Blind Face Restoration via Deep Multi-scale Component Dictionaries

(Supplementary Material)

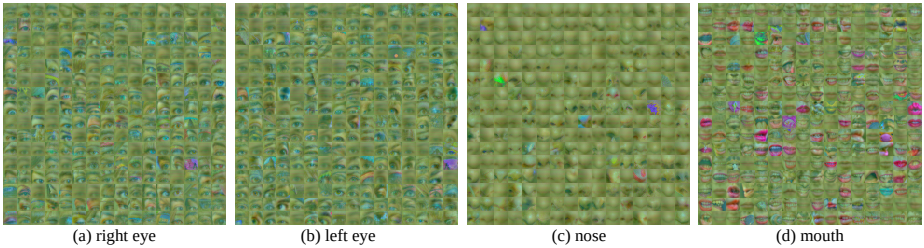
In the supplemental materials, we first give the analyses about running times in Section A. Then, some feature visualizations including dictionaries and confidence score are provided in Section B. Results of additional variant of our DFDNet about the cluster is analyzed in Section C. The network architecture of our DFDNet is presented in details in Section D. More restoration results on real-world low-quality images are demonstrated in Section E. Section F demonstrates more visual results on $\times 4$ and $\times 8$. Finally, the visual results of DFDNet variants are shown in Section G.

A Running Time

It takes nearly 2 days to generate the four scale dictionaries with 256 clusters and 33 *ms* (including 1 *ms* for facial landmark detection by Dlib [4]) for DFDNet to handle a 256×256 image in inference phase, which is comparable with GFRNet [5] (31 *ms*) and faster than GWAINet [1] (94 *ms*).

B Feature Visualization

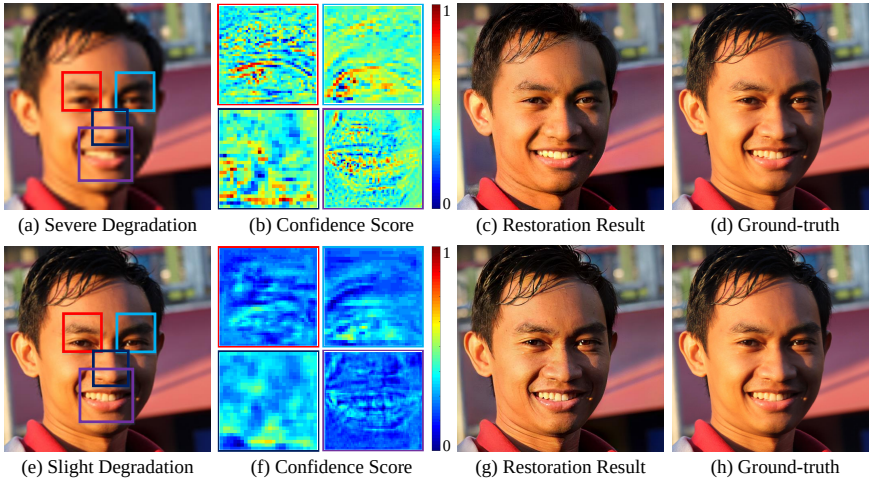
In this section, we conduct two visual experiments to show what the features represent. For the first one, we visualize the conducted dictionary clusters. To achieve this goal, we fix the parameters of VggFace model and only update the input noise by minimizing the distance between the network output and each cluster, which is inspired from the neural style transfer task [2]. Some visual results from scale-1 dictionary are shown in Fig. A. We can observe that each cluster has different texture, shape (open or close) and pose, which nearly covers the most common face structure.



(a) right eye (b) left eye (c) nose (d) mouth

Figure A: Feature visualization of our conducted component dictionaries on scale-1.

For the second one, we visualize the confidence map for the same image with different degradation level. We synthesize two degraded images with slight and severe degradation. Visualization of the confidence score on scale-1 is shown in Fig. B. It can be seen that when the degradation is severe, the confidence values are high, indicating that more features should come from the selected cluster, and vice versa. The confidence score can well learn the differences between the input and selected cluster, and thereby adaptively fuse the dictionary features to the images with different degradation level.



