WaveFace: Authentic Face Restoration with Efficient Frequency Recovery

Supplementary Material

A. Experimental Details

A.1. Conditioning scheme

In Sec. 5.2, we discuss the effectiveness of several widely-used conditioning schemes. In this section, we will provide more details about "AdaGN" and "Cross-Att.".

AdaGN. Adaptive group normalization (AdaGN) conditions the denoising network at the latent level, where the timestep and latent vector are incorporated into each residual block after group normalization:

$$\text{AdaGN}(\boldsymbol{h}, t, \boldsymbol{z}_{ll}) = \boldsymbol{z}_s(t_s\text{GroupNorm}(h) + t_b), \quad (16)$$

where $\boldsymbol{z}_s \in \mathbb{R}^d = \text{Affine}(\boldsymbol{z}_{ll})$ refers to identity features of LQ images $\boldsymbol{x}_0$ extracted by ArcFace [5] after the affine transformation. $(t_s, t_b) \in \mathbb{R}^{2\times d} = MLP(\psi(t))$ is the output of a Multi-Layer Perceptron (MLP) with a sinusoidal encoding function ψ. d denotes the dimension of embeddings.

Cross-Att. Cross-attention (CA) layer can improve the model performance via the inner relationship between inputs from multiple modalities [7, 25]. In the paper, cross-attention is adopted to complement the denoised HQ sample at each timestep $\boldsymbol{y}_t$ with its LQ counterpart $\boldsymbol{x}_0$:

$$\text{CA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad (17)$$

$$Q = W_Q \cdot \boldsymbol{x}_0, \quad K = W_K \cdot \boldsymbol{y}_t, \quad V = W_V \cdot \boldsymbol{y}_t, \quad (18)$$

where W_Q, W_K, W_V are learnable projection matrices [29]. CA$(\cdot)$ refers to the cross-attention operation, which is the pixel-wise dot product between HQ feature maps V and corresponding attention scores.

B. Additional Experimental Results

B.1. Discussion on Non-Reference Metric

In Sec. 5.3, we report the performance of state-of-the-art (SOTA) methods on synthetic and real-world datasets in terms of FID score. In this section, we further consider two commonly used non-reference metrics: NIQE [21] and NRQM [20]. The results are reported in Tab. B.1 and Tab. B.2, respectively. As can be seen, despite the superiority of our method in terms of identity preserving and facial detail recovery, it shows inferior performance on these non-reference metrics. We also notice that images restored by some methods even unreasonably beat ground truth (GT) on CelebA-Test.

To explore the reason behind this, the qualitative comparison is conducted between a generative prior-based method (VQFR [9]) and diffusion model-based methods (DiffFace [39] and ours). As shown in Fig. B.1, although the image generated by VQFR provides better sharpness, it contains many artifacts. For example, the hair and eyelashes present an unnatural woolen texture, which deteriorates the image's authenticity. On the contrary, diffusion model-based methods can yield more photorealistic faces with human-like textures.

Methods	CelebA	LFW	WebPhoto	WIDER
Input	14.114	8.575	12.664	13.498
GPEN [37]	7.760	3.853	4.498	4.105
GFP-GAN [31]	4.171	3.954	4.248	3.880
VQFR [9]	3.775	3.574	3.606	3.054
RestoreFormer [34]	4.436	4.145	4.459	3.894
CodeFormer [43]	4.680	4.520	4.708	4.165
DR2 [35]	4.998	4.736	6.159	5.171
DifFace [39]	4.500	4.220	4.666	4.688
DM	4.898	4.784	4.860	4.988
WaveFace	4.421	4.133	4.383	4.963
GT	4.373	-	-	-

Table B.1. Quantitative comparisons on synthetic and real-world datasets (-Test) in terms of **NIQE↓**.

Methods	CelebA	LFW	WebPhoto	WIDER
Input	6.042	2.810	2.044	1.358
GPEN [37]	8.514	8.482	7.584	8.112
GFP-GAN [31]	7.985	7.782	7.750	7.990
VQFR [9]	8.657	8.564	8.457	8.792
RestoreFormer [34]	8.495	8.572	8.133	8.537
CodeFormer [43]	8.339	8.217	7.457	8.370
DR2 [35]	6.906	6.049	4.423	5.219
DifFace [39]	7.724	6.322	4.929	4.728
DM	7.121	7.125	7.091	7.167
WaveFace	7.732	7.753	6.749	6.541
GT	7.909	-	-	-

Table B.2. Quantitative comparisons on synthetic and real-world datasets (-Test) in terms of **NRQM↑**.

