Super-Identity Convolutional Neural Network for Face Hallucination -Supplementary Material-

Kaipeng Zhang[1][0000−0001−6105−6532], Zhanpeng Zhang[2][0000−0002−1709−4176], Chia-Wen Cheng[1,3][0000−0002−5753−6530], Winston H. Hsu[1*][0000−0002−3330−0638], Yu Qiao[4][0000−0002−1889−2567], Wei Liu[5][0000−0002−3865−8145], and Tong Zhang[5][0000−0002−5511−2558]

[1] National Taiwan University, Taipei, Taiwan
whsu@ntu.edu.tw
[2] SenseTime Group Limited, China
[3] The University of Texas at Austin, Texas, USA
[4] Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, China
[5] Tencent AI Lab, China

In this supplementary document, we present additional results to support our work. First, we introduce the network structure of our face recognition model (i.e. CNN_R). Then, we present more visual comparisons with state-of-the-art algorithms. After that, we evaluate our method for 24×28 inputs with 4× upscaling factor. Besides, we also evaluate the performance of different hallucination network (i.e., $CNNR_H$) architectures. At last, we evaluate the performance on un-aligned faces.

1 Face Recognition Network Architecture

The network structure of CNN_R is shown in Tab. 1. It is a ResNet-like [2] architecture consisting of 27 convolutional layers. The identity representation is extracted from the fully connected layers (i.e., FC1). We use A-Softmax [5] loss to train this network and more details are introduced in Sec. 3.3 on our paper.

2 Visual Comparisons

In this section. We present supplementary visual comparisons (as shown in Fig. 1) over different methods on hallucination test dataset for Sec. 4.4 on our paper. It is clear that our method achieves the best visual results and more details for this experiment are introduced in Sec. 4.4 on our paper.

3 Evaluation on Higher Input Resolution

For more comprehensive analysis, in this section, we trained our model for 24×28 inputs with 4× upscaling factor. Specifically, we modify the hallucination network (i.e., CNN_H) by removing the first DB, DeConv and Conv layers. As shown

