Restoration by Generation with Constrained Priors

Supplementary Material

A. Additional Results on Blind Face Restoration

A.1. Standard Benchmark

We provide additional qualitative comparisons on Wider-Test dataset in Figure 3 and the Deblur-Test dataset in Figure 4. Wider-Test contains 970 images selected from the Wider-Face[9] dataset which is initially collected for face detection that contains many real-world low-quality face images. The Deblur-Test dataset contains 67 real-world motion-blur images from [4]. Both of the datasets are aligned using the same way in FFHQ[3]. All the previous methods we compare are synthetic-data-based methods. As our method does not utilize synthetic data which previous methods rely on, our method shows good generalizability in handling different kinds of real-world degraded images.

A.2. More Distortion Types

We further provide results on more distortion types e.g., JPEG compression and scratches in Figure 1. This further demonstrates that our method is able to perform better on out-of-distribution input low-quality images as we don't utilize synthetic data for training.

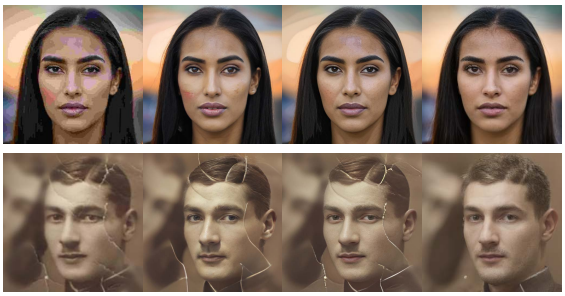


Figure 1. Results on JPEG compression(top) and scratches(bottom).

B. Additional Results on Personalized Blind Face Restoration

In this section, we present more personalized restoration results involving additional subjects. This includes Subject B (a man), as well as public figures such as Biden and Hermione. These results are showcased in Figure 5. We compare our method with previous methods CodeFormer[10] and DR2(+VQFR) [8] which are two single-image-based restoration methods that rely on synthetic data as well as ASFFNet[5] that requires a reference dataset. We use the same personal album in ASFFNet and our method. Our findings demonstrate a superior preservation of identity, along with a high level of quality in the results.

C. Anchor Images and Constrained Generative Space

In this section, we discuss both the generative album and the personalized album utilized by the model to restrict the generative space for restoration. Additionally, we visualize this constrained space through unconditional generation from the fine-tuned model.

For the generative album, we employ the input low-quality image with skip guidance to produce anchor images. These images are then used to fine-tune the model. We display the generated album, along with randomly generated images from the fine-tuned model, in Figure 6. The generated album contains images similar to the input but with enhanced quality, though not as high as those produced by the pre-trained model. This is likely due to the influence of skip guidance. Nevertheless, the model fine-tuned with this album is capable of generating high-quality images that still bear resemblance to the original input.

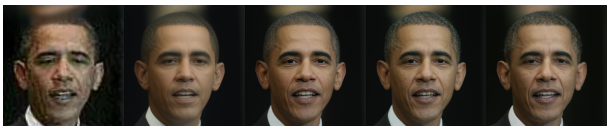
Regarding the personalized album, we collect around 20 real high-quality images to act as anchor images. Examples of these images, along with images randomly generated by the personalized model, are presented in Figure 7. The images produced by the personalized model exhibit both diversity and identity preservation, attributes learned from the personal album.

D. More Ablation Studies

Noise Step K and Constraining Prior with Generative Album. Here we provide supplemental results to Figure 8 from the main paper, which analyzes the effect of noise step K and the effectiveness of using a generative album to constrain the prior. Results are shown in Figure 8. From the results we can see that as K increases both results using either the constrained prior or not would have better quality but would not be less faithful to the input image. However, with constrained prior, we can see that the loss in faithfulness is considerably less than the ones not using the constrained prior.

