Supplementary Material for "High-Fidelity Image Inpainting with GAN Inversion"

Yongsheng Yu[1,2], Libo Zhang[1,2,3], Heng Fan[4], and Tiejian Luo[2]

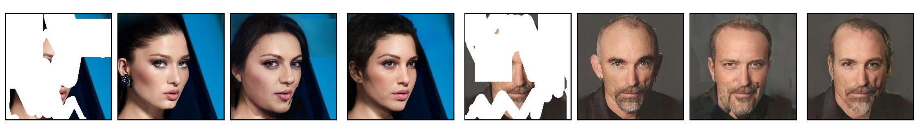
[1] Institute of Software, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Nanjing Institute of Software Technology
[4] Department of Computer Science and Engineering, University of North Texas
yuyongsheng19@mails.ucas.ac.cn; libo@iscas.ac.cn; heng.fan@unt.edu; tjluo@ucas.ac.cn

1 Implementation Details

In this work, the updating factor τ of soft-update mean latent is set to 0.001. In respect of the overall loss in Eq. 7, we use $\lambda_{\text{msr}} = 0.5, \lambda_{\text{fid}} = 0.005$. We train the encoder using Adam optimizer and set the batch size to 8 and the initial learning rate to $1e^{-4}$. For more diverse masking, we simply renew the mask generation based on [2] with controllable coverage and random square. Moreover, we practically notice that noise plays a trivial role in this work. To reduce variables, we set noise randomly sampled from a Gaussian distribution for each image generation.

2 Ablation Study

We visualize the comparison between $\mathcal{F}\&\mathcal{W}^+$ and W^+ in the Fig. 1. We can observe that our method with $\mathcal{F}\&\mathcal{W}^+$ settles the "gapping" issue and achieves better both qualitative results.



Masked | Original | w. $\mathcal{W}^+$ | w. $\mathcal{F}\&\mathcal{W}^+$ | Masked | Original | w. $\mathcal{W}^+$ | w. $\mathcal{F}\&\mathcal{W}^+$

Fig. 1. Visually comparing $\mathcal{F}\&\mathcal{W}^+$ and $\mathcal{W}^+$. Please zoom in.

The role of $\mathcal{L}_{\text{msr}}$ is to supervise the generated image from decoder and make final generation close to the original image. We conduct an ablation on λ_{msr} and the results are shown in Tab 1.

Table 1. Ablation of λ_{msr} on Places2.

	0.1	0.3	0.5	0.7
SSIM↑	0.629	0.644	0.652	0.647

3 Compared with Diffusion-based Method

The score-based diffusion models have recently shown high performance in many image generation tasks, including inpainting. We implement the recent Score-SDE [3] by official code and pre-trained CelebA-HQ weights (256 resolution). We show the comparison results in Figure 2 and Table 2. Noteworthy, Score-SDE takes about 314 seconds (on 1×A100 GPU) to infer an image.

Table 2. Quantitative comparison results on the *all* and *extreme* mask settings.

CelebA-HQ		SSIM	FID	LPIPS
all	Score-SDE	0.786	15.43	0.138
	Ours	**0.867**	**7.71**	**0.089**
extreme	Score-SDE	0.428	24.76	0.337
	Ours	**0.652**	**13.21**	**0.214**

Fig. 2. Qualitative comparison with diffusion-based Score-SDE approach.

4 Visual Results

We provide more qualitative results in Fig. 3 (each column arrange by Places2 [5], Metfaces [1], and Scenery [4] datasets from left to right) to evidence the effectiveness of our method.

Fig. 3. More qualitative results. Please zoom in.

References

1. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: NeurIPS (2020)
2. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: CVPR. pp. 7757–7765 (2020)
3. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
4. Yang, Z., Dong, J., Liu, P., Yang, Y., Yan, S.: Very long natural scenery image prediction by outpainting. In: ICCV. pp. 10560–10569 (2019)
5. Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI pp. 1452–1464 (2018)