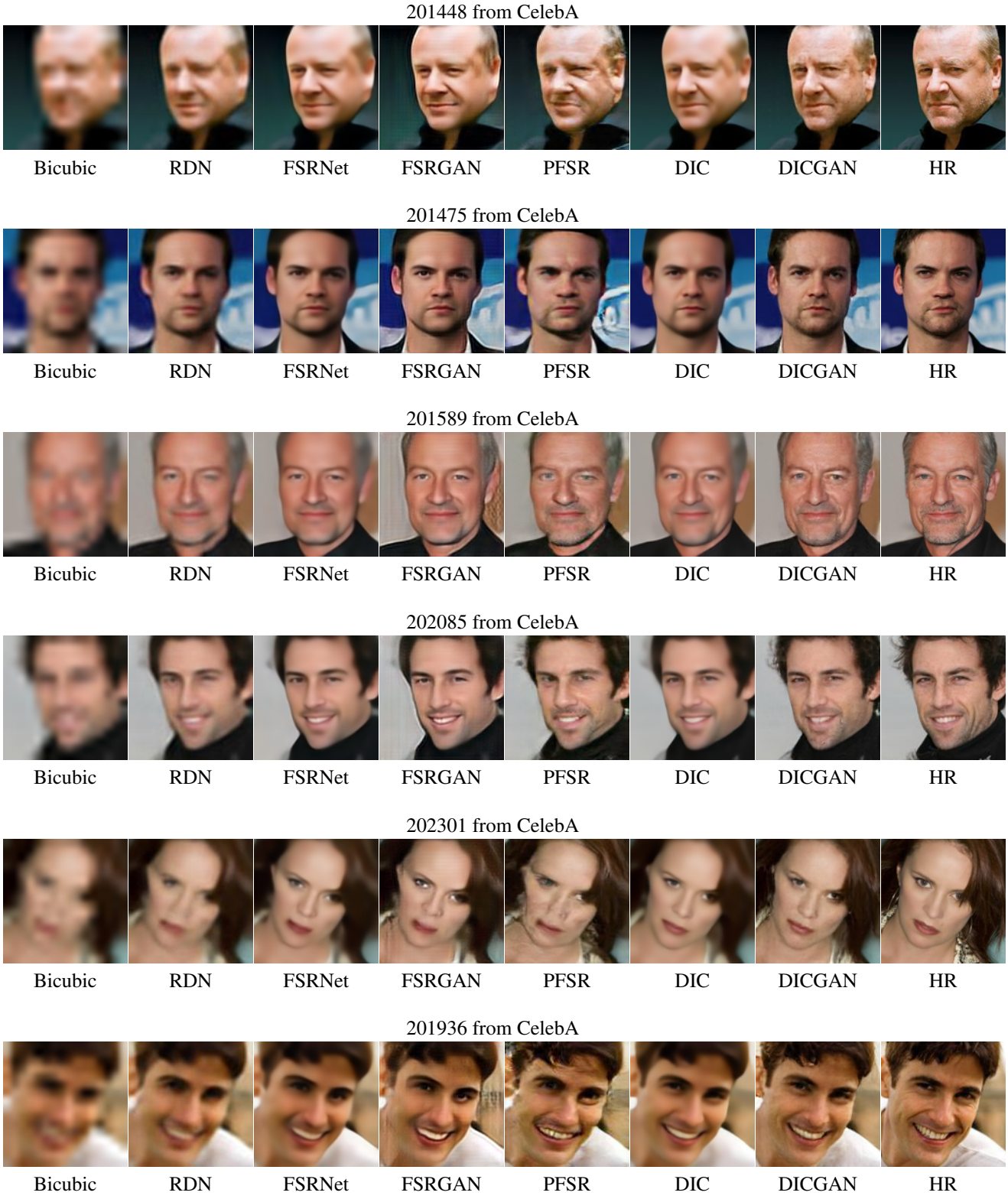

Figure 2. Qualitative comparison with state-of-the-art face super-resolution methods.

Figure 3. Qualitative comparison with state-of-the-art face super-resolution methods.