(a) Severe Degradation (b) Confidence Score (c) Restoration Result (d) Ground-truth

(e) Slight Degradation (f) Confidence Score (g) Restoration Result (h) Ground-truth

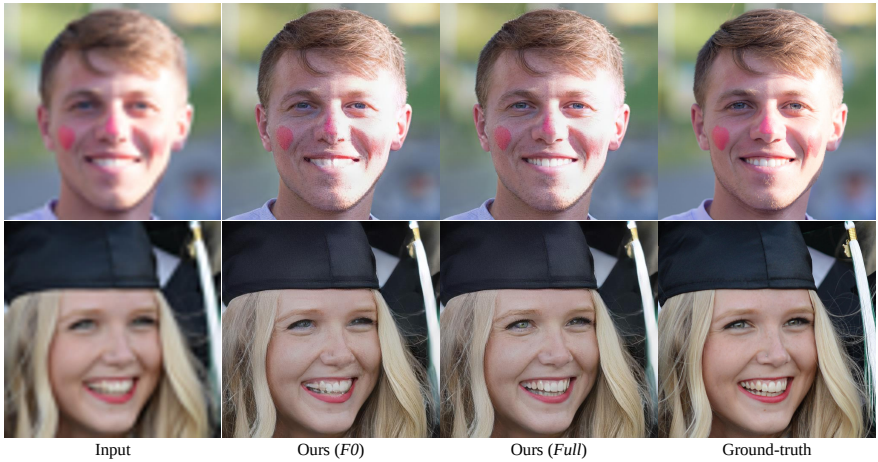
Figure B: Visualization of confidence score for images with different degradation level.

C Results of Filling 0 on the Matched Cluster

In this section, we explore the effectiveness of our proposed dictionaries in the restoration process by directly filling 0 on the matched cluster, which is defined as Ours (*F0*). It can be seen from Table A that the quantitative performance is severely degraded by a large margin, mainly indicating that our proposed dictionary does play an important role in the reconstructing process. Visual results of Ours (*F0*) contain obviously artifacts and the component regions only have coarse structure, while Ours (*Full*) have more realistic details (see Fig. C).

Table A: Quantitative comparisons on Ours (*F0*) and Ours (*Full*).

Methods	×4			×8		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours(*F0*)	25.12	.890	.251	22.97	.838	.259
Ours(*Full*)	27.54	.923	.114	23.73	.872	.239



Input Ours (*F0*) Ours (*Full*) Ground-truth

Figure C: Comparisons on the variant of our DFDNet by filling 0 on the matched cluster.

D Network Architecture of DFDNet

Our DFDNet adopts VggFace as the encoder, whose parameters are fixed in the training phase. Details of VggFace are shown in Table B. The remaining architecture of DFDNet is shown in Table C. Our Confidence Score for each component is constructed as [Conv., SN, LReLU, Conv., SN, Sigmoid], which is not shown here. Conv. (DiConv.) (o, k, s) denotes a convolutional (dilated convolutional) layer, where o, k and s are the output dimension, kernel size and stride (dilation rate), respectively. SN represents spectral normalization, and LReLU (c) is leaky ReLU with negative slope c. Maxpooling (d) and Bilinear Upsample (d) denote the way of downsample and upsample with scale factor d.

Table B: Details of VggFace.

Input	Degraded Input I^d (256×256)
Scale-1 Feature Extraction	Conv.(64,3,1), ReLU() Conv.(64,3,1), ReLU() Maxpooling(2) Conv.(128,3,1), ReLU() Conv.(128,3,1)
Scale-2 Feature Extraction	ReLU(), Maxpooling(2) Conv.(256,3,1), ReLU() Conv.(256,3,1), ReLU() Conv.(256,3,1), ReLU() Conv.(256,3,1)
Scale-3 Feature Extraction	ReLU(), Maxpooling(2) Conv.(256,3,1), ReLU() Conv.(512,3,1), ReLU() Conv.(512,3,1), ReLU() Conv.(512,3,1)
Scale-4 Feature Extraction	ReLU(), Maxpooling(2) Conv.(512,3,1), ReLU() Conv.(512,3,1), ReLU() Conv.(512,3,1), ReLU() Conv.(512,3,1)
Output	F_{vgg}

Table C: Details of DFDNet.