Skip Guidance for Generative Album. We analyze the effectiveness of our proposed Skip Guidance in generating a generative album from a degraded input image. The album should contain images close to the input yet of high quality, serving as anchor images for the constrained generative space. Figure 9 shows that without guidance (i.e., direct sampling of the album from y_K), we obtain high-quality images that do not closely resemble the input, thus failing to effectively constrain the generative space. Conversely, applying skip guidance too frequently lowers sample quality due to the guidance's approximate nature, potentially leading to a constrained space filled with low-quality images.

Size of Personal Album. In Fig.2, we present results using personal albums of varying sizes. Generally, larger albums enable the model to better preserve identity and details, though the improvements diminish with size increase.



Input Size=1 Size=4 Size=8 Size=16

Figure 2. Ablation on Size of Personal Album.

E. Theoretical Analysis

In this section, we provide the theoretical analysis for the intuition we claimed in the paper.

1. *Adding noise to high-quality and low-quality images can progressively align their distributions, making them more similar over time.*
   For a clean image x_0 and low-quality image y_0, we can analyze the distribution of their noisy versions: $q(x_t|x_0)$ and $q(y_t|y_0)$ which are

$$q(x_t|x_0) \sim N(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (1)$$
$$q'(y_t|y_0) \sim N(\sqrt{\bar{\alpha}_t}y_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (2)$$

   in which $\bar{\alpha}_t$ is a hyper-parameter in the diffusion process and will decrease as t increases. Therefore we can compute the KL divergence between these two distributions:

$$KL(q, q') = \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}(x_0 - y_0)^2. \quad (3)$$

   When more noise is added ($t \uparrow$), $KL(q, q')$ will decrease. Therefore the two distributions get more similar as more noise is added.
2. *The larger t is, the larger the generative space $p(x_0|x_t)$ spans.*
   In this case, we compute the entropy H of $p(x_0|x_t)$ to show how large the generative space spans. Let's consider $q(x_t|x_0)$ first:

$$q(x_t|x_0) \sim N(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (4)$$

   Thus we can compute the entropy:

$$H(q(x_t|x_0)) = \frac{1}{2}\log(2\pi(1-\bar{\alpha}_t)) + \frac{1}{2} \quad (5)$$

   For a deterministic denoising process (such as DDIM), we can have $H(p(x_0|x_t)) = H(q(x_t|x_0))$. Therefore $H(p(x_0|x_t))$ increases as t increases, showing the generative space spans larger.

F. Implementation Details

We provide the model details trained on datasets (256×256 and 512×512) along with the training/inference parameters in Table 1. Due to the lack of 512×512 model trained with diffusion models and its slow speed in both training and inference, we use 256×256 for standard benchmarks while for personalized restoration, we utilize a 512×512 model.

We first train an unconditional generative model using the model architecture based on [1, 6]. After this, we will get a powerful generative prior that can output high-quality images. Then we finetune the model using either the generative album or the personal album. For the generative album, we first generate the images with the skip guidance to ensure that our images follow the input. Then we finetune the model using either the generative album or the personal album. Finally, we restore the images using the constrained prior.

	256×256	512×512
Model Details		
Diffusion Steps	1000	1000
Channels	128	256
Channels Multiple	$1, 1, 2, 2, 4, 4$	$0.5, 1, 1, 2, 2, 4, 4$
Heads Channels	128	64
Attention Resolution	16	$32, 16, 8$
Dropout	0.1	0.1
Training Details		
Batch Size[1]	256	32
Iterations	200k	2320k
Learning Rate	10^{-4}	10^{-4}
Optimizer	Adam	Adam
Weight Decay	0.0	0.0
Generative Album		
Noise Step K	600	-
Skip Guidance	20	-
Finetuning Details		
Batch Size	4	4
Iterations[2]	3000	5000
Learning Rate	10^{-5}	10^{-5}
Inference Details		
Noise Step K	200	300

Table 1. **Implementation details**. [1] for AFHQ-Dog and AFHQ-Cat (256×256), the iterations are 50k and 100k respectively. [2] for personalized finetuning, we use 5000 iterations.

