RestoreFormer: High-Quality Blind Face Restoration from Undegraded Key-Value Pairs
Supplementary Materials

Zhouxia Wang[1] Jiawei Zhang[2] Runjian Chen[1] Wenping Wang[1] Ping Luo[1]
[1] The University of Hong Kong, [2] SenseTime Research

The supplementary material includes:

1. Detailed structures of encoder $\mathbf{E}_d$ and decoder $\mathbf{D}_d$ in RestoreFormer (Sec. 1).
2. Detailed structures of the networks in ablation study. (Sec. 2)
3. More experiment results and analysis. (Sec. 3)
4. Details of user study. (Sec. 4)
5. Limitations, broader Impact, and ethic statement. (Sec. 5)

1. Structures of Encoder and Decoder

The detailed structures of encoder $\mathbf{E}_d$ and decoder $\mathbf{D}_d$ are shown in Table 1 and Table 2, respectively. Noted that $\mathbf{E}_h$ and $\mathbf{D}_h$ for attaining HQ Dictionary have the same structures of $\mathbf{E}_d$ and $\mathbf{D}_d$.

2. Frameworks in Ablation Study

Exp 1: Degraded + None. As shown in Figure 1 (d), similar to the proposed RestoreFormer shown in Figure 1 (a), the degraded representation $\boldsymbol{Z}_d$ is extracted from a corrupted image $\boldsymbol{I}_d$ with an encoder $\mathbf{E}_d$. However, since there is no priors, $\boldsymbol{Z}_d$ is sent into the decoder $\mathbf{D}_d$ directly for the reconstruction of a high-quality face $\hat{\boldsymbol{I}}_d$
Exp 2: Degraded + MHSA. As shown in Figure 1 (e), compared to Figure 1 (d), we add two MHSAs described in Figure 1 (f) between the encoder $\mathbf{E}_d$ and decoder $\mathbf{D}_d$ for further attaining contextual information in $\boldsymbol{Z}_d$.
Exp 3: Prior + MHSA. As shown in Figure 1 (b), in this setting, two MHCA blocks in Restorformer (Figure 1 (a)) are replaced by two MHSA blocks. In this experiment, the degraded information $\boldsymbol{Z}_d$ will not involve in the following restoring process.
Exp 4: SFT. As shown in Figure 1 (c), in this setting, the MHCAs in the proposed RestoreFormer are replaced with a Spatial Feature Transform (SFT) [9] layer while fusing the degraded information $\boldsymbol{Z}_d$ and the high-quality facial priors $\boldsymbol{Z}_p$. Specifically, a pair of affine transformation parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) is generated from $\boldsymbol{Z}_d$ by two convolutional layers. After that, the fusing result of $\boldsymbol{Z}_d$ and $\boldsymbol{Z}_p$ is attained by scaling and shifting $\boldsymbol{Z}_p$, formulated by:

$$\begin{aligned} \boldsymbol{\alpha}, \boldsymbol{\beta} &= Conv(\boldsymbol{Z}_d), \\ \boldsymbol{Z}'_f &= \alpha \odot \boldsymbol{Z}_p + \beta \end{aligned} \tag{1}$$

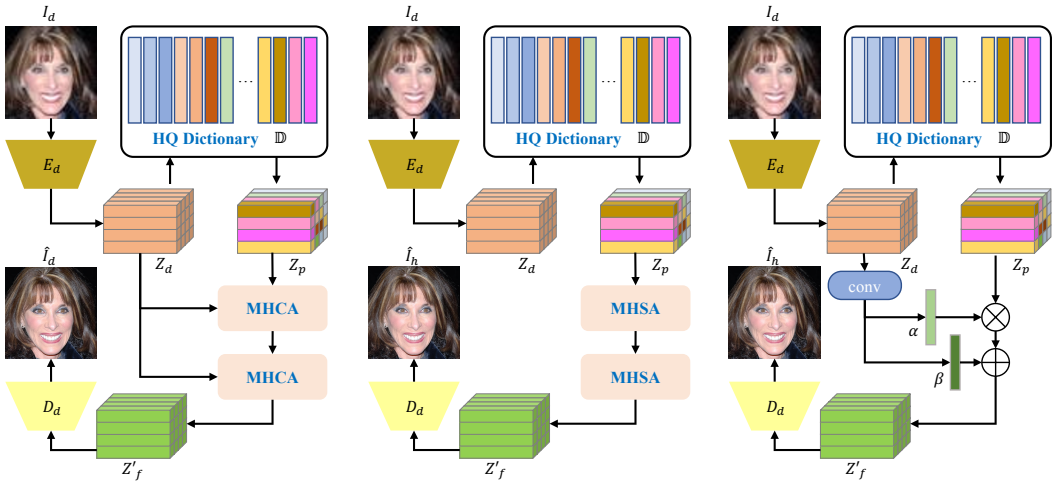
where $\odot$ is an element-wise multiply operation.
Exp 5: MHCA-D. Different from MHCA-P in Figure 1 (g)
More results of these experiments are shown in Sec 3.