Input	F_{vgg}
4 Dilated Resblocks	DiConv.(512,3,1), SN, LReLU(0.2) DiConv.(512,3,1), SN DiConv.(512,3,2), SN, LReLU(0.2) DiConv.(512,3,2), SN DiConv.(512,3,3), SN, LReLU(0.2) DiConv.(512,3,3), SN DiConv.(512,3,4), SN, LReLU(0.2) DiConv.(512,3,4), SN
Decoder	Conv.(512,3,1), SN, LReLU(0.2) Dilated ResBlock Conv.(512,3,1), SN Bilinear Upsample(2) Conv.(256,3,1), SN, LReLU(0.2) Dilated ResBlock Conv.(256,3,1), SN Bilinear Upsample(2) Conv.(128,3,1), SN, LReLU(0.2) Dilated ResBlock Conv.(128,3,1), SN Bilinear Upsample(2) Conv.(64,3,1), SN, LReLU(0.2) Dilated ResBlock Conv.(64,3,1), SN, LReLU(0.2) Conv.(3,3,1), Tanh()
Output	Result $\hat{I}$ (256×256)

E More Visual Results on Real LQ Images

Among these competing methods, only GFRNet [5] can well handle blind face restoration. Thus we mainly compare DFDNet with it on restoring real-world LQ images that were collected from *Google Image* with resolution lower than 80×80. For fair comparison, we also conduct the identity-belonging HQ reference for GFRNet, which is not required in our DFDNet. The restoration results are shown in Fig. D. One can see that our DFDNet can generate plausible and realistic details, which are superior to GFRNet [5]. Moreover, we also retrain our DFDNet512 to handle high-resolution images. Results are shown in Fig. E. We can see that our DFDNet512 can also perform well on high-resolution results and generalize to different degraded images, *i.e.*, old photos, diverse poses, ages, *etc*, which are mainly attributed to the deep multi-scale component dictionaries.

Figure D: Visual comparison on real-world LQ images. Close-up in the right bottom is the guidance for GFRNet [5]. Best view it by zooming in the screen.