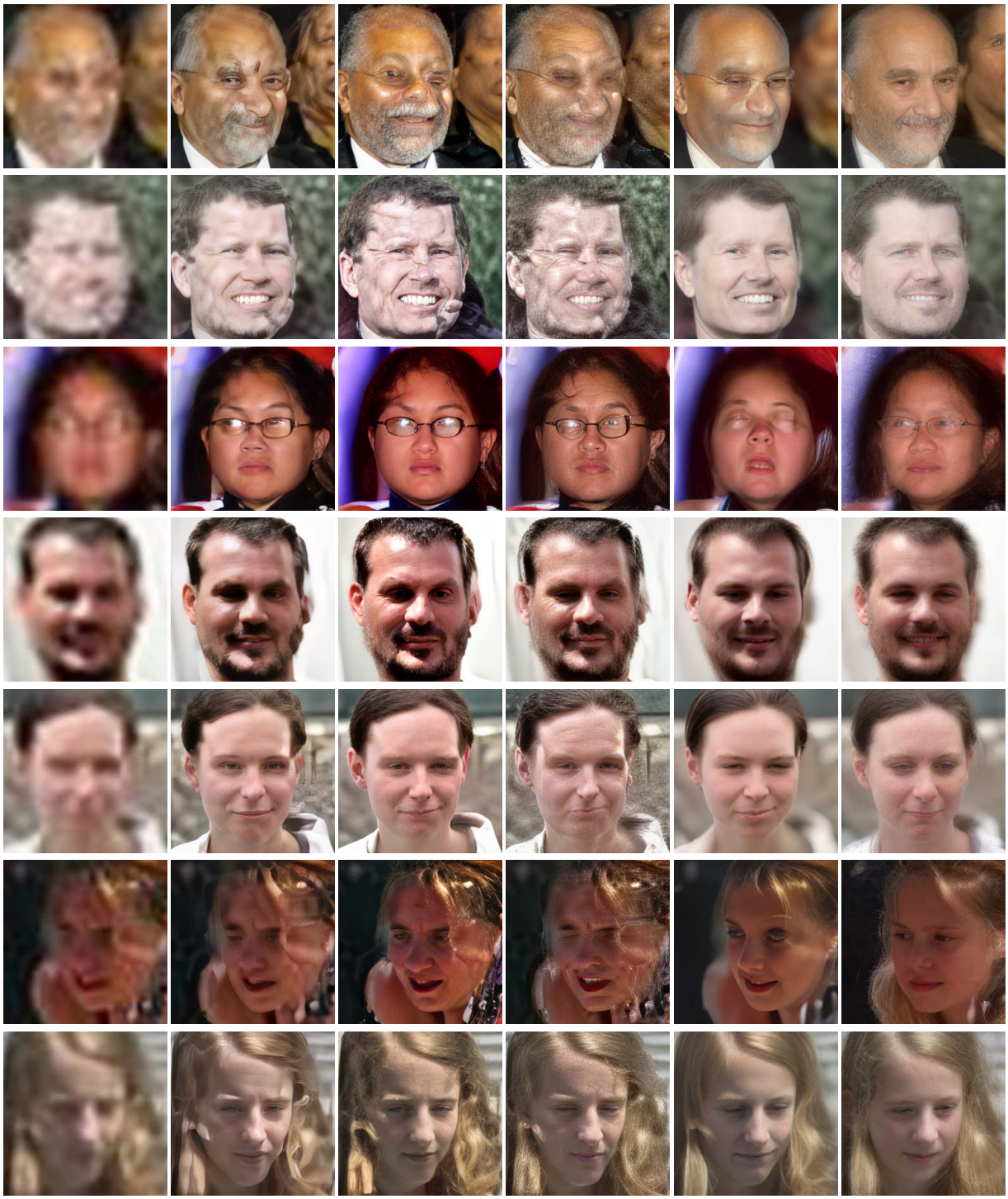


Figure 3. **More qualitative comparison with previous methods on Wider-Test.**

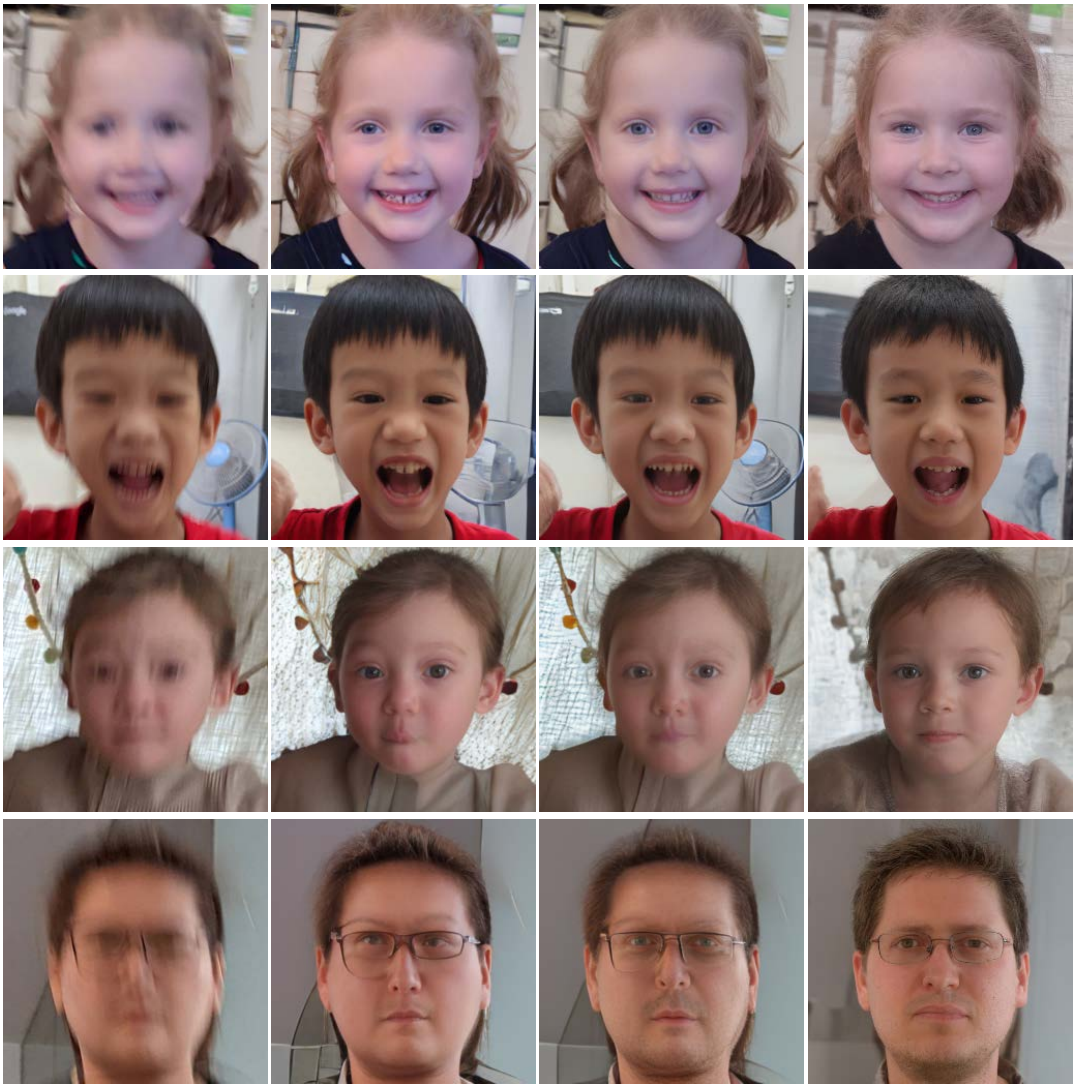


Figure 4. **More qualitative