layer_name	out_size	Blocks	layer_name	out_size	Blocks
conv_in	512×512	$3 \times 3, 64, stride1$			
block_0	512×512	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 64, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 64, stride1 \end{bmatrix} \times 2$	block_3	64×64	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \end{bmatrix} \times 2$
down_0	256×256	$Conv: 3 \times 3, 64, stride2$	down_3	32×32	$Conv: 3 \times 3, 256, stride2$
block_1	256×256	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \end{bmatrix} \times 2$	block_4	32×32	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \end{bmatrix} \times 2$
down_1	128×128	$Conv: 3 \times 3, 128, stride2$	down_4	16×16	$Conv: 3 \times 3, 256, stride2$
block_2	128×128	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \end{bmatrix} \times 2$	block_5	16×16	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 512, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 512, stride1 \end{bmatrix} \times 2$
down_2	64×64	$Conv: 3 \times 3, 128, stride2$	conv_out	16×16	$3 \times 3, C, stride1$

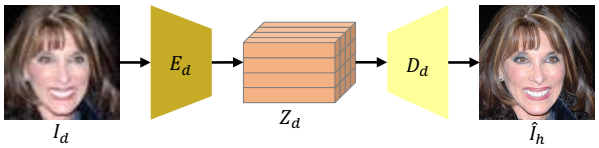
Table 1. Detailed structures of Encoder $\mathbf{E}_d$. In each block_x, there are two blocks that consist of two GroupNorm layers, two nonlinear layers implemented with $in * \mathrm{sigmoid}(in)$, where in is the input, and two convolutional layers (in $Conv: x \times x, y, stride\ z$, x is the size of kernel, y is the size of output, and z is the size of stride). In conv_out, C is the length of the elements in HQ Dictionary.

layer_name	out_size	Blocks	layer_name	out_size	Blocks
conv_in	16×16	$3 \times 3, 512, stride1$			
block_0	16×16	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 512, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 512, stride1 \end{bmatrix} \times 3$	block_3	128×128	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \end{bmatrix} \times 3$
up_0	32×32	$nearest\ upsampling$ $Conv: 3 \times 3, 512, stride1$	up_3	256×256	$nearest\ upsampling$ $Conv: 3 \times 3, 128, stride1$
block_1	32×32	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \end{bmatrix} \times 3$	block_4	256×256	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 128, stride1 \end{bmatrix} \times 3$
up_1	64×64	$nearest\ upsampling$ $Conv: 3 \times 3, 256, stride1$	up_4	512×512	$nearest\ upsampling$ $Conv: 3 \times 3, 128, stride1$
block_2	64×64	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 256, stride1 \end{bmatrix} \times 3$	block_5	512×512	$\begin{bmatrix} GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 64, stride1 \\ GroupNorm: 32 \\ Nonlinear \\ Conv: 3 \times 3, 64, stride1 \end{bmatrix} \times 3$
up_2	128×128	$nearest\ upsampling$ $Conv: 3 \times 3, 256, stride1$	conv_out	512×512	$3 \times 3, 3, stride1$

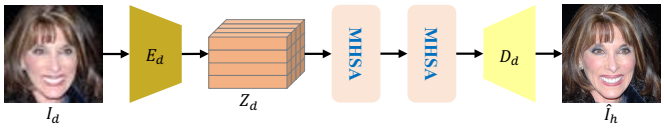
Table 2. Detailed structures of Decoder $\mathbf{D}_d$. In each block_x, there are three blocks that consist of two GroupNorm layers, two nonlinear layers implemented with $in * \mathrm{sigmoid}(in)$, where in is the input, and two convolutional layers (in $Conv: x \times x, y, stride\ z$, x is the size of kernel, y is the size of output, and z is the size of stride).



