OSDFace: One-Step Diffusion Model for Face Restoration
Supplementary Material

Jingkai Wang[1*], Jue Gong[1*], Lin Zhang[1], Zheng Chen[1]
Xing Liu[2], Hong Gu[2], Yutong Liu[1†], Yulun Zhang[1†], Xiaokang Yang[1]
[1]Shanghai Jiao Tong University, [2]vivo Mobile Communication Co., Ltd

A. Overall

This supplementary material provides additional results to support the main manuscript. First, in Section B, we analyze the parameter size of VRE and inference time. Next, in Section C, we present experiments that integrate VRE into OSEDiff [9]. These experiments demonstrate VRE's strong image understanding capabilities. In Section D, we validate our method on downstream face recognition tasks. The results show that our method outperforms others when used as a pre-processing step. Additionally, we analysis the limitation and future work in Section E. Finally, in Section F, we provide more visual comparisons with state-of-the-art methods.

B. Parameters and Inference Time

Table 1 clearly shows that OSDFace achieves high inference speed and low computational cost compared to other one-step diffusion models. The VRE prompt embedder in OSDFace significantly reduces the parameter count and MACs. This reduction is notable when compared to the prompt embedder used in OSEDiff [9], *i.e.*, DAPE [10] with CLIP text encoder.

Additionally, generating text embeddings from input images does not conflict with generating latent vectors through a VAE encoder. Therefore, we can introduce a parallel mechanism that could speed up both OSEDiff [9] and OSDFace. Using parallel acceleration, our OSDFace could further reduce inference time by 14% on top of its fast performance. All tests are conducted on an NVIDIA A6000 GPU.

C. Integrating VRE into OSEDiff

The existing representative one-step diffusion (OSD) image restoration model, OSEDiff [9], does not focus on face restoration tasks. In order to assess its applicability to face restoration, we retrained it using the same dataset and experimental settings as OSDFace, resulting in OSEDiff*. Furthermore, we integrated the proposed VRE into OSEDiff*, creating the enhanced model OSEDiff*+VRE.

As shown in Tab. 2, Tab. 3, and Fig. 1, OSEDiff*+VRE performs well in both quantitative metrics and visual quality. The incorporation of VRE significantly reduces information loss during the image-text-embedding process, ensuring more accurate data representation. Visual results indicate that OSEDiff*+VRE prevents common issues like gender misclassification and unwanted artifacts. Additionally, it reliably captures subtle facial expressions from the input images. Besides, the IQA metrics demonstrate a competitive advantage by consistently reducing the distribution differences from the reference data. These experimental results demonstrate that our proposed VRE substantially enhances face restoration performance, particularly when applied to OSD models.

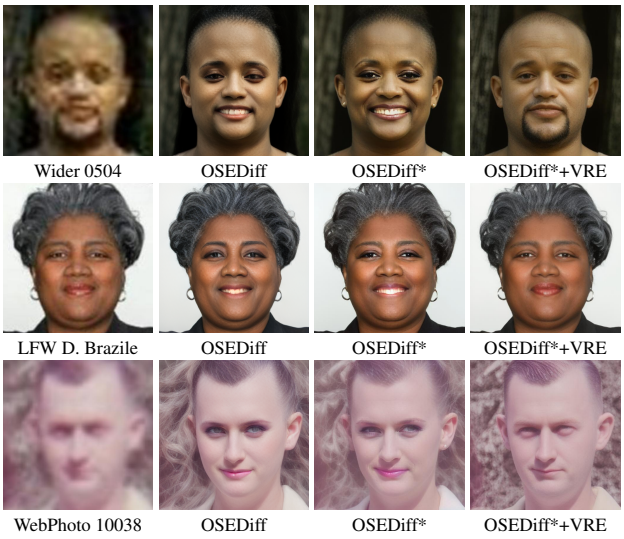

Figure 1. Visual comparisons of various versions of OSEDiff [9]. OSEDiff*+VRE shows enhanced visual quality.

Methods	Prompt Embedder		Inference Time (ms)	
	Param (M)	MACs (G)	Serialized	Parallelized
OSEDiff [9]	353.41	141.45	130.75	125.54
OSDFace (ours)	28.63	99.47	119.00	102.83

Table 1. Complexity comparison during inference. "parallelized" refers to the parallel execution of the prompt embedder and VAE encoder, while "serialized" denotes a fully sequential execution approach. We provide the number of parameters (Param), multiply-accumulate operations (MACs), and time during inference. All models are evaluated with 512×512 input image.

*Equal contribution.
†Corresponding authors.