comparison with previous methods on Deblur-Test.**

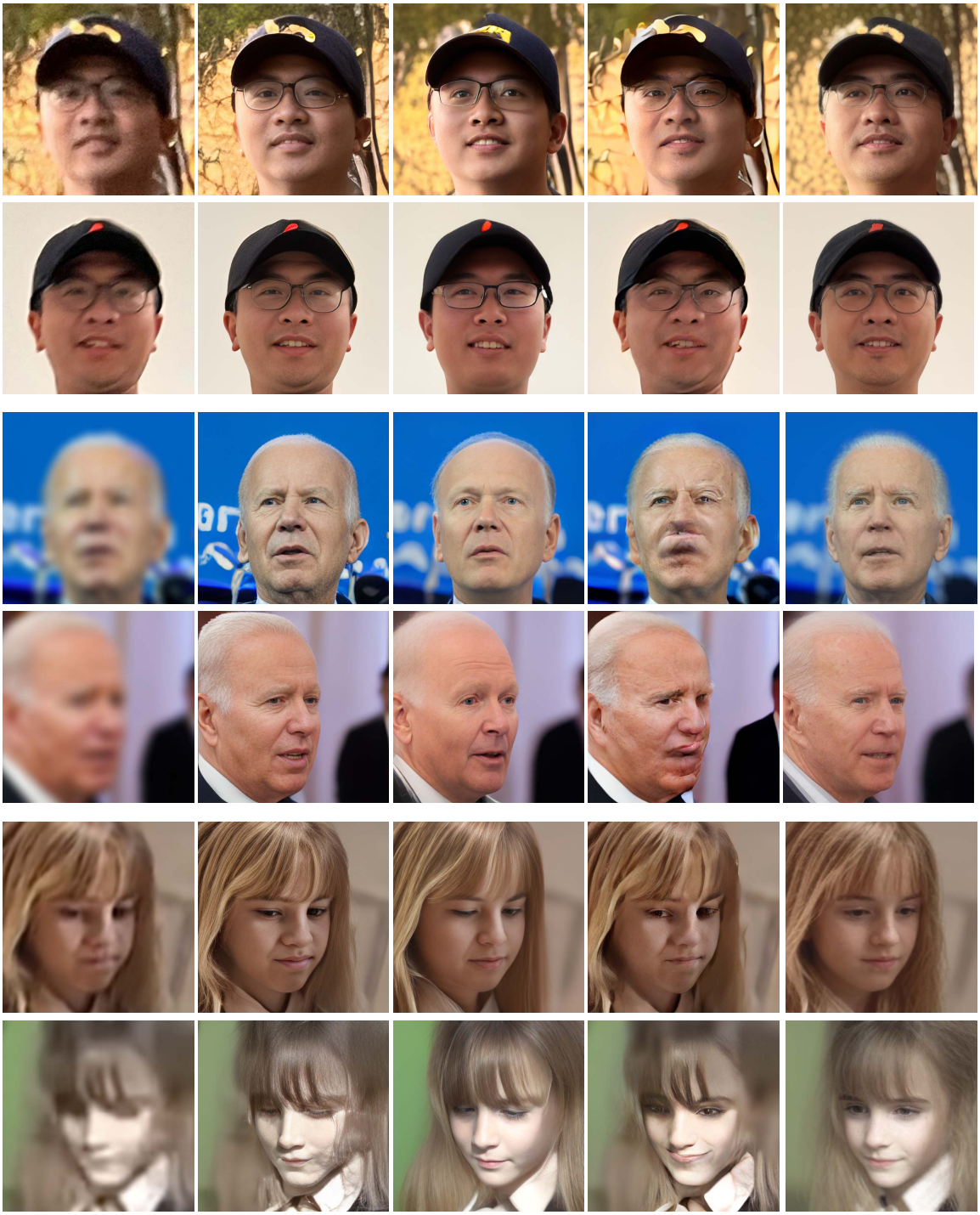


Figure 5. **More Qualitative Comparison on Personalized Face Restoration.** We present three subjects here. For each subjects, we compare two real-world low-quality images with previous methods. The subjects from the top to bottom are: subject B referenced in the main paper, Biden and Hermione.