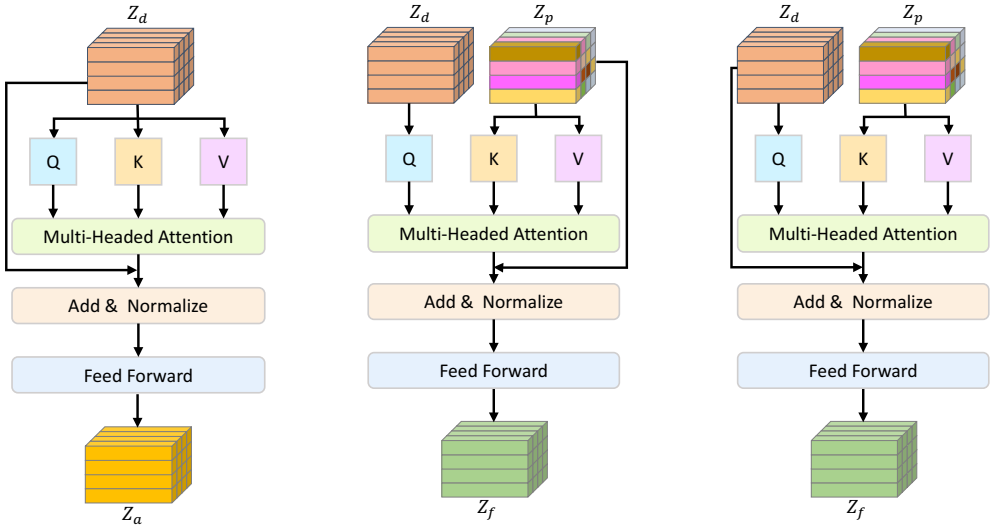
(a) **RestoreFormer** (b) **Exp 3: Prior + MHSA** (c) **Exp 4: SFT**



(d) **Exp 1: Degraded + None**



(e) **Exp 2: Degraded + MHSA**



(f) **MHSA** (g) **MHCA-P (MHCA)** (h) **Exp 5: MHCA-D**

Figure 1. **Frameworks.** (a) is the frameworks of the proposed RestoreFormer. (b) is the framework of exp 3 in the ablation study. It replaces the MHCA (Transformer with Multi-Head Cross-Attention) in (a) with MHSA (Transformer with Multi-Head Self-Attention) and the degraded information $\boldsymbol{Z}_d$ will not involve in MHSA and the following restoring process. (c) is the framework of exp 4 in the ablation study. It replaces the MHCA in (a) with SFT (Spatial Feature Transform Layer) [9]. (d) is the framework of exp 1 in the ablation study. There is no facial prior and transformer. (e) is the framework of exp 2 in the ablation study. There is no facial prior but it will model the contextual information of face with MHSA. (f) and (g) are the structures of MHSA and MHCA (named it as MHCA-P since its shortcut connects $\boldsymbol{Z}_p$ and the attended output). MHCA-P is adopted in Restoreformer. (h) is the framework of exp 5 in the ablation study. Different from (g), its shortcut connects $\boldsymbol{Z}_d$ and the attended output.

