Faithful Extreme Rescaling via Generative Prior Reciprocated Invertible Representations – Supplementary Material –

Zhixuan Zhong[1], Liangyu Chai[1], Yang Zhou[1], Bailin Deng[2], Jia Pan[3], Shengfeng He[1*]
[1] School of Computer Science and Engineering, South China University of Technology
[2] School of Computer Science and Informatics, Cardiff University
[3] Department of Computer Science, The University of Hong Kong

In this supplementary material, we first provide the implementation details of our **G**enerative prior **R**eciproc**A**ted **I**nvertible rescaling **N**etwork (GRAIN) in Sec. 1. And then we provide more visual results of GRAIN under different settings in Sec. 2.

1. Implementation Details

We adopt the pre-trained StyleGANv2 model [3] (trained on FFHQ [2], LSUN-Cat [7], or LSUN-Church [7] for diverse data domains) to produce the generative prior, whose weights are fixed during training. In the default setting, we downscale the HR image to the LR version with resolution 16×16 through our invertible encoder and then upscale it to 1024×1024 to better investigate the invertibility under extreme rescaling .

All training stages share a common component $\mathcal{L}_{base}$ as discussed in the main paper:

$$\mathcal{L}_{base} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_{LPIPS} + \lambda_3 \mathcal{L}_{id}, \tag{1}$$

where $\lambda_1 = 1.0$, $\lambda_2 = 0.8$ and $\lambda_3 = 0.1$ (for the domains other than human face, $\lambda_3 = 0$). We also apply adversarial loss $\mathcal{L}_{adv}$ with a constant $\lambda_{adv} = 0.01$ to encourage image generation with realistic details. In addition, in *Stage 1* and *Final Stage*, we adopt a $\mathcal{L}_2$ loss between the generated LR image and ground-truth LR image with $\lambda_{LR} = 0.1$ to produce semantically reasonable LR images.

Table 1 shows the model size (the number of parameters) and training/inference time of GPEN [6], GLEAN [1] and our method, on a machine with a Geforce RTX 3090. Note that we finetuned GPEN in 512×512-resolution due to the lack of released training code, hence it shows the best performance because of its lower scaling factor. Although our inference time is slightly higher than GLEAN