To further investigate the performance of diffusion models on these two metrics, we randomly sample the same number of images as the corresponding dataset with a benchmark pre-trained diffusion model[1] and evaluate the quality by NIQE and NRQM. Some generated images are depicted in Fig. B.2 and quantitative results are denoted as "DM" in Tab. B.1 and Tab. B.2. As shown in both tables, even images generated by the benchmark diffusion

[1] https://github.com/openai/improved-diffusion

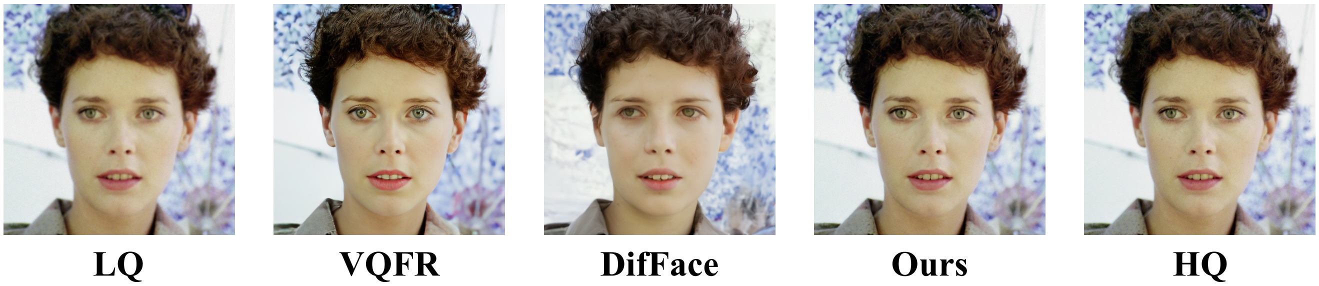

Figure B.1. Qualitative comparison between generative prior-based and diffusion model-based methods.

Figure B.2. Samples generated by benchmark diffusion model.

model underperform those by GAN-based methods on these non-reference metrics, which challenges the common finding that diffusion models beat GANs in image synthesis [7, 23, 25].

Both qualitative and quantitative results illustrate that these two non-reference metrics cannot well represent the performance of BFR methods. More investigations are called to study the appropriate evaluation metrics for BFR.

B.2. Degradation types

Apart from the classical degradation model (Eq. (15)), we adopt the second-order degradation process proposed by RealESRGAN [32], where classical degradations are applied repeatedly to mimic real-world degradation. Following the settings in Sec. 5.1, we train both LCD and HFR modules on FFHQ [14] with RealESRGAN degradations. To evaluate the model, a corresponding evaluation set is synthesized based on 3000 CelebA-HQ images, namely CelebA-Test-RESR.

The performances of SOTA methods and our method on CelebA-Test-RESR and real-world datasets are reported in Tab. B.3 and Tab. B.4, respectively. As can be seen, the model trained on data with RealESRGAN degradations outperforms that on classical degradations, which demonstrates that RealESRGAN degradations can better imitate real-world degradations. The qualitative comparison between SOTA methods and ours is illustrated in Fig. B.4. Our method (WaveFace) is able to deliver authentic results with both identity information and fine-grained details well preserved. For example, our method restores more details of the earrings in 2nd column.

Table B.3. Quantitative comparison on **CelebA-Test-RESR** for blind face restoration. "Deg." refers to the angle between identity embeddings of restored images and HQ counterparts. Best performances are **highlighed**.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	Deg.↓
Input	18.886	0.449	0.574	48.968	39.910
VQFR [9]	18.167	0.516	0.459	11.911	35.104
DifFace [39]	18.321	0.540	0.489	12.353	43.773
WaveFace	**19.126**	**0.576**	**0.436**	**11.336**	**32.863**

B.3. Denoising Process Visualization

We illustrate the denoising process of the diffusion model used in our low-frequency conditional denoising (LCD) module and the unconditional one adopted in DifFace [39] in Fig. B.3. Both models are trained on FFHQ [14] and