Methods	Wider-Test					LFW-Test					WebPhoto-Test				
	C-IQA↑	M-IQA↑	MUSIQ↑	NIQE↓	FID↓	C-IQA↑	M-IQA↑	MUSIQ↑	NIQE↓	FID↓	C-IQA↑	M-IQA↑	MUSIQ↑	NIQE↓	FID↓
OSEDiff [9]	0.6298	0.4951	70.559	4.9388	50.274	0.6326	0.5037	73.401	4.7196	57.800	0.6457	0.5108	72.593	5.2611	117.510
OSEDiff*	0.6193	0.4752	69.101	5.0869	47.883	0.6186	0.4879	71.707	4.8002	51.048	0.6254	0.4823	69.816	5.3253	109.236
OSEDiff*+VRE	0.6637	0.4834	68.259	5.0490	41.490	0.6608	0.5015	70.826	4.8956	46.911	0.6410	0.4646	66.912	5.5233	95.566
OSDFace (ours)	0.7284	0.5229	74.601	3.7741	34.648	0.7203	0.5493	75.354	3.8710	44.629	0.7106	0.5162	73.935	3.9864	84.597

Table 2. Quantitative comparison on real-world datasets with one-step diffusion methods. C-IQA stands for CLIPIQA, and M-IQA stands for MANIQA. The best and second best results are colored with red and blue, respectively.

Methods	LPIPS↓	DISTS↓	MUSIQ↑	NIQE↓	Deg.↓	LMD↓	FID(FFHQ)	FID(HQ)↓
OSEDiff [9]	0.3306	0.2170	71.467	5.1241	67.390	6.4141	73.484	37.210
OSEDiff*	0.3496	0.2200	69.981	5.3280	67.403	7.4082	81.362	37.131
OSEDiff*+VRE	0.3368	0.2420	69.089	5.3241	63.758	6.5365	67.785	36.356
OSDFace (ours)	0.3365	0.1773	75.640	3.8840	60.071	5.2867	45.415	17.062

Table 3. Quantitative comparison on the synthetic CelebA-Test dataset with one-step diffusion methods. The best and second best results are colored with red and blue, respectively.

Figure 2. Visualization of the atmospheric turbulence [1] range from 20,000 to 40,000.

D. Validation on Face Recognition

Face restoration, as a fundamental low-level vision task, could enhance downstream face recognition tasks to achieve better performance. We use the LFW [4] dataset as a benchmark for comparison, which includes 3,000 positive pairs and 3,000 negative pairs. Following DAEFR [7], we evaluate the face recognition accuracy using the ArcFace [2] model under different degradation levels. Specifically, we employ unseen atmospheric turbulence degradation [1] to simulate diverse degradation levels, with propagation lengths ranging from 20,000 to 40,000, as illustrated in Fig. 2.

The experimental results in Fig. 3 demonstrate the superior performance of our method across various degradation levels. As degradation severity increases, our method significantly improves precision at the same recall level. The ROC curve shows that OSDFace makes fewer errors at specific true positive rates. Besides, OSDFace widens the gap between positive and negative predictions, thereby improving classifier performance. These findings indicate that our method provides substantial enhancements to downstream face recognition tasks.

E. Limitations and Future Work

We briefly analyze the limitations and future work. **(1)** Color shift: OSDFace sometimes over-enhances contrast or saturation in less degraded regions, causing color shifts in restored face images. Although AdaIN [5] can fix this during inference, we aim for an end-to-end, color shift-free restoration. Future work will explore content-aware color regularization to improve color preservation. **(2)** Texture in complex regions: OSDFace struggles with realistic skin textures and fine details in complex regions like limbs or fingers. This arises from the model's focus on face features, with limited training data for non-facial parts with similar skin textures. Future work will explore semantic information extraction and domain-specific priors to improve the handling of these areas. **(3)** Generalization to low-degradation images: OSDFace was not trained on minimal degradation images but still shows some generalization. However, finer skin texture restoration remains a focus, requiring higher resolution input and output faces, HD training data, and texture-sensitive architectures.

F. Additional Visual Comparisons

These comparisons demonstrate that our proposed OSDFace generates high-quality faces and effectively preserves identities, even with severely degraded input images. Compared to other methods, OSDFace more accurately recovers finer details and produces more realistic faces. To illustrate these advantages further, we select various representative images with unique characteristics, which can be regarded as different face categories. These images are briefly analyzed below.

Synthetic dataset. Visualized results are presented in Fig. 4, Fig. 5, and Fig. 6. Compared to other methods, OSDFace produces more natural-looking restorations with greater detail. This is especially evident in the hair, whether long, short, straight, or curly. Additionally, our method effectively restores occluded regions, such as an arm covering the mouth or bangs obscuring the eyes. For profile views, OSDFace naturally recovers facial contours. In some ground truth images with blurred backgrounds, OSDFace performs well, even achieving higher quality and greater detail than the original HQ images. In scenarios with complex backgrounds, many VQ-based methods, such as VQFR [3], CodeFormer [13], and DAEFR [7], fail to restore natural backgrounds. These methods often produce wallpaper-like outputs, exhibit color distortions, or even blend the person's clothing with the background. In contrast, OSDFace, which combines VQ Dict and diffusion model, successfully generates harmonious faces.

