Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance

– Supplementary Material –

Anonymous Author(s)
Affiliation
Address
`email`

In this supplementary material, we provide additional discussions and results. In Sec. A, we present additional implementation details including the inference requirements, choice of hyperparameters involved in the inference process, and discussions on the pre-trained restorer for blind face restoration. In Sec. B, we provide more results on various tasks, *i.e.*, blind face restoration, old photo restoration, reference-based restoration, face colorization and inpainting. Sec. C and Sec. D discuss the limitations and potential negative societal impacts of our work, respectively.

A Implementation Details

A.1 Inference Requirements

The pre-trained diffusion model we employ is a 512×512 denoising network trained on the FFHQ dataset [5] provided by [20]. The inference process is carried out on NVIDIA RTX A5000 GPU.

A.2 Inference Hyperparameters

During the inference process, there involves hyperparameters belonging to three categories. **(1) Sampling Parameters:** The parameters in the sampling process (*e.g.*, gradient scale s). **(2) Partial Guidance Parameters:** Additional parameters introduced by our partial guidance, which are mainly relative weights for properties involved in a certain task (*e.g.*, α that controls the relative importance between the structure and color guidance in face colorization). **(3) Optional Parameters:** Parameters for optional quality enhancement (*e.g.*, the range for multiple gradient steps to take place $[S_{start}, S_{end}]$). While it is principally flexible to tune the hyperparameters case by case, we provide a set of default parameter choices for each homogeneous task in Table 1.

Table 1: Default hyperparameter settings in our experiments.

Task	**Sampling**		**Partial Guidance**					**Optional**			
	s_{norm}	s	Unmasked Region	Lightness	Color Statistics	Smooth Semantics	Identity Reference	$N = 2$	$N = 3$	Perceptual Loss	GAN Loss
Restoration		0.1	-	-	-	$\mathcal{L}_{res}$	-	$T \sim 0.5T$	$T \sim 0.7T$	1e-2	1e-2
Colorization	✓	0.01	-	$\mathcal{L}_l$	$0.01\mathcal{L}_c$	-	-	-	-	-	-
Inpainting	✓	0.1	$\mathcal{L}_{inpaint}$	-	-	-	-	-	-	-	-
Ref-Based Restoration		0.1	-	-	-	$\mathcal{L}_{res}$	$100\texttt{sim}(v_{\hat{x}_0}, v_r)$	$T \sim 0.5T$	$T \sim 0.7T$	1e-2	1e-2

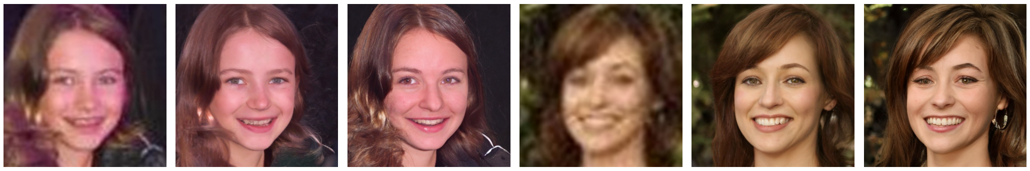
A.3 Restorer Design

Network Structure. In the blind face restoration task, given an input low-quality (LQ) image y_0, we adopt a pre-trained face restoration model f to predict smooth semantics as partial guidance. In this work, we employ the $\times 1$ generator of Real-ESRGAN [17] as our restoration backbone. The network follows the basic structure of SRResNet [6], with RRDB being its basic blocks. In a $\times 1$

generator, the input image is first downsampled 4 times by a pixel unshuffling [14] layer before any convolution operations. In our work, we deal with 512×512 input/output pairs, which means that most computation is done only in a 128×128 resolution scale. To employ it as the restorer f, we modify some of its settings. Empirically we find that adding x_t and t as the input alongside y_0 can enhance the sample quality in terms of sharpness. Consequently, the input to f is a concatenation of y_0, x_t, and t, with t embedded with the sinusoidal timestep embeddings [15].