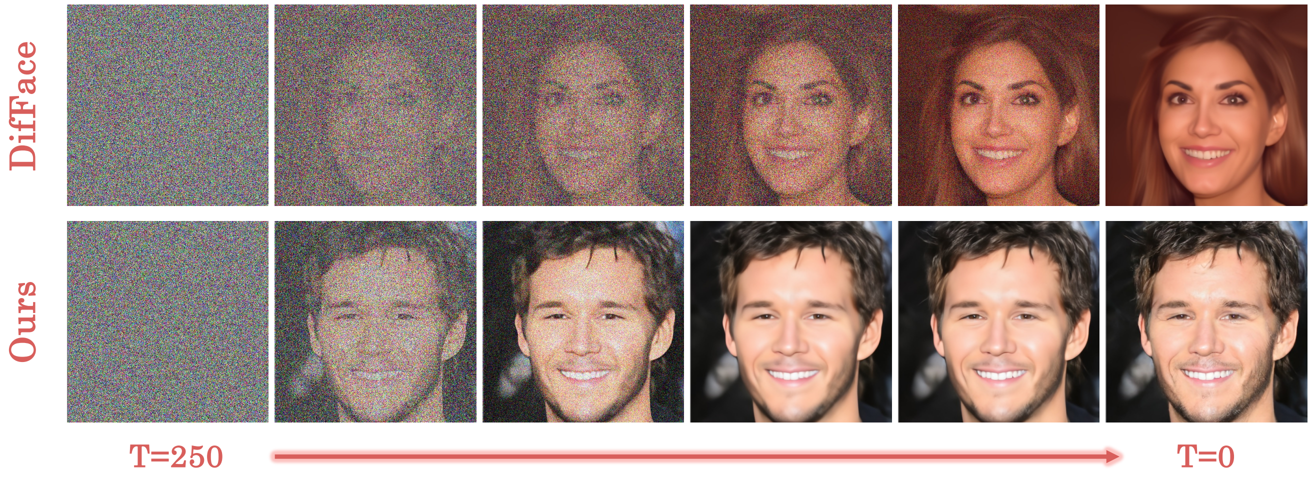

Figure B.3. Visualization of denoising process of the unconditional diffusion model adopted in DifFace [39] and our conditional one.

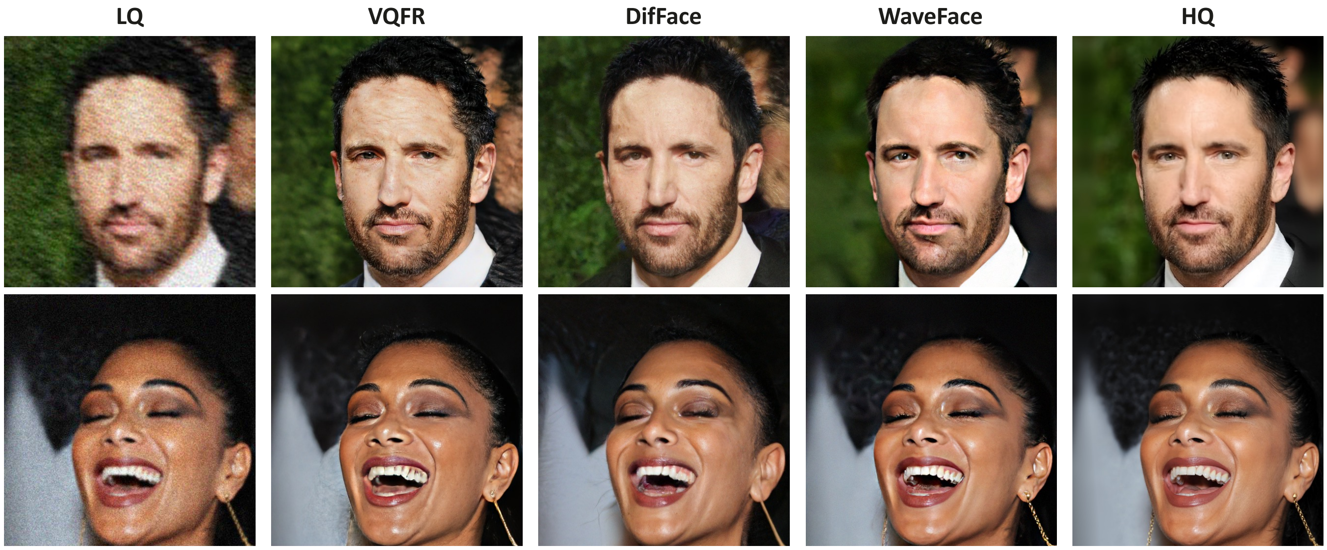

Figure B.4. Qualitative comparison on CelebA-Test-RESR.