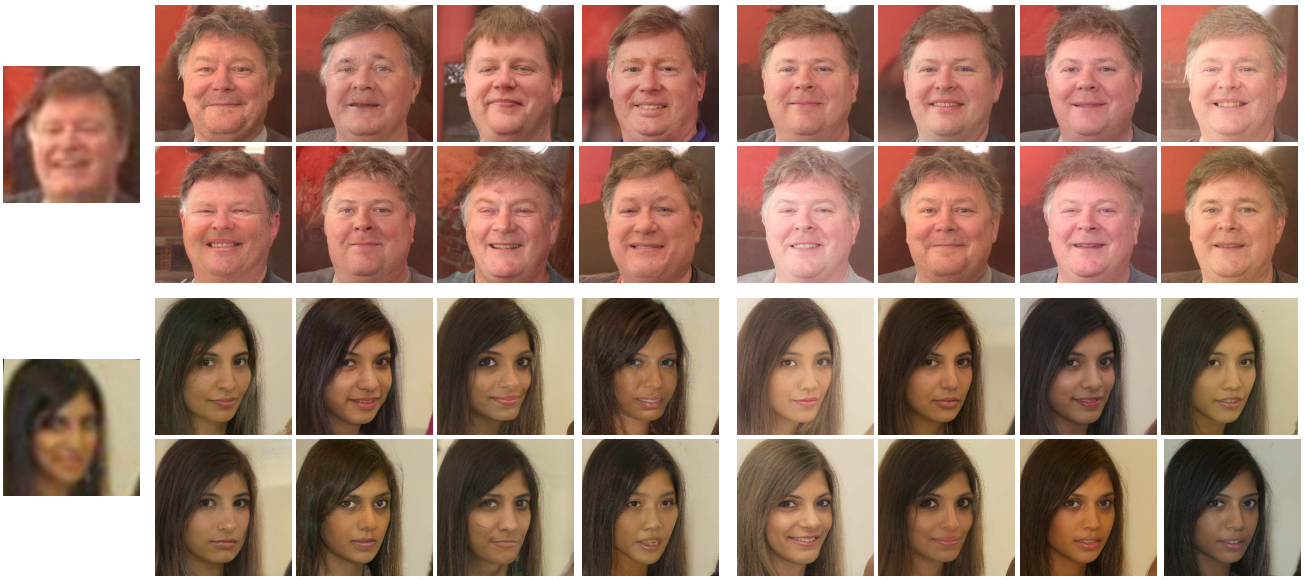


Figure 6. **Generative Album & Unconditional Generation from Fine-tuned Model.** The generative album is generated with the input image as guidance. Model fine-tuned with this album can then generate high-quality images that are close to the original input.

Figure 7. **Personal Album & Unconditional Generation from Personalized Model.** We provide two sets of results.

Figure 8. **More Results of Ablation on Noise Step K and Constraining with Generative Album.**

Figure 9. **The impact of skip guidance frequency on the generative album.** Absence of skip guidance results in divergence from the input, while overly frequent guidance produces low-quality anchor images.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[2] 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 *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 3, 4

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[4] Wei-Sheng Lai, Yichang Shih, Lun-Cheng Chu, Xiaotong Wu, Sung-Fang Tsai, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Face deblurring using dual camera fusion on mobile phones. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 1

[5] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 1, 5

[6] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[7] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[8] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. 1, 3, 5

[9] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 1

[10] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 1, 3, 4, 5