Training Details. f is implemented with the PyTorch framework and trained using four NVIDIA Tesla V100 GPUs at 200K iterations. We train f with the FFHQ [5] and CelebA-HQ [4] datasets and form training pairs by synthesizing LQ images I_l from their HQ counterparts I_h, following a common pipeline with a second-order degradation model [17, 7, 16, 19]. Since our goal is to obtain *smooth semantics* without hallucinating unnecessary high-frequency details, **it is sufficient to optimize the model f solely with the MSE loss**.

Model Analysis. To investigate the most effective restorer for blind face restoration, we compare the sample quality with restorer being $f(y_0)$ and $f(y_0, x_t, t)$, respectively. Here, $f(y_0, x_t, t)$ is the one trained by ourselves as discussed above, and $f(y_0)$ is SwinIR [9] from DifFace [20], which is also trained with MSE loss only. As shown in Fig.1, when all the other inference settings are the same, we find that the sample quality with restorer $f(y_0, x_t, t)$ is higher in terms of sharpness compared with that of $f(y_0)$. One may choose to sacrifice a certain degree of sharpness to achieve higher inference speed by substituting the restorer with $f(y_0)$, whose output is constant throughout T timesteps.



Input | Output with Restorer $f(y_0)$ | Output with Restorer $f(y_0, x_t, t)$ | Input | Output with Restorer $f(y_0)$ | Output with Restorer $f(y_0, x_t, t)$

Figure 1: Visual comparison of the restoration outputs with different restorers f in blind restoration. We observe that including x_t and t as the input to f enhances the sharpness of the restored images.

B More Results

B.1 More Results on Blind Face Restoration

In this section, we provide quantitative and more qualitative comparisons with state-of-the-art methods, including **(1)** task-specific CNN/Transformer-based restoration methods: PULSE [11], GFP-GAN [16], and CodeFormer [21] and **(2)** diffusion-prior-based methods: GDP [2], DDNM [18] and DifFace [20].

To compare our performance with other methods quantitatively, we adopt FID [3] and NIQE [12] as the evaluation metrics and test on three real-world datasets: LFW-Test [16], WebPhoto-Test [16], and WIDER-Test [21]. LFW-Test consists of the first image from each person whose name starting with `A` in the LFW dataset [16], which are 431 images in total. WebPhoto-Test is a dataset comprising 407 images with medium degradations collected from the Internet. WIDER-Test contains 970 severely degraded images from the WIDER Face dataset [16]. As shown in Table 2, our method achieves best or second-best scores across all three datasets for both metrics. Although GFP-GAN achieves the best NIQE scores across datasets, notable artifacts can be observed as shown in Fig. 2. Meanwhile, our method shows exceptional robustness and produces visually pleasing outputs without artifacts.

Table 2: Quantitative comparison on the *real-world* **LFW-Test**, **WebPhoto-Test**, and **WIDER-Test**. Red and blue indicate the best and the second best performance, respectively.

Dataset	Metric	CNN/Transformer-based Methods			Diffusion-prior-based Methods			
		PULSE [11]	GFP-GAN [16]	CodeFormer [21]	GDP [2]	DDNM [18]	DifFace [20]	**Ours**
LFW-Test	FID↓	84.02	72.45	74.10	118.04	122.43	67.98	71.62
	NIQE↓	4.98	3.90	4.52	8.60	9.24	5.47	4.15
WebPhoto-Test	FID↓	88.18	91.43	86.19	163.28	161.35	90.58	86.18
	NIQE↓	4.84	4.13	4.65	10.61	10.76	4.48	4.34
WIDER-Test	FID↓	71.31	40.93	40.26	193.20	153.99	38.54	39.17
	NIQE↓	4.83	3.77	4.12	14.33	11.68	4.44	3.93