Table B.4. Quantitative comparisons on three **real-world datasets (-Test)** in terms of **FID**$_{\downarrow}$. Best performances are **highlighed**.

Methods	**LFW**	**WebPhoto**	**WIDER**
Input	124.974	170.112	199.961
WaveFace (Classical)	43.175	81.525	36.913
WaveFace (RealESRGAN)	**46.711**	**78.438**	**35.750**

tested on CelebA-Test [13]. We take DDIM [28] as the sampling scheme, which takes 250 steps to sample an image from a pure Gaussian noise. It can be easily observed that conditional DM (Ours) achieves a faster sampling convergence due to the prior knowledge provided by LQ counterparts.

B.4. More Qualitative Comparisons

More qualitative comparison results are illustrated in Fig. B.5 $\sim$ Fig. B.8. For CelebA-Test (Fig. B.5), with increasing degradation applied (from top to bottom), our method can generate authentic facial images while well preserving the identity. Qualitative comparison on real-world datasets: LFW-Test (Fig. B.6), WebPhoto-Test(Fig. B.7) and WIDER-Test (Fig. B.8) shows that our method (last column) can restore photorealistic images without destroying style and color of the original image. Meanwhile, more fine-grained facial details are recovered such as the texture of the beanie (Row 2 in Fig. B.6), beard (Row 2 in Fig. B.7), and wrinkles (Row 1 in Fig. B.8).