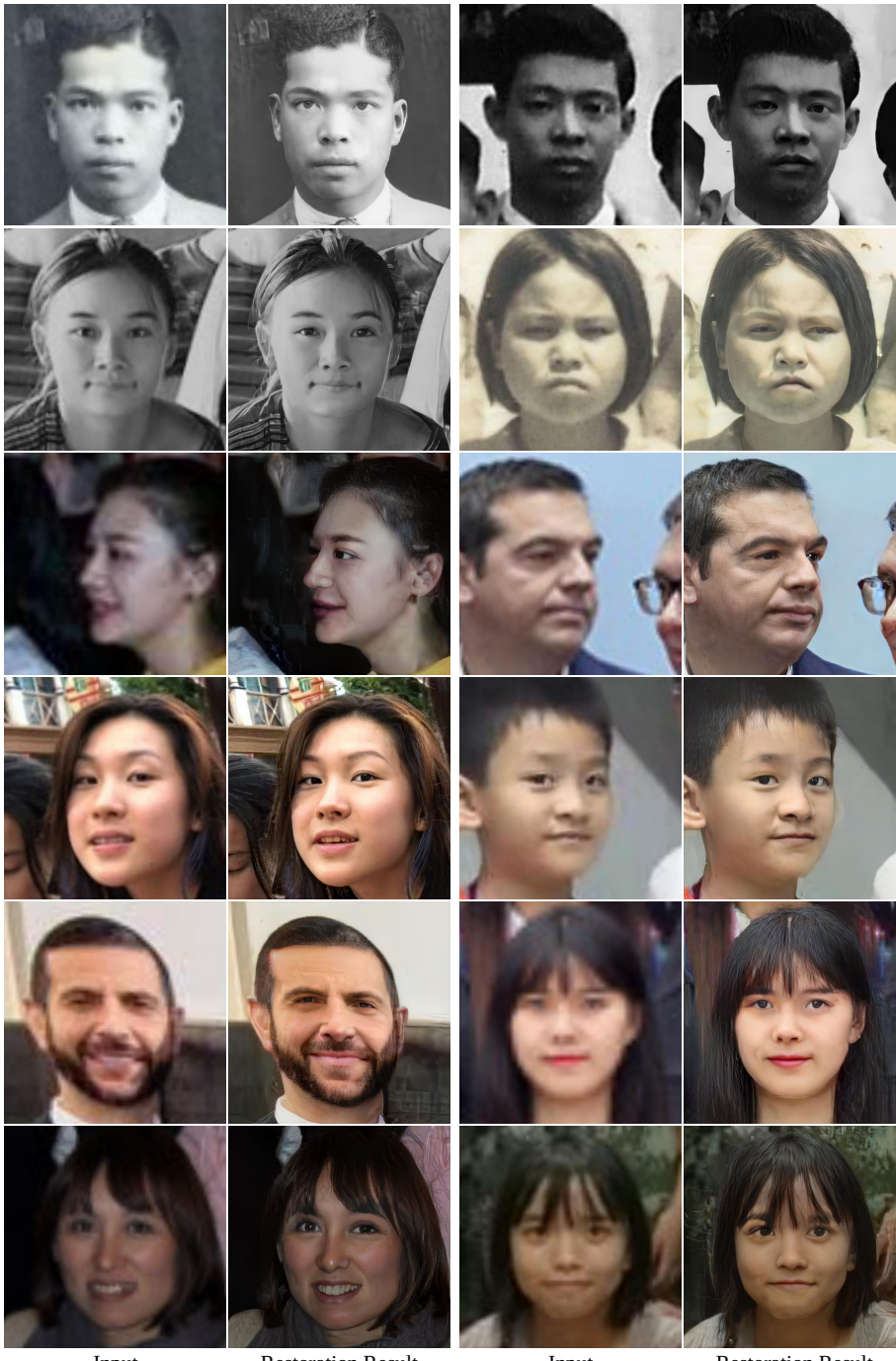


Figure E: High-resolution restoration results of DFDNet512 on real-world LQ images.

F More Visual Results on ×4 and ×8

In this section, we report the visual restoration results on selected methods (*i.e.*, *RCAN [7], *ESRGAN [6], WaveletSR [3], GWAINet [1], and GFRNet [5]) with top quantitative performance to give more visual comparisons. Figs. F and G show that our DFDNet outperforms to all the competing methods in generating rich and realistic details, especially in the semantic component regions.