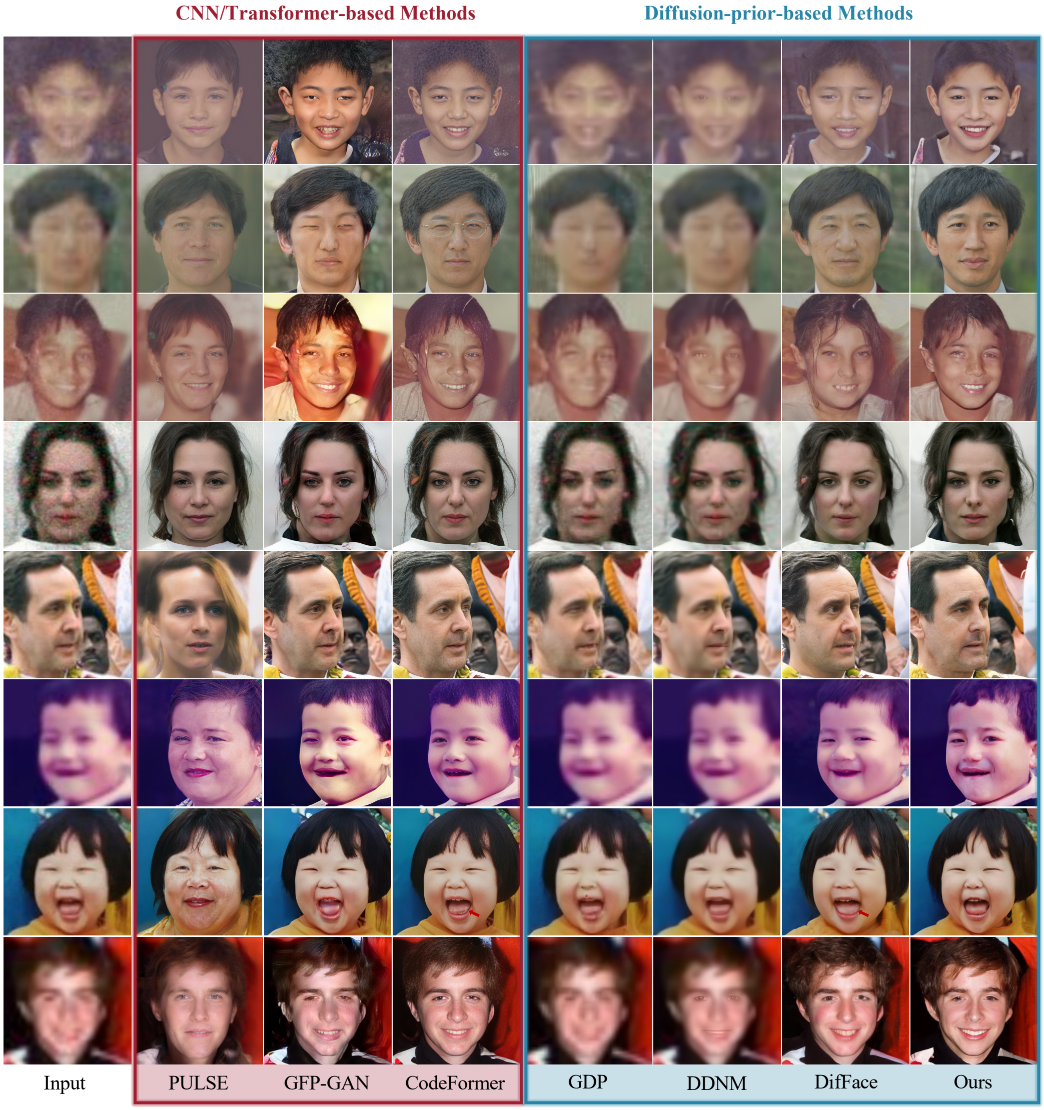


Figure 2: **Comparison on Blind Face Restoration.** Input faces are corrupted by real-world degradations. Our method produces high-quality faces with faithful details. (**Zoom in for best view**)

B.2 More Results on Old Photo Restoration

We provide more visual results of old photo restoration on challenging cases both with and without scratches, as shown in Fig. 3. The test images come from both the CelebChild-Test dataset [16] and the Internet. We compare our method with GFP-GAN (v1) [16] and DDNM [18]. Our method demonstrates an obvious advantage in sample quality esepecially in terms of vibrant colors, fine details, and sharpness.