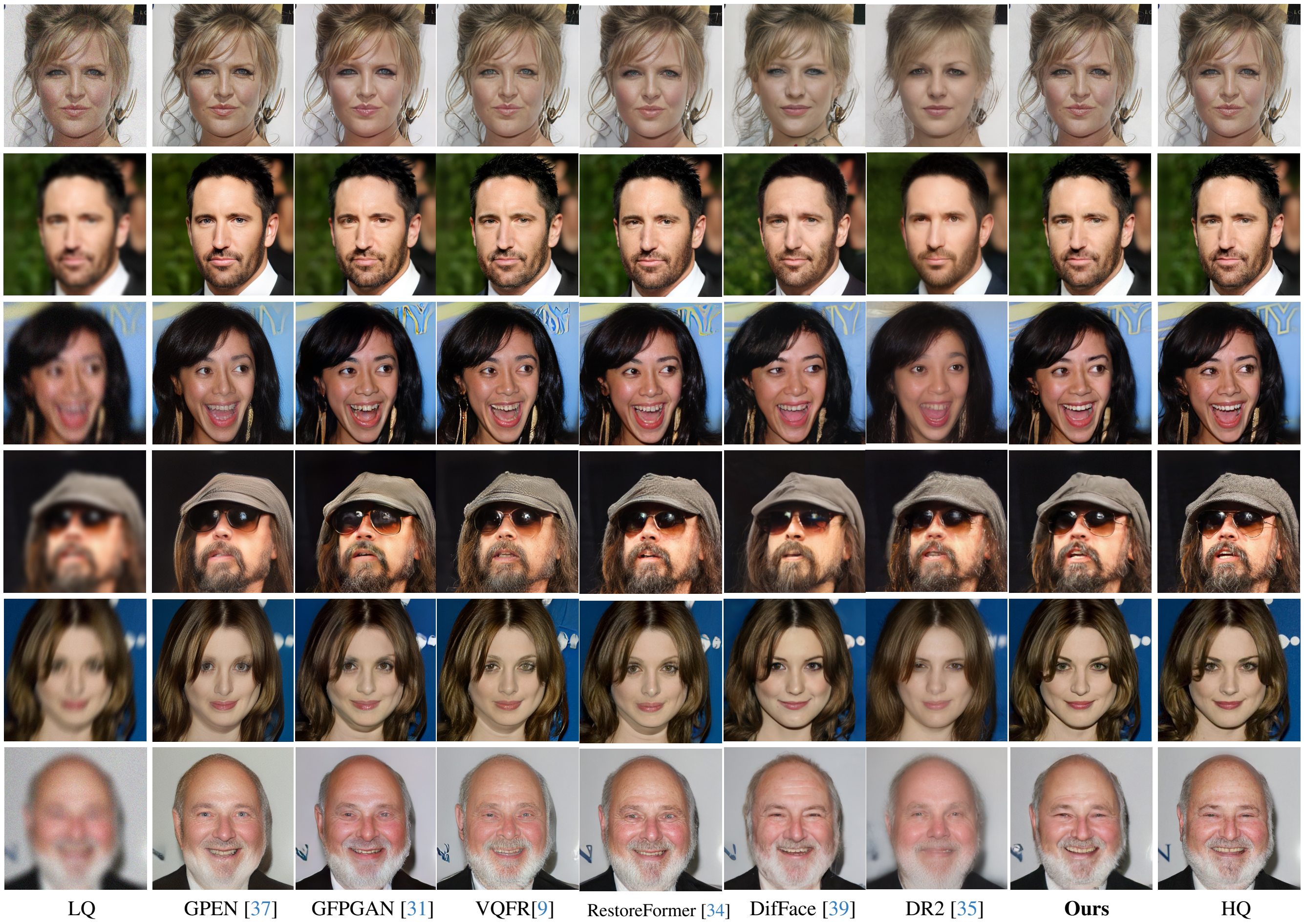

Figure B.5. More qualitative comparison results on **CelebA-Test**. (Zoom in for best view).

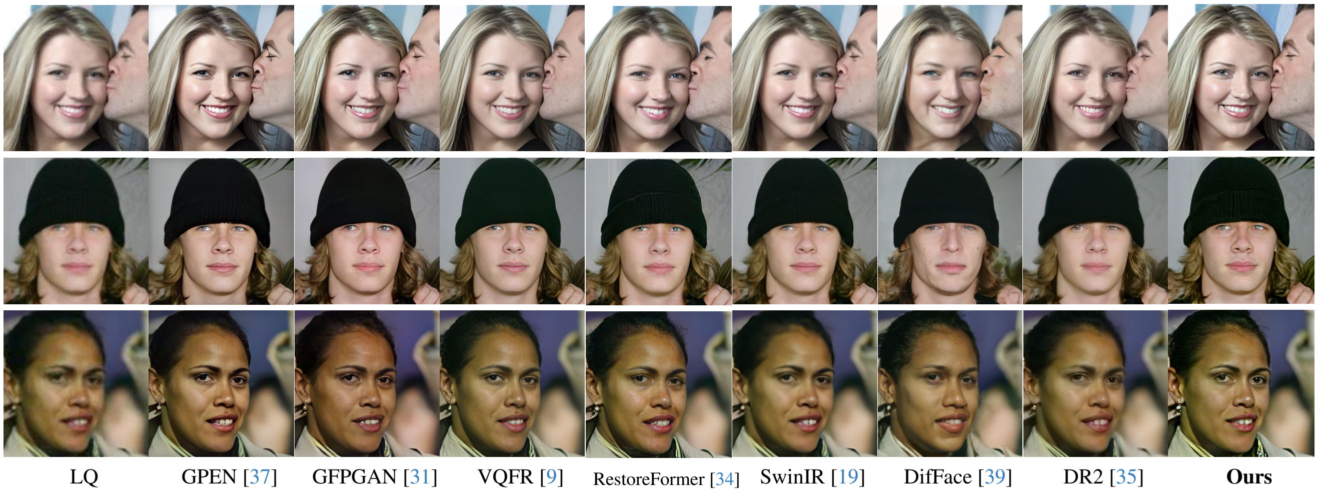

Figure B.6. More qualitative comparison results on **LFW-Test**. (Zoom in for best view).