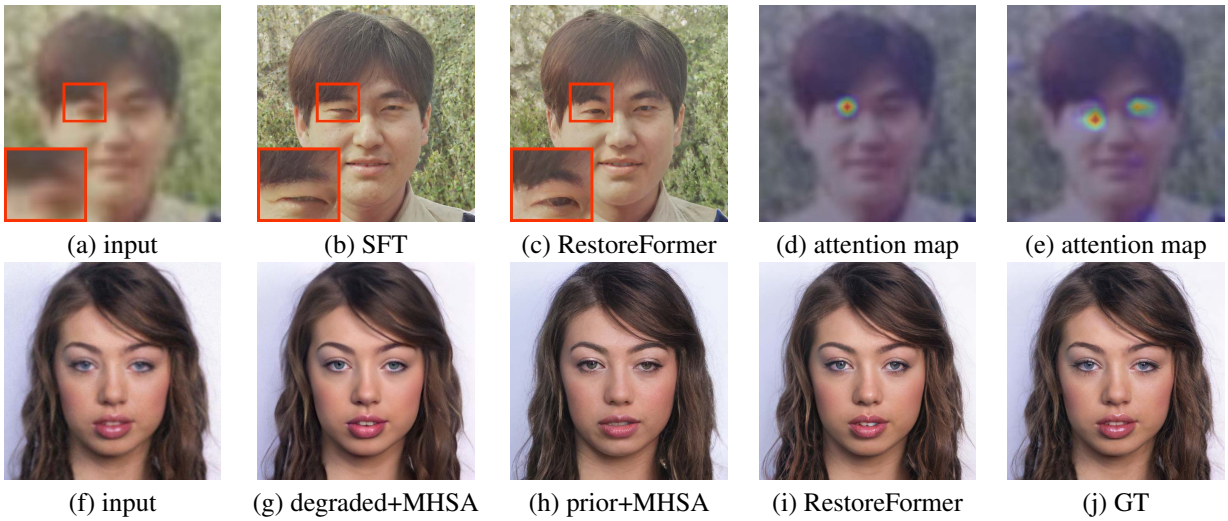
3. More Experiments Analysis

3.1. Importance of MHCA

The ablation study in the script has discussed the importance of MHCA in RestoreFormer. It can not only capture the abundant contextual information in the face, but also model the interaction between the degraded information and high-quality facial priors and finally attain a high-quality face with realness and fidelity. Due to the limited length of the script, our results shown in Figure 6 in the script are too small. Therefore, we show it more clearly in Figure 2.

Since SFT [9] fuses the degraded information and priors locally, the left eye in Figure 2 (b) is weird. In contrast, MHCA in RestoreFormer can globally capture the contextual information in the face. As shown in Figure 2, (d) and (e) are the attention map of left eye. While restoring, it not only focuses on itself but the information in the right eye. Therefore, its result (Figure 2 (c)) is relatively more natural.

Besides, RestoreFormer also tend to model the interaction between the degraded information and high-quality facial priors with MHCA for restoring a high-quality face with realness and fidelity. As Figure 2 shown, although the degradation of the input in (f) is relatively slight, the result attained from degraded information (shown in (g)) is lack of details, especially the hair area. And the result attained from high-quality priors (shown in (h)) looks real but less similar to the original person. On the contrary, the results of RestoreFormer in (i) owns more detaisl than (g) and more similar to the original person than (h).



(a) input (b) SFT (c) RestoreFormer (d) attention map (e) attention map

(f) input (g) degraded+MHSA (h) prior+MHSA (i) RestoreFormer (j) GT

Figure 2. (c) is the restored face of (a) by RestoreFormer. (b) replaces the MHCA with SFT to validate the effectiveness of the facial contextual information. (d) and (e) are two attention maps for the left eye in RestoreFormer. (f) to (j) are to validate the effectiveness of fusing the information from degraded image and prior. (g) and (h) use self-attention, *i.e.* MHSA, to process either degraded information from the input or prior information from HQ Dictionary. While our RestoreFormer can utilize these two sources of information to restore a face (i) that looks more visually pleasant than (g) and more similar to the ground truth (j) than (h). **Zoom in for better view.**

3.2. Advantages of HQ Dictionary

Compared to the component dictionaries proposed in DFDNet [3], which are extracted by an off-line VGG [6] model and mainly focus on eyes, nose, and mouth, the HQ Dictionary adopted in RestoreFormer is reconstruction-oriented and covers all the areas of the face. Therefore, the HQ Dictionary can provide more facial details for the restoration of degraded faces. The results shown in Figure 3 can demonstrate the advantages of the proposed HQ Dictionary. DFDNet [3] works well in eyes, nose, and mouth when the degradation of the input face is relatively small (Figure 3 (a) (b) (c)). However, the **hair** area of DFDNet [3] contains less details. Also, DFDNet [3] even cannot attain high-quality eyes, nose, and mouth when it meets a face with severe degradations (Figure 3 (d)). On the contrary, thanks to the reconstruction-oriented and comprehensive HQ Dictionary, the results of Restoreformer are more visually pleasant. They own more facial details not only in the eyes, nose, and mouth but also in the hair area.

Figure 3. Results of DFDNet [3] and RestoreFormer on real-world data. The proposed RestoreFormer can restore more high-quality details in both important facial components (such as eyes, nose, and mouth) and other areas (such as hair), while DFDNet [3] mainly focuses on eyes, nose, and mouth. **Zoom in for a better view and focus on the hair area.**

3.3. MHCA-P v.s MHCA-D

As shown in Figure 4, the results of MHCA-D, whose shortcut is a skip connection between attended feature $\boldsymbol{Z}_{mh}$ and degraded feature $\boldsymbol{Z}_d$, contain artifacts and have an unnatural look. On the contrary, MHCA-P, whose shortcut is a skip connection between attended feature $\boldsymbol{Z}_{mh}$ and prior $\boldsymbol{Z}_p$, attains more facial details and has a more visual pleasant look in its results. It seems that the high-quality facial priors play a more important role in the final restored results. Therefore, MHCA-P is adopted in RestoreFormer.