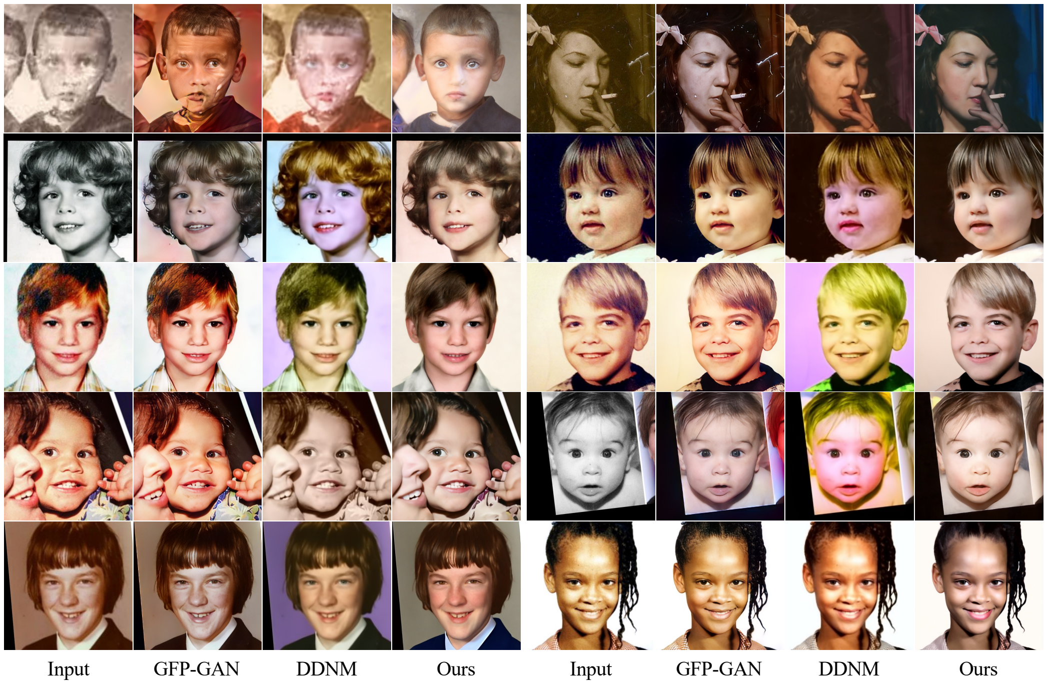


Figure 3: **Comparison on Old Photo Restoration on Challenging Cases.** Our method is able to produce high-quality restored outputs with natural color and complete faces.

B.3 More Results on Reference-Based Restoration

We provide more visual results on the reference-based restoration in Fig. 4, which is our exploratory extension based on blind face restoration. Test images come from the CelebRef-HQ dataset [8], which contains $1,005$ entities and each person has 3 to 21 high-quality images. With identity loss added, we observe that our method is able to produce personal characteristics similar to those of the ground truth.

B.4 More Results on Face Inpainting

In this section, we provide more qualitative comparisons with state-of-the-art methods in Fig. 5, including **(1)** task-specific methods: GPEN [19] and CodeFormer [21] and **(2)** diffusion-prior-based methods: GDP [2] and DDNM [18]. Since the pre-trained diffusion model DDNM employs [10] is trained on the CelebA-HQ dataset [4], we take the CelebRef-HQ [8] dataset for testing. Our method is able to recover challenging structures such as glasses. Moreover, diverse and photo-realism outputs can be obtained by setting different random seeds.