Real-world dataset. More visual comparisons on real-world datasets are shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10. Our OSDFace demonstrates strong capabilities in detail generation and boundary distinction. Some images contain multiple closely positioned faces, such as image 0026 in the Wider-Test and Damon Stoudamire in the LFW-Test. Our method successfully restores each individual face. In Wider 0003, only OSDFace successfully generates complete glasses and clearly separates the arm from the face. For faces with varying skin tones, our method consistently maintains the realism of the images. Furthermore, our approach accurately restores facial accessories, including patterns on hats (Wider 0026), bandages (Daniel Osorno in LFW-Test), and earrings (Wider 0173). In old photo restoration scenarios, our OSDFace also effectively handles unknown degradations.

References

[1] Nicholas Chimitt and Stanley H Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated zernike coefficients. *Optical Engineering*, 2020. 2

[2] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 4

[3] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022. 2, 5, 6, 7, 8, 9, 10, 11

[4] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. 2, 4

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[6] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. In *ECCV*, 2024. 5, 6, 7, 8, 9, 10, 11

[7] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin CK Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In *ICLR*, 2024. 2, 5, 6, 7, 8, 9, 10, 11

[8] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE TPAMI*, 2023. 5, 6, 7, 8, 9, 10, 11

[9] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024. 1, 2, 5, 6, 7, 8, 9, 10, 11

[10] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 1

[11] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 5, 6, 7, 8, 9, 10, 11

[12] Zongsheng Yue and Chen Change Loy. DifFace: Blind Face Restoration with Diffused Error Contraction . *IEEE TPAMI*, 2024. 5, 6, 7, 8, 9, 10, 11

[13] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 2, 5, 6, 7, 8, 9, 10, 11