Figure 4. Comparison between MHCA-D and MHCA-P (adopted in RestoreFormer). MHCA-P, whose shortcut is a skip connection between attended feature $\boldsymbol{Z}_{mh}$ and prior $\boldsymbol{Z}_p$, attains more facial details and has a more visual pleasant look in its results. On the contrary, the results of MHCA-D, whose shortcut is a skip connection between attended feature $\boldsymbol{Z}_{mh}$ and degraded feature $\boldsymbol{Z}_d$, contain artifacts and have an unnatural look.

3.4. Running Time and Model Size

We also have a comparison with the recently state-of-the-art methods on the running time and model size. The running time is tested on GeForce GTX 1060 and the evaluating results are shown in Table 3. It shows that the running time and model size of RestoreFormer are comparable to the other methods

Methods	DFDNet [3]	PSFRGAN [1]	GFP-GAN [8]	**RestoreFormer**
time/s	0.6429	0.8389	0.07538	0.2288
Memory/M	918	258	587	455

Table 3. **Running time and model size**. The running time and model size of RestoreFormer are comparable to the recently state-of-the-art methods.

3.5. More results

Figure 5, Figure 6, Figure 7, and Figure 8 are more restored results of several state-of-the-art methods and the proposed Restoreformer on four datasets: LFW-Test [2], CelebChild-Test [8], WebPhoto [8], and CelebA-Test [4].