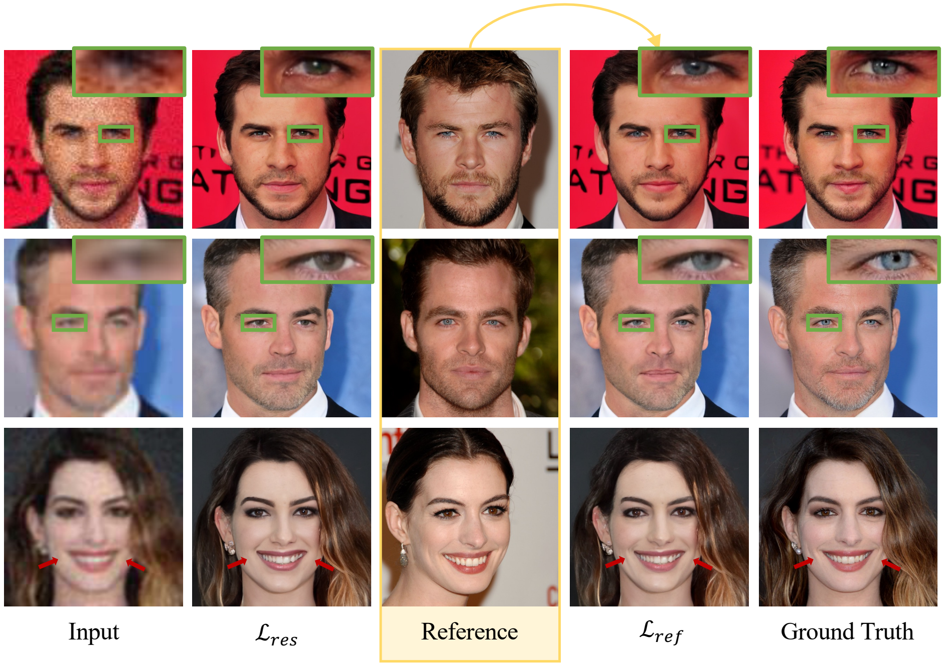


Figure 4: **Comparison on Reference-Based Face Restoration.** Our method produces personal characteristics which are hard to recover without reference.