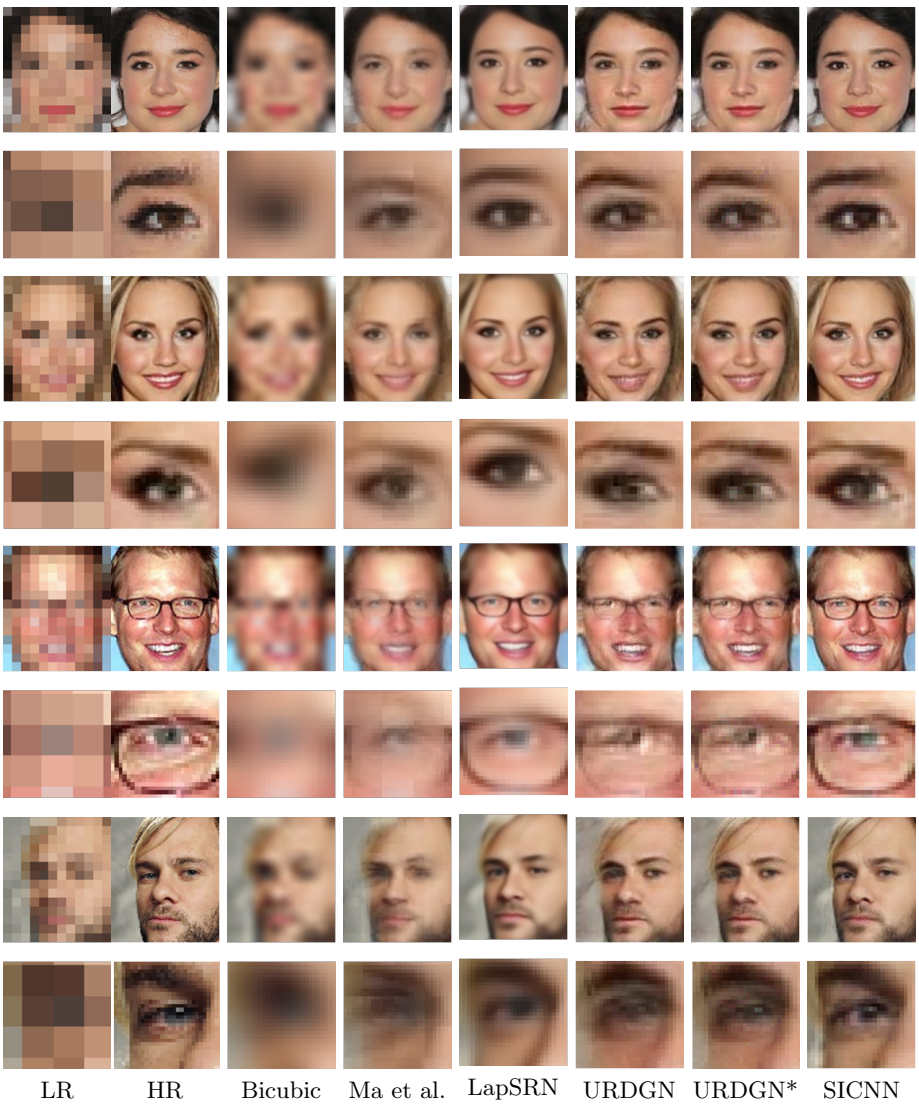

Fig. 1. Hallucination visual comparison with state-of-the-art methods (Ma et al. [6], LapSRN [4], URDGN [8] and URDGN* trained with the additional perceptual loss, see Sec 4.4 on our paper for more details) on hallucination test dataset. It is clear that our method achieves the best hallucination visual quality. Please *zoom in* for better comparison.

Layer Name	Output Size	Structure
Input	96×112	-
Conv1a	94×110	3×3, 64, pad 0
Conv1b	92×108	3×3, 64, pad 0
Avepool1	46×54	3×3, stride 2
Residual_block1	46×54	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 1$
Conv2	44×52	3×3, 128, pad 0
Avepool2	22×26	3×3, stride 2
Residual_block2	22×26	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
Conv3	20×24	3×3, 256, pad 0
Avepool3	10×12	3×3, stride 2
Residual_block3	10×12	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 5$
Conv4	8×10	3×3, 512, pad 0
Avepool4	4×5	3×3, stride 2
Residual_block4	4×5	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$
FC1	512	4×5, 512

Table 1. The architecture for our face recognition CNN (CNN_R). It follows the residual block structure [2]. We use PReLU [1] activation function after each convolution layer. The output of FC1 is the identity representation.

in Fig. 2, our method performs very well visual quality in higher resolution inputs with 4x upscaling factor.

For identity recovery and identity recognizability evaluation, our method also achieves very good results: Average identity similarity: 0.8868, LFW accuracy: 99.21%, YTF accuracy: 94.86%, which are very close to the performance on HR faces.

4 Evaluation on Different Hallucination Architectures

In this section, we try different hallucination architectures (i.e. CNN_H) including different growth rates with the same depth (i.e. numbers of layers) and the network structure used in URDGN [8]. From Tab. 2 and Fig. 3, we observe that different network structures show slight differences of hallucination performances while different training approaches and methods are more important for this semantic face hallucination (see Sec. 4.5 and Sec. 4.6 on our paper).

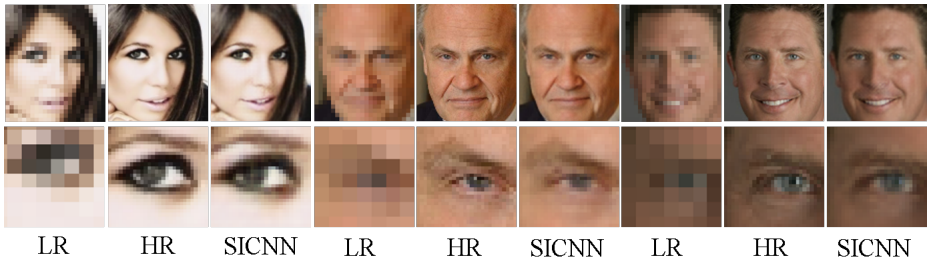

Fig. 2. Hallucination visual results for 24×28 inputs with 4× upscaling factor. Please *zoom in* for better comparison.