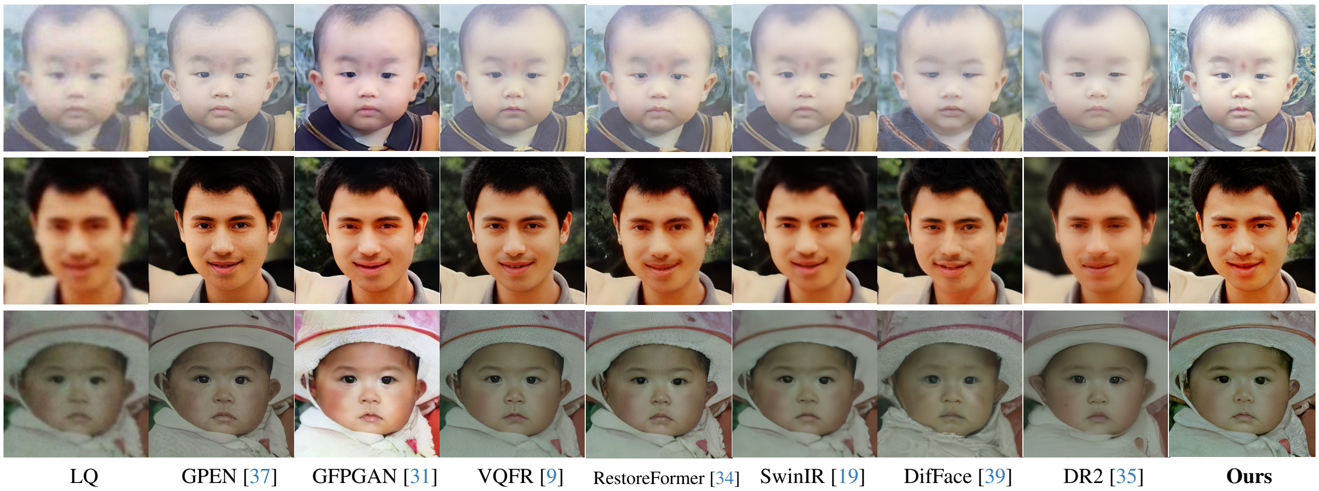

Figure B.7. More qualitative comparison results on **WebPhoto-Test**. (Zoom in for best view).

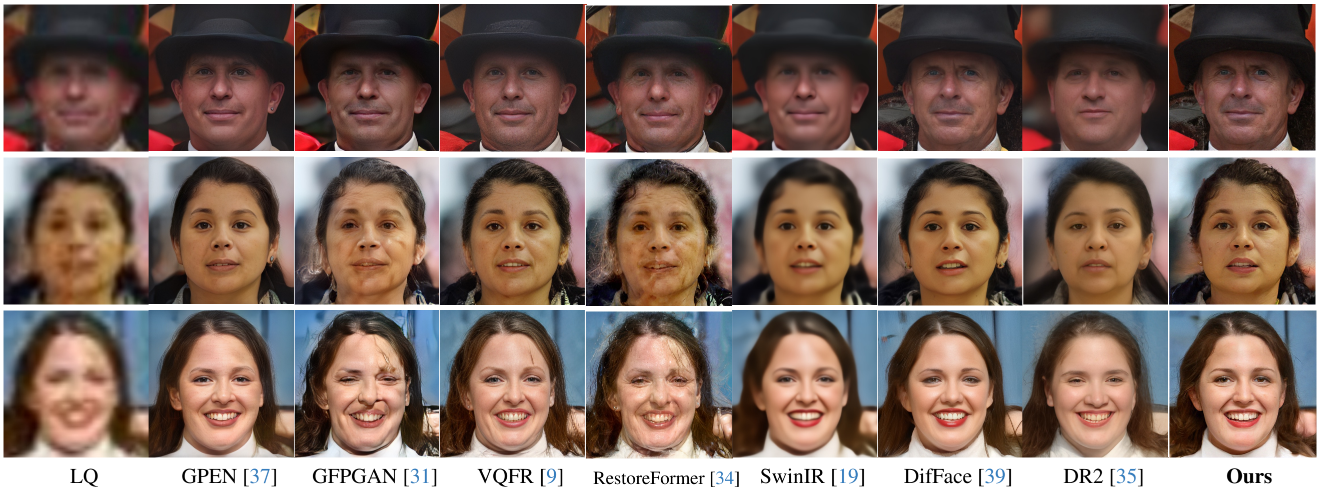

Figure B.8. More qualitative comparison results on **WIDER-Test**. (Zoom in for best view).