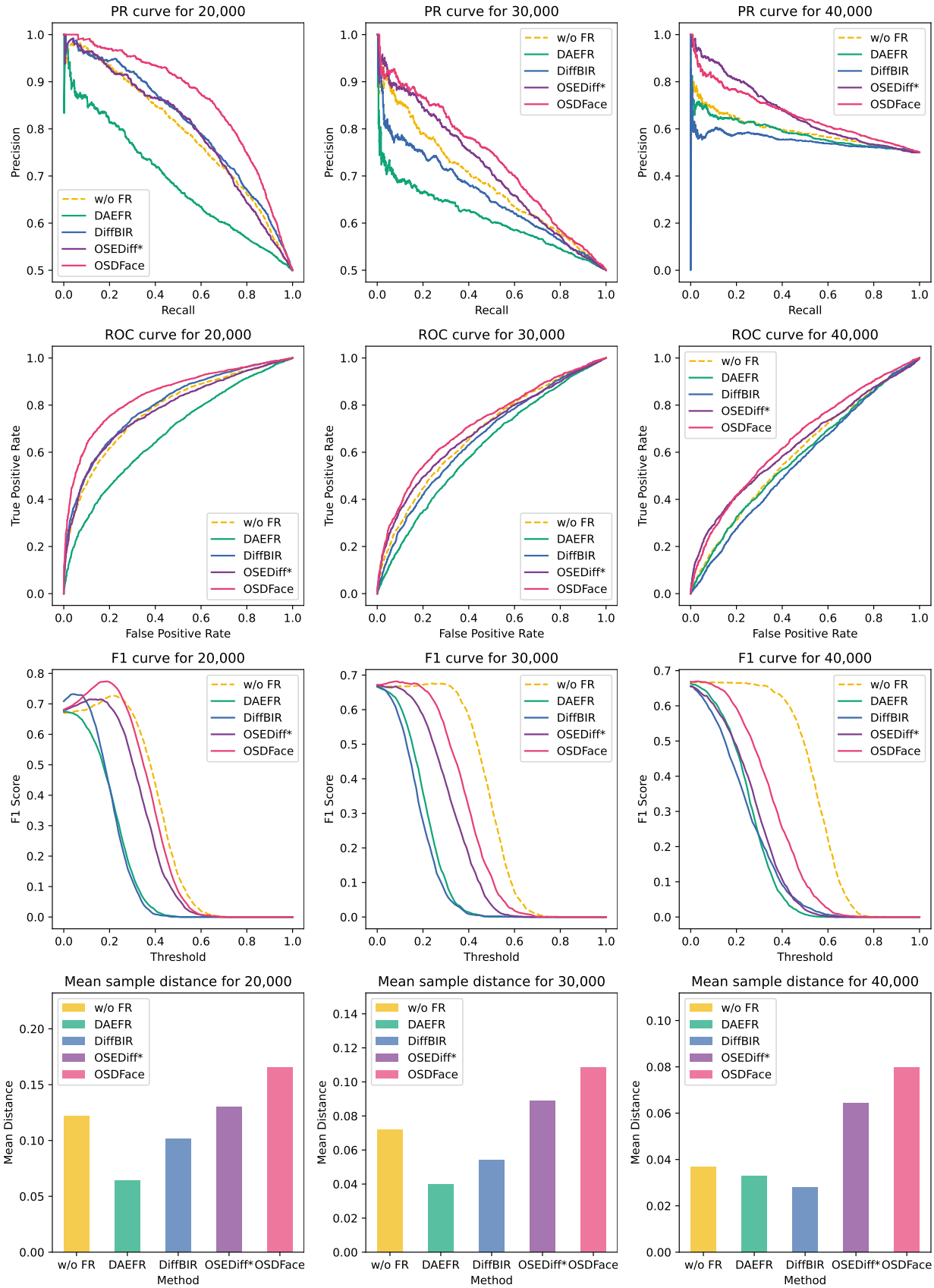

Figure 3. Quantitative results on the LFW dataset [4] for face recognition using the official ArcFace [2] MS1MV3 R50 model. The evaluated metrics include precision-recall (PR) curves, receiver operating characteristic (ROC) curves, F1 scores, and mean sample distance histograms. The mean sample distance is defined as the difference between the average cosine similarity of predicted positive pairs and predicted negative pairs. "w/o FR" refers to the absence of the face restoration process. Atmospheric turbulence parameters range from 20,000 to 40,000.

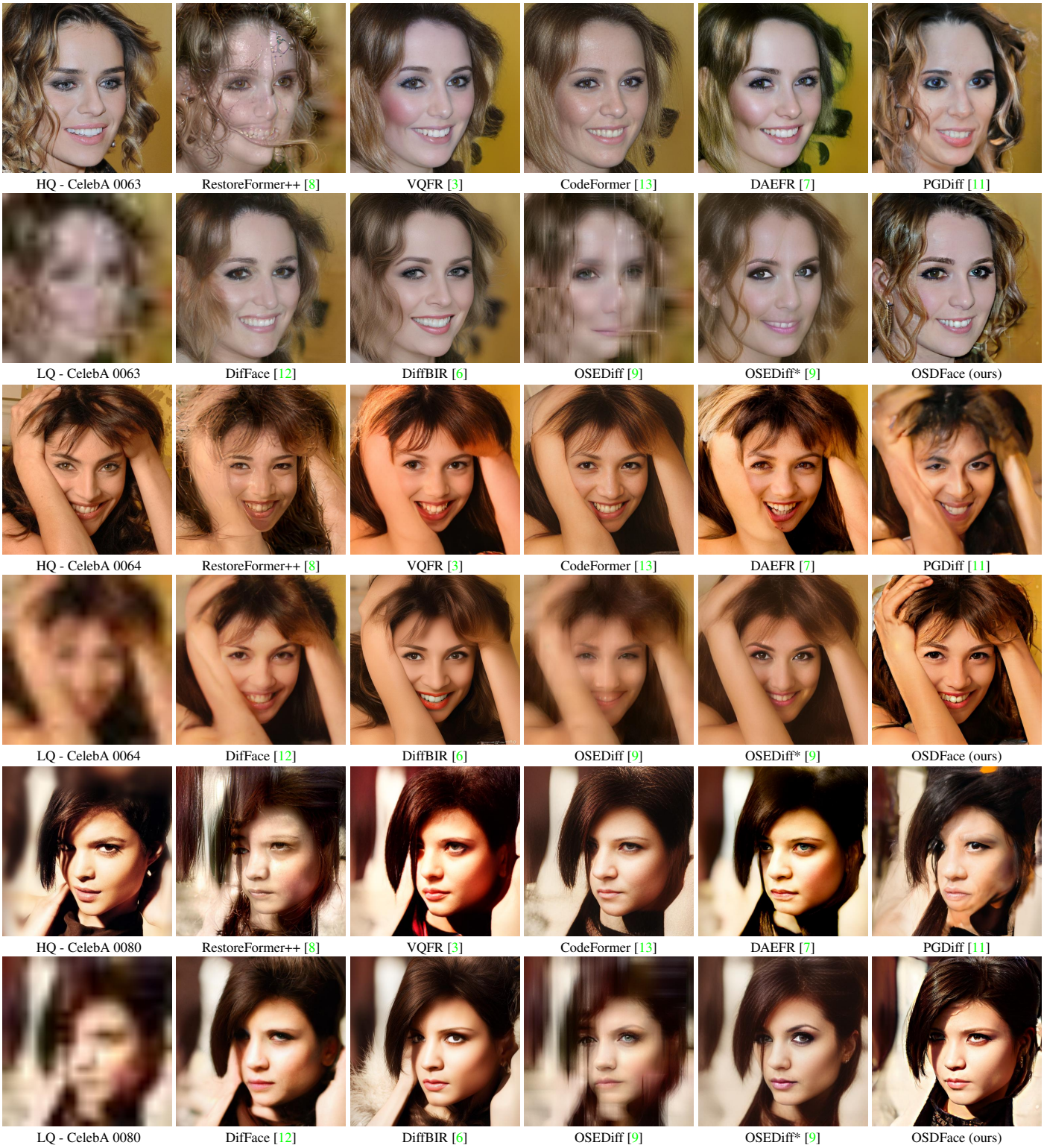

Figure 4. More visual comparison of the synthetic CelebA-Test dataset in challenging cases. Please zoom in for a better view.