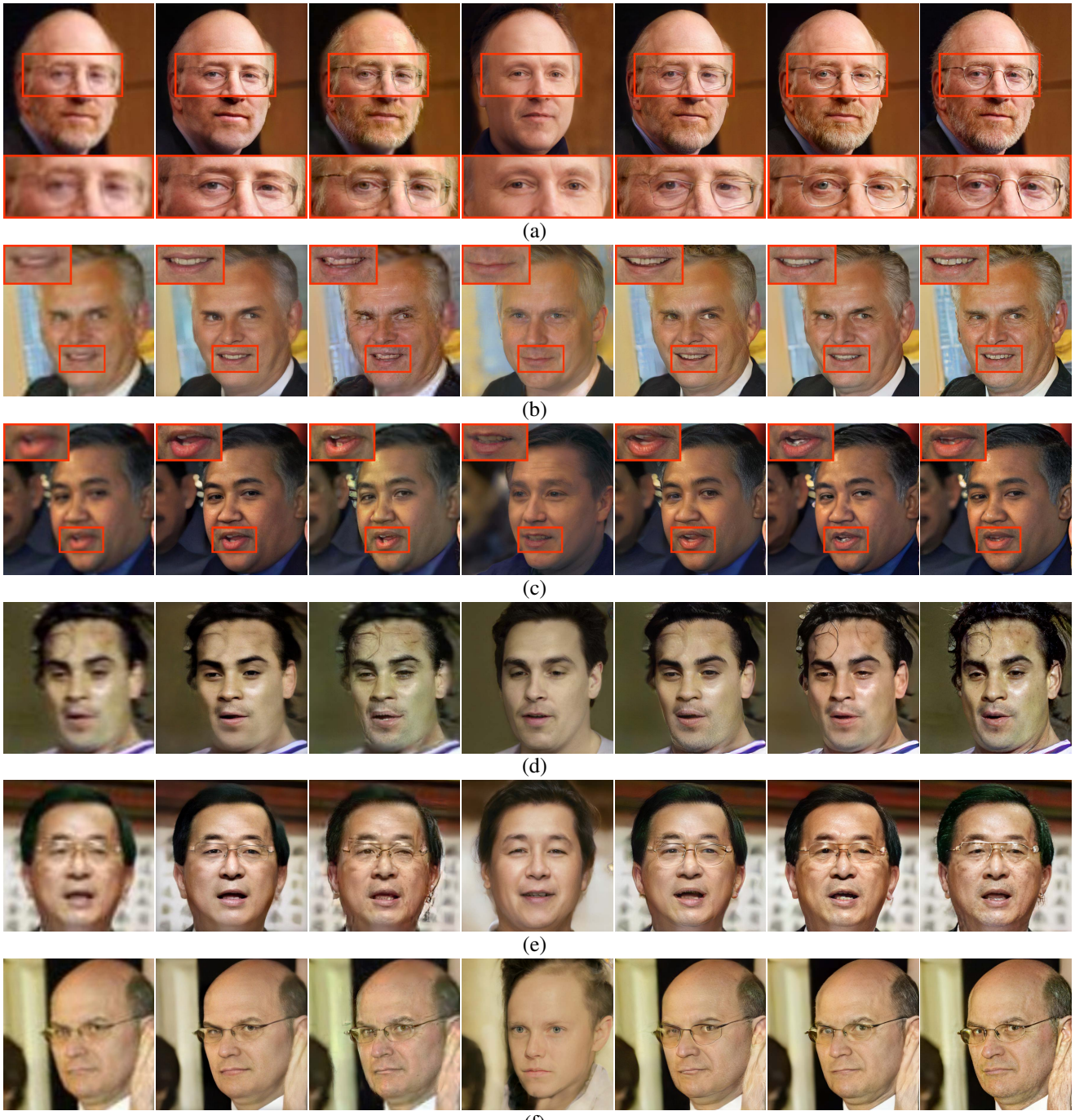


Figure 5. Qualitative comparison on **LFW-Test** [2]. The results of our RestoreFormer have a more realistic overview and contain more details in eyes, mouth, hair, and even the glasses. Like the results in (a), the glasses restored by RetoreFormer are more complete compared to other methods. **Zoom in for a better view**.