Method	UR-DGN*	SICNN ($k = 32$)	SICNN ($k = 16$)	SICNN ($k = 8$)	SICNN (UR-DGN)
Identity Similarity	0.5267	0.5978	0.5959	0.5911	0.5803
LFW Acc.	98.01%	98.25%	98.26%	98.23%	98.20%
YTF Acc.	93.54%	93.82%	93.84%	93.82%	93.78%

Table 2. Quantitative identity recovery and identity recognizability comparison with URDGN* and our method of different hallucination architectures. k denotes the growth rate in the dense block [3]. URDGN* is trained by the additional perceptual loss. See Sec 4.4 on our paper for more details.

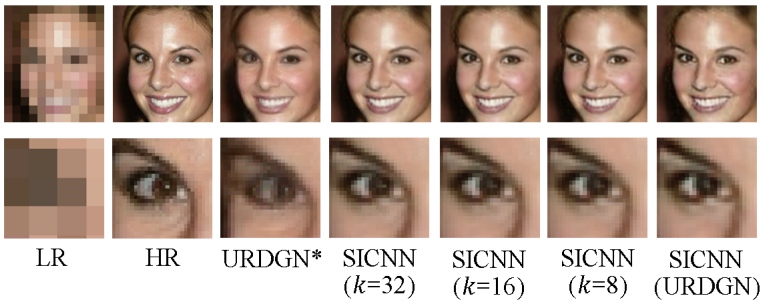

Fig. 3. Hallucination visual comparison with URDGN* and our method of different hallucination architectures (i.e. CNN_H). k denotes the growth rate in the dense block (i.e., DB) [3]. Please *zoom in* for better comparison. URDGN* is trained by the additional perceptual loss. See Sec 4.4 on our paper for more details.

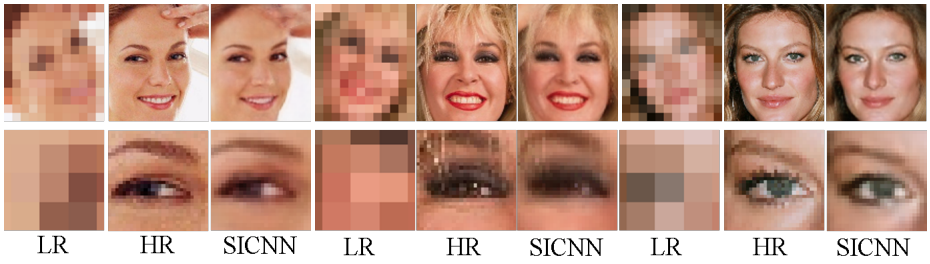

Fig. 4. Hallucination visual examples on un-aligned faces using our method with the spatial transform component. Please *zoom in* for better comparison.

5 Evaluation on Un-Aligned Faces

For the fair comparison with other face hallucination methods, we use aligned faces for evaluation on our paper. In this section, we try to add an additional component, spatial transform network, like TDN [9] to solve un-aligned faces.

Following TDN, we train our model using aligned HR faces and un-aligned LR faces which are cropped from the images based on the detection results. And, we use three additional spatial transform networks (Conv(3,64) + Conv(3,64) + FC(128) + FC(6)) after each dense block (i.e. DB).

From Fig. 4, we observe that our method with the spatial transform component can achieve good performance on un-aligned faces. In addition to alignment, it is worth noting that our proposed training approaches and loss function can be easily incorporated into other methods for different situations, such as noisy or blur faces [10, 7].

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. pp. 1026–1034 (2015)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
3. Huang, G., Liu, Z., Weinberge, r.K., Maaten, L.v.d.: Densely connected convolutional networks (2017)
4. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. CVPR (2017)
5. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition (2017)
6. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. PR **43**(6), 2224–2236 (2010)
7. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: CVPR. pp. 251–260 (2017)
8. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: ECCV. pp. 318–333 (2016)

9. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: AAAI. vol. 2, p. 3 (2017)
10. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: CVPR. pp. 3760–3768 (2017)