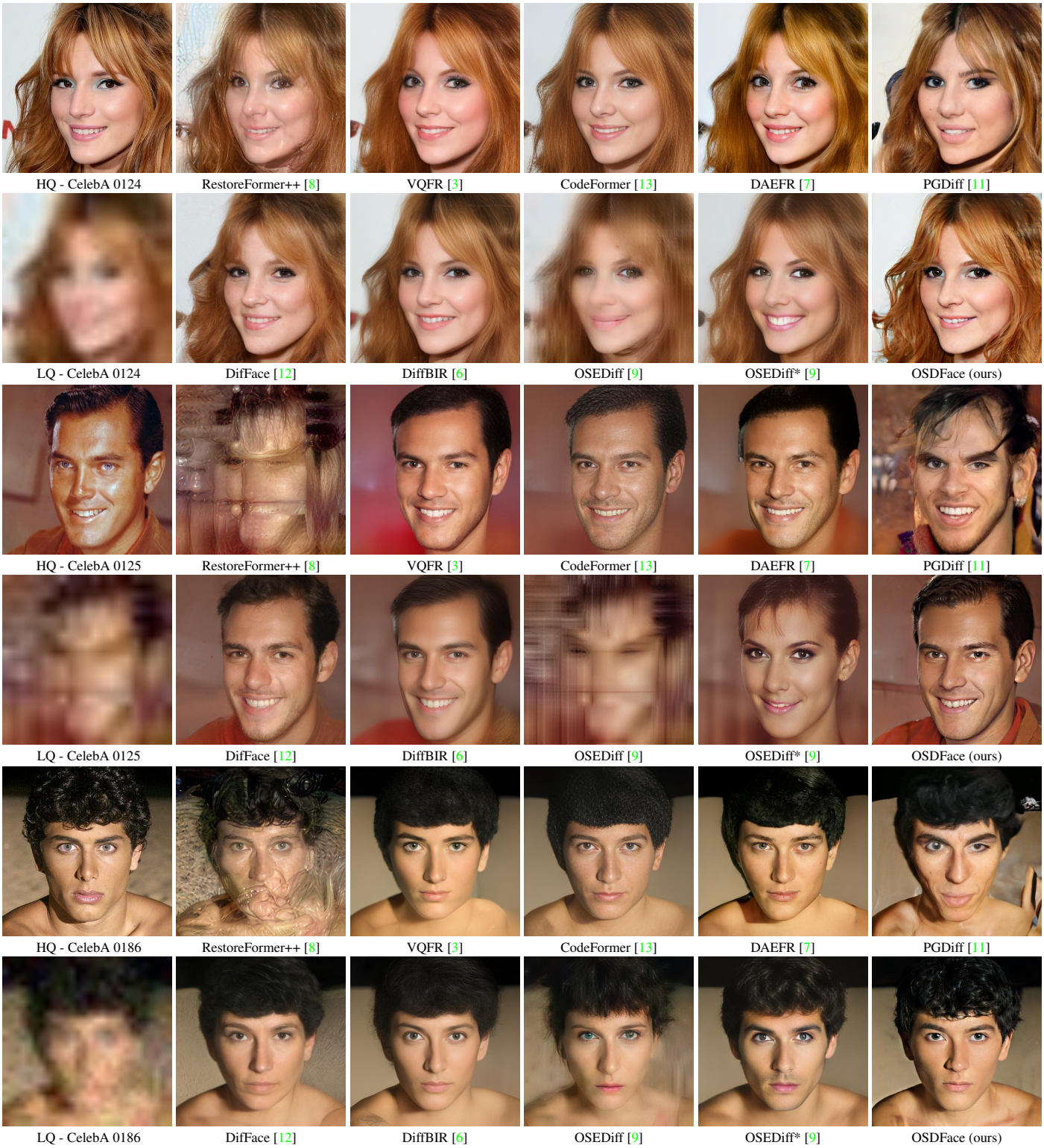

Figure 5. More visual comparison of the synthetic CelebA-Test dataset in challenging cases. Please zoom in for a better view.