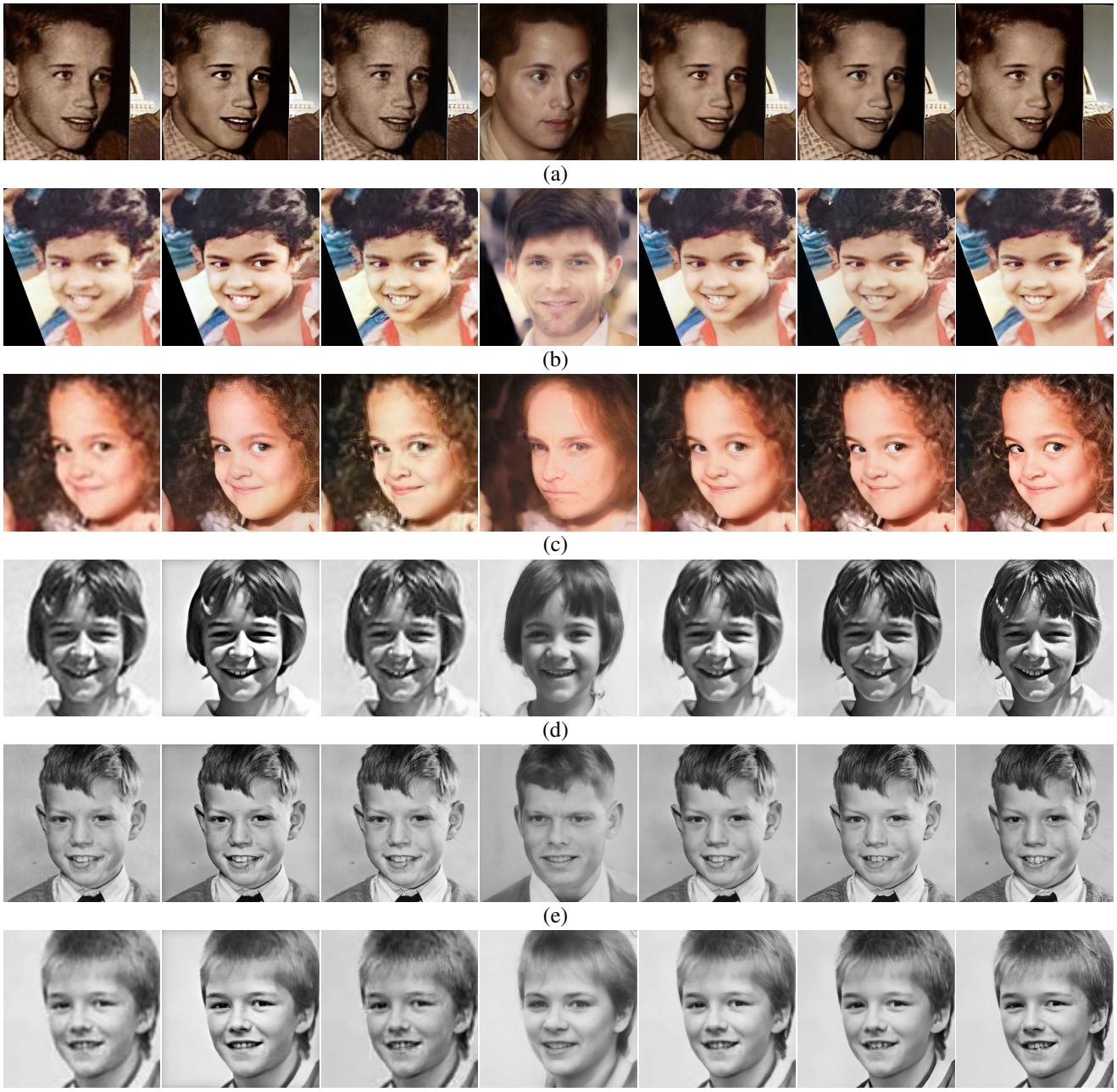


Figure 6. Qualitative comparison on **CelebChild-Test** [8]. The results of our RestoreFormer have a more realistic overview and contain more details in eyes, mouth, and hair. Noted that since there are many black-and-white old photos in this dataset and the methods based on generative priors will lead to colorization. For a more pleasant and consistent visualization, we turn all the results into a gray format for visualization. **Zoom in for a better view**.

Figure 7. Qualitative comparison on **WebPhoto-Test** [8]. The results of our RestoreFormer have a more realistic overview and contain more details in eyes, mouth, and hair. Besides, the results of RestoreFormer tend to be more consistent in some symmetric areas, such as eyebrows and eyes. Like the eyebrows in (a) and the eyes (b) and (c), they have different sizes in the results of Wan *et al*. [7], PSFRGAN [1], and GFP-GAN [8]. On the contrary, they look well in the results of RestoreFormer. **Zoom in for a better view**.