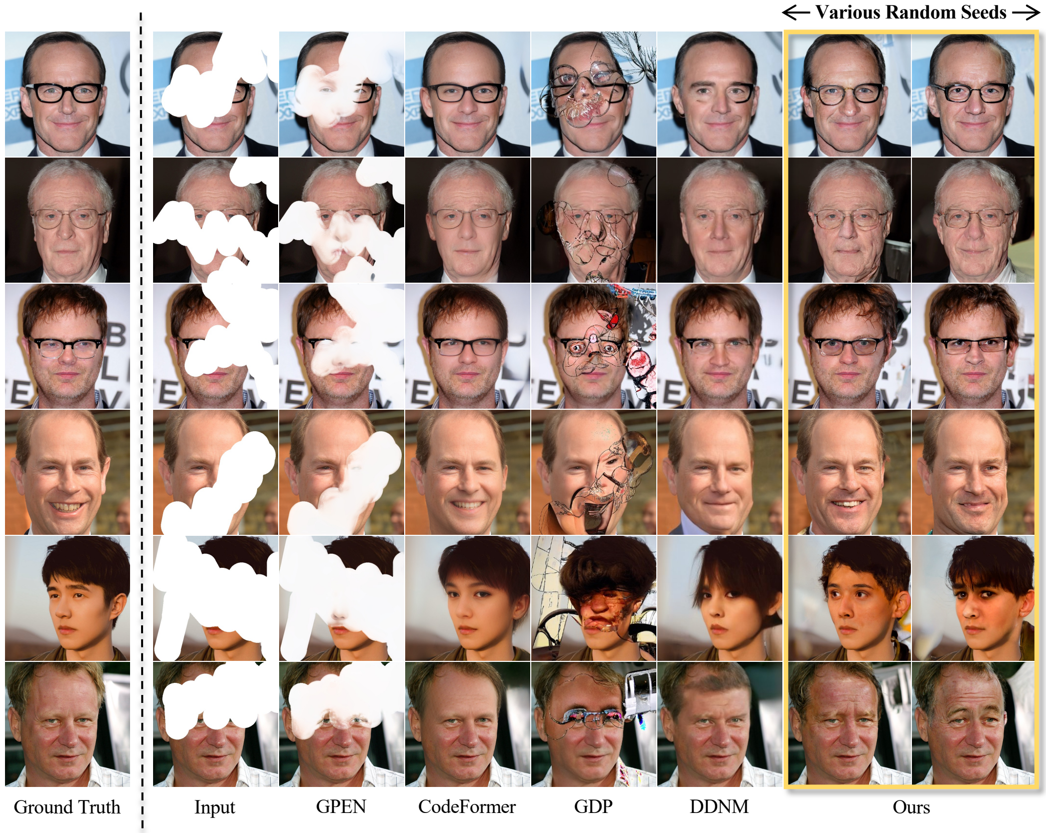


Figure 5: **Comparison on Face Inpainting on Challenging Cases.** Our method produces natural outputs with pleasant details coherent to the unmasked regions. Moreover, different random seeds give various contents of high quality.

B.5 More Results on Face Colorization

In this section, we provide more qualitative comparisons with state-of-the-art methods in Fig. 6, including **(1)** task-specific methods: CodeFormer [21] and **(2)** diffusion-prior-based methods: GDP [2] and DDNM [18]. Even though the test images come from the CelebA-HQ dataset [4], our method still produces more vibrant colors and finer details than DDNM. Moreover, our method demonstrates a desirable diversity by guiding with various color statistics.