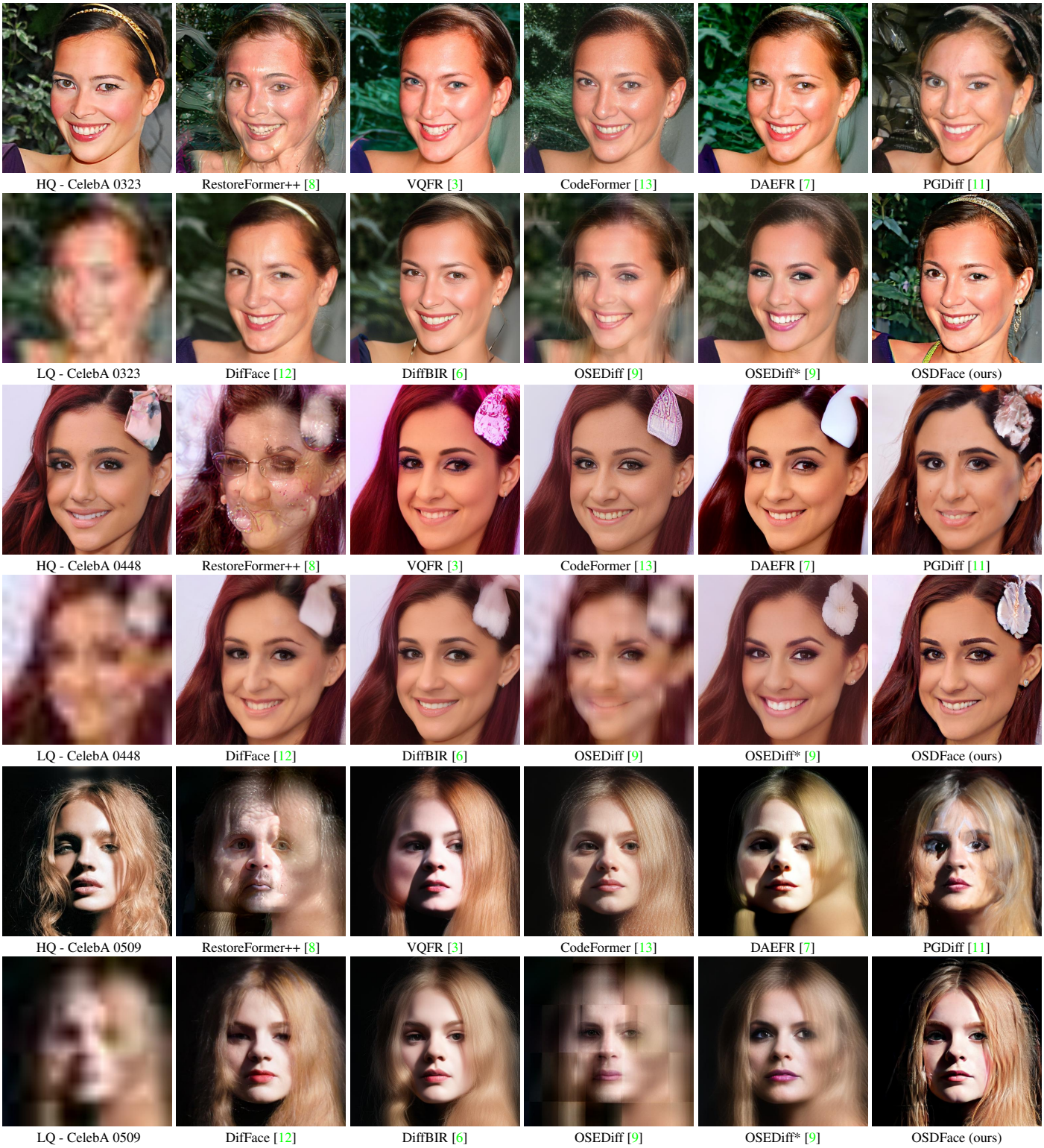

Figure 6. More visual comparison of the synthetic CelebA-Test dataset in challenging cases. Please zoom in for a better view.

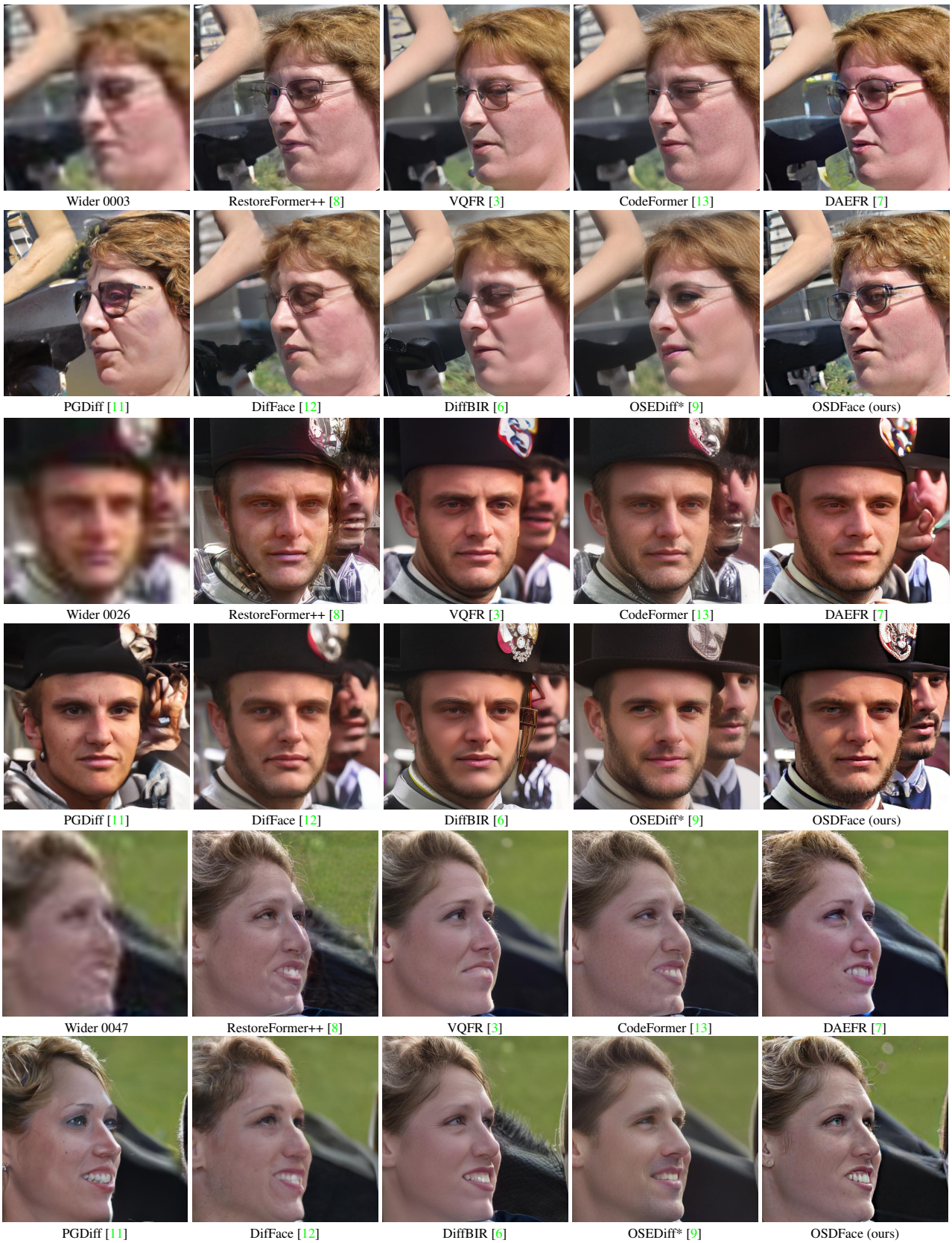

Figure 7. More visual comparison of the real-world Wider-Test dataset in challenging cases. Please zoom in for a better view.

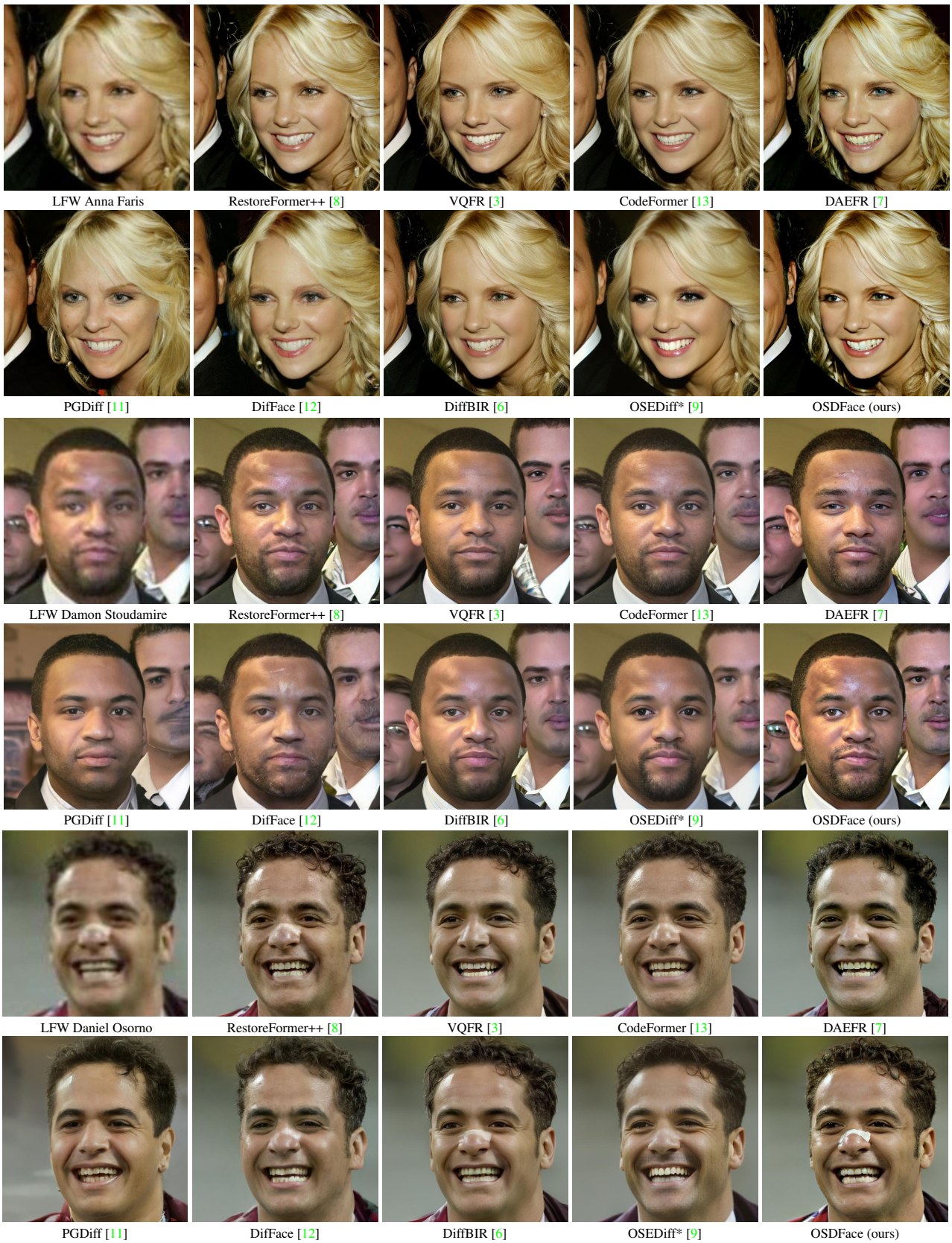

Figure 8. More visual comparison of the real-world LFW-Test dataset in challenging cases. Please zoom in for a better view.

Figure 9. More visual comparison of the real-world WebPhoto-Test dataset in challenging cases. Please zoom in for a better view.

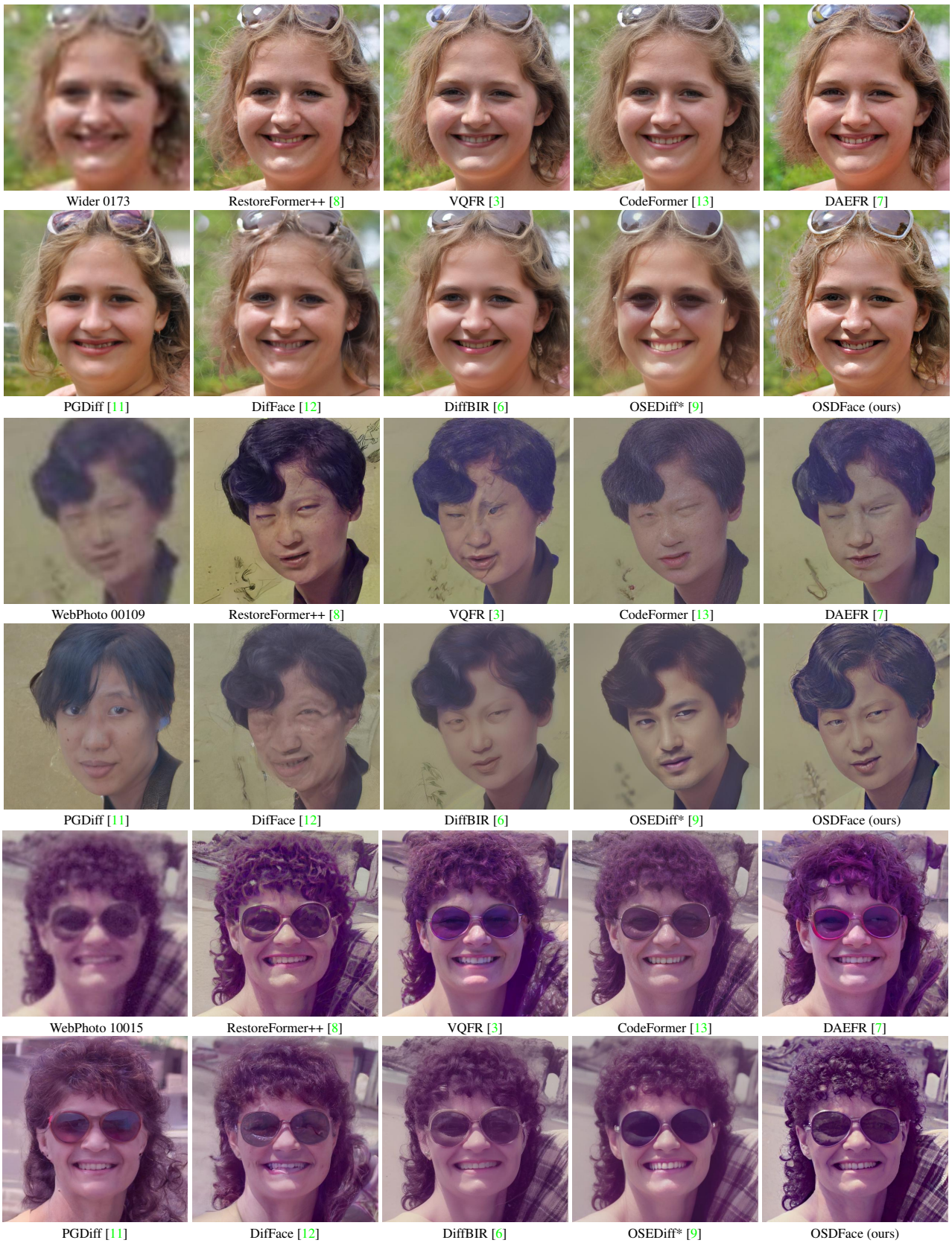

Figure 10. More visual comparison of the real-world datasets in challenging cases. Please zoom in for a better view.