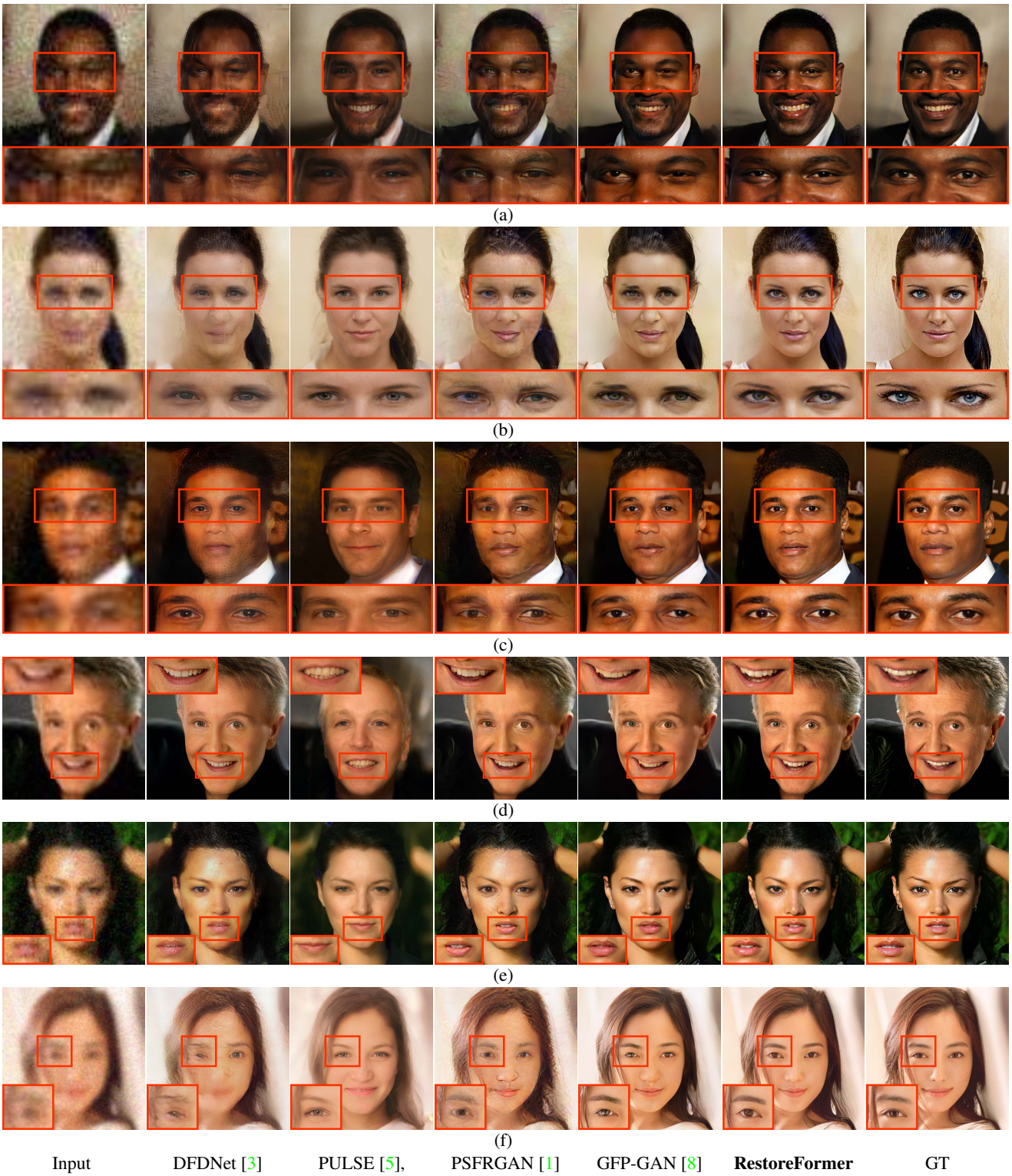


Figure 8. Qualitative comparison on the **CelebA-Test** [4]. Compared to the other methods, the results of our RestoreFormer have a more realistic overview and contain more details. Specifically, the result of RestoreFormer in (a) has eyes nearly in the same size and looks more natural compared to the result of GFP-GAN [8]. The left eyebrow in (b) looks more complete compared to other methods. Furthermore, the left eye restored by RestoreFormer in (f) is more natural while the one generated by other methods exists artifact or blink. Noted that although the results of PULSE [5] are natural, their identities are far away from their GT. **Zoom in for a better view**.

4. User Study

We will show more details of the user study in the script. First, we randomly select 200 samples from LFW-Test [2] and WebPhoto-Test [8] (each dataset provides 100 samples). Then, we process these 200 samples with the proposed RestoreFormer and three recently state-of-the-art methods: DFDNet [3], PSFRGAN [1], and GFP-GAN [8], and attain the restored results. The results of RestoreFormer will pair-wisely compare with the results of the three methods, respectively. Therefore, there are 600 pairs for the user study. We recruit 100 volunteers and each of them will give a one from two selection for 50 pairs randomly selected from the 600 pairs. So, each pair will be compared by about 8.33 times.

5. Limitations, Broader Impact, and Ethic Statement

Limitations. As shown in Figure 9, similar to the existing methods, when meeting the samples with obstacles and non-frontal poses, the restored results of RestoreFormer also exist twists and artifacts. This is because the training data of these methods are mainly front face without obstacles. One possible way for the future to alleviate this problem is to augment the training data with more hard samples and finally build a more robust model for real-world blind face restoration.

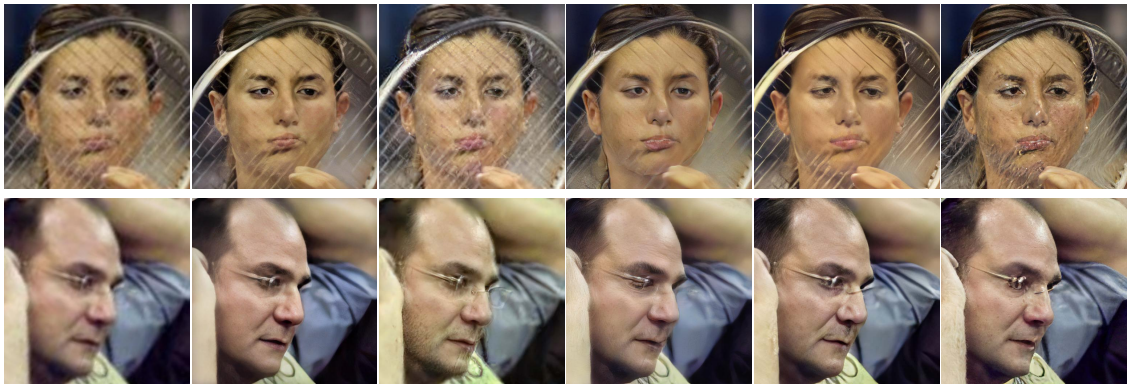


Figure 9. **Limitations.** Similar to the existing methods, RestoreFormer also cannot handle the samples with obstacles and side face well.

Broader Impact. Since the high-quality facial priors are statistics of training data, when applying them to the face restoration, the final results may contain inexistent content that leads to negative social impacts. This issue can be a future study for discussing how to control these generating content for minimizing its potential negative social impacts and maximize its advantages about promoting the development of society.

Ethic Statement. The face images adopted in the work all are collected by the existing works and we cite them well.

References

[1] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 6, 7, 8, 9, 10, 11

[2] 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. 7, 11

[3] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 5, 6, 7, 8, 9, 10, 11

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7, 10

[5] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 7, 8, 9, 10

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[7] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, 2020. 7, 8, 9, 11

[8] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 6, 7, 8, 9, 10, 11

[9] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 1, 4, 5