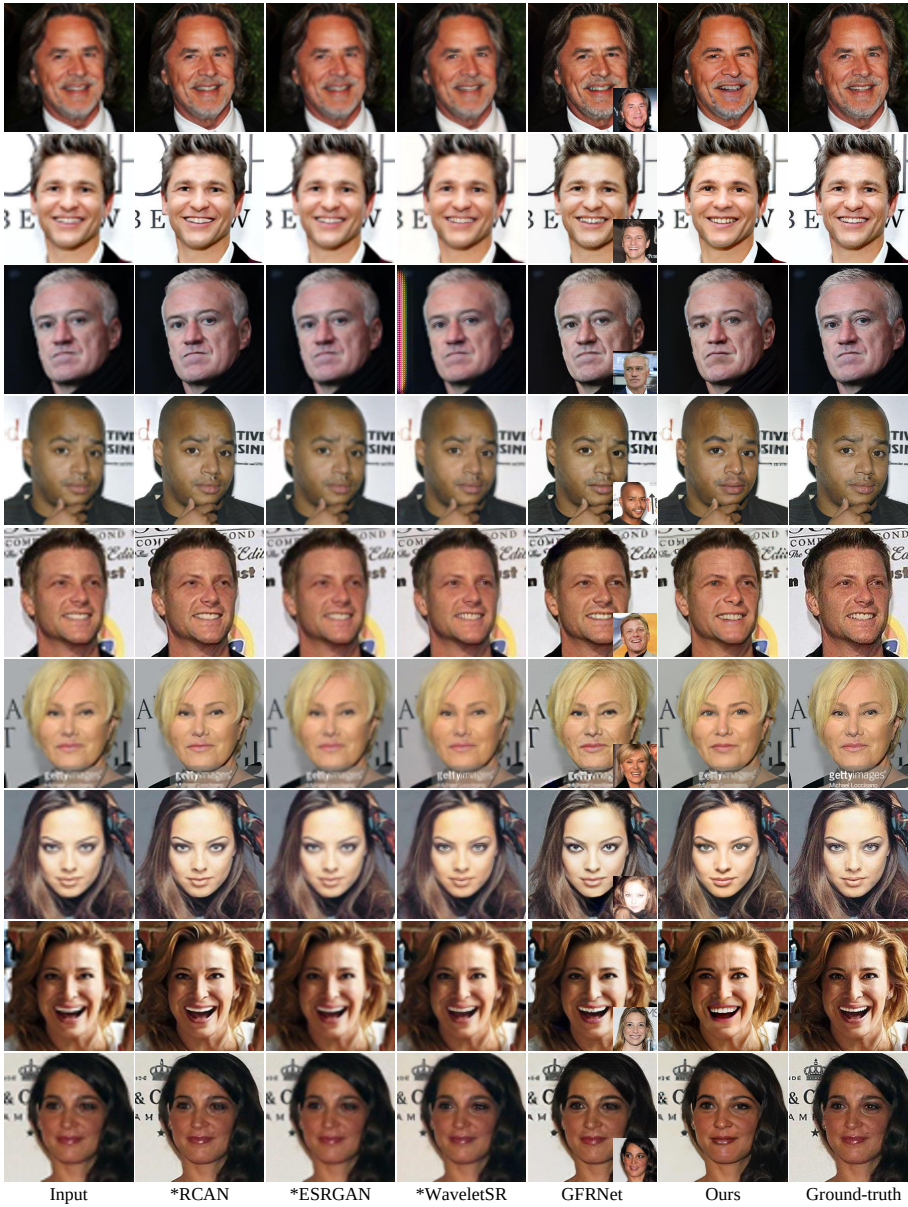


Figure F: More restoration results on ×4 SR compared with the competing methods.

Figure G: More restoration results on $\times 8$ SR compared with the competing methods. Best view it by zooming in the screen.

G More Visual Results on DFDNet Variants

In this section, we demonstrate more visual results in ablation study, including the effect of cluster numbers, the progressive manner of multi-scale component dictionaries in DFT blocks, the component AdaIN, as well as the Confidence Score in Figs. H and I.

Figure H: More results of our DFDNet with different numbers of dictionary clusters.

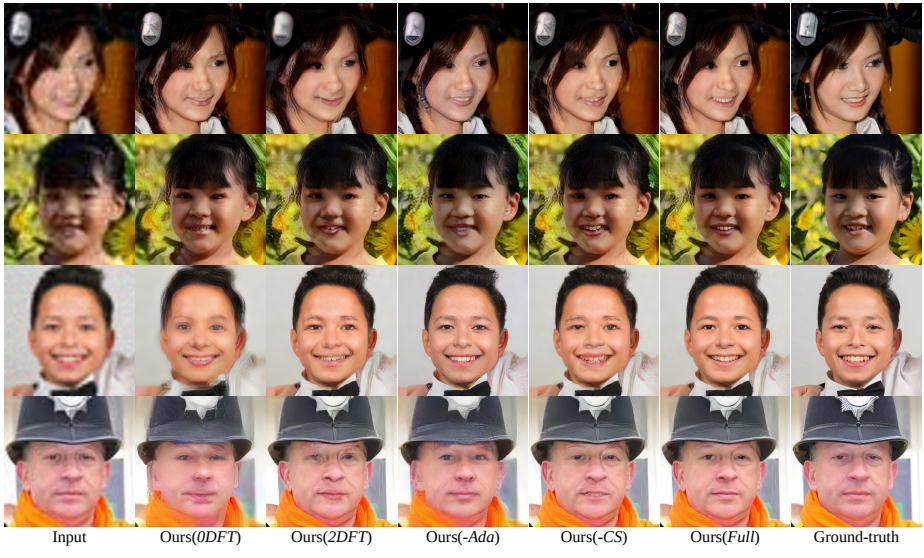


Figure I: More results of our DFDNet variants.

References

1. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: CVPRW (2019)
2. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
3. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: ICCV (2017)
4. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research (2009)
5. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: ECCV (2018)
6. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCVW (2018)
7. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)