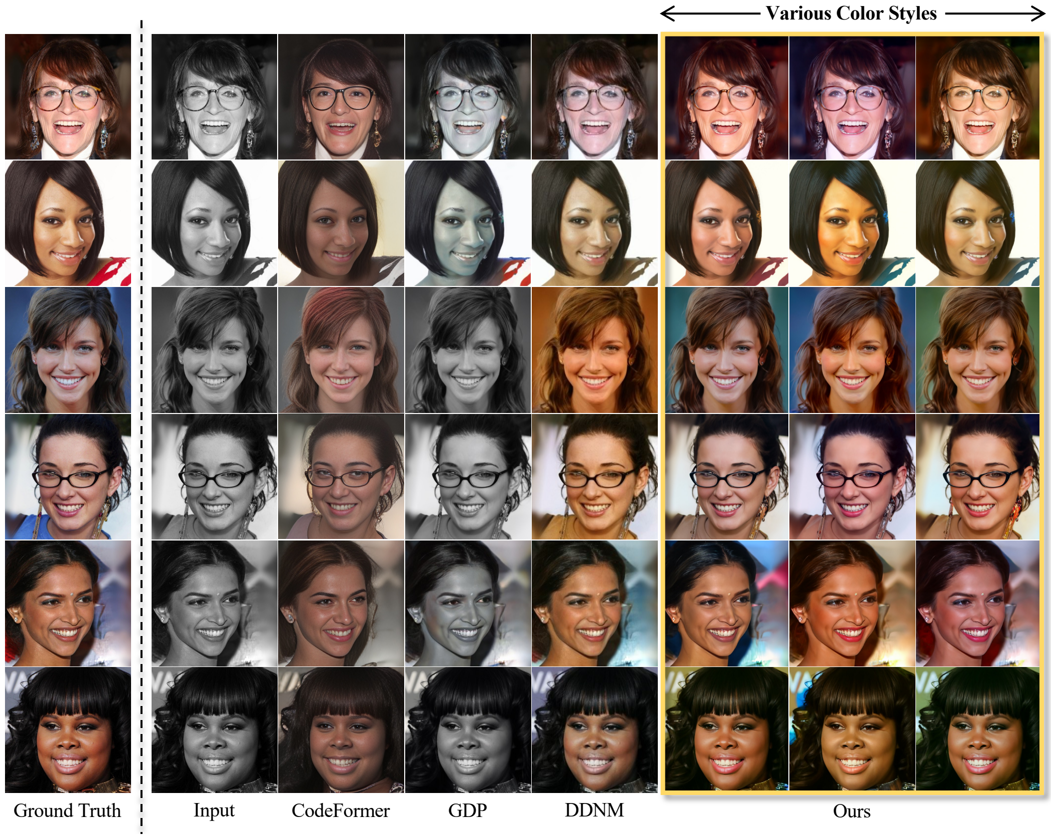
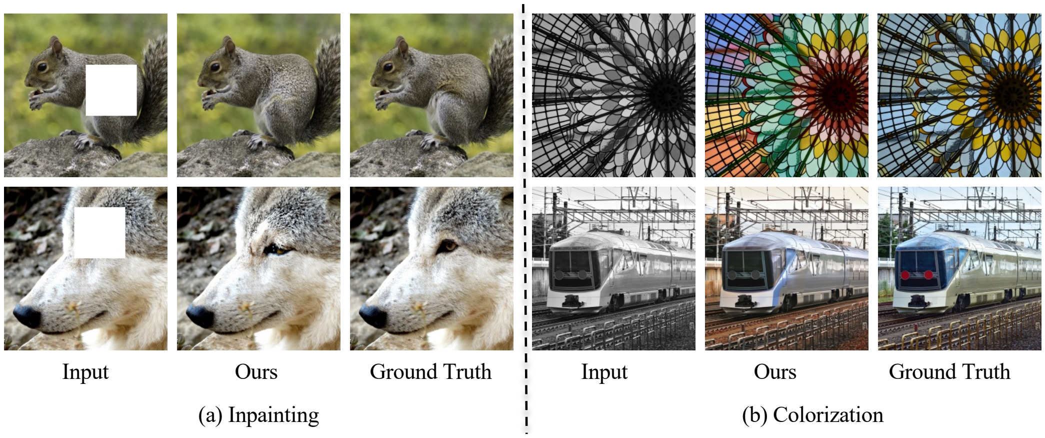


Figure 6: **Comparison on Face Colorization.** Our method produces diverse colorized outputs with various color statistics given as guidance.

C Limitations

As our partial guidance is based on a pre-trained diffusion model, our performance largely depends on the capability of the model in use. In addition, since a face-specific diffusion model is adopted in this work, our method is applicable only on faces in its current form. Nevertheless, this problem can be resolved by adopting stronger models trained for generic objects. For example, as shown in Fig. 7, we employ an unconditional 256 × 256 diffusion model trained on the ImageNet dataset [13] provided by [1], and achieve promising results on inpainting and colorization. Further exploration on natural scene restoration will be left as our future work.



(a) Inpainting

(b) Colorization

Figure 7: Extension on natural images for the inpainting and colorization tasks. By employing an unconditional 256 × 256 diffusion model trained on the ImageNet dataset [13] provided by [1], our method achieves promising results.

D Broader Impacts

This work focuses on restoring images corrupted by various forms of degradations. On the one hand, our method is capable of enhancing the quality of images, improving user experiences. On the other hand, our method could generate inaccurate outputs, especially when the input is heavily corrupted. This could potentially lead to deceptive information, such as incorrect identity recognition. In addition, similar to other restoration algorithms, our method could be used by malicious users for data falsification. We advise the public to use our method with care.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

[2] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[7] 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.

[8] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *TPAMI*, 2022.

[9] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV*, 2021.

[10] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[11] 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.

[12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

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

[17] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021.

[18] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023.

[19] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *CVPR*, 2021.

[20] Zongsheng Yue and Chen Change Loy. DifFace: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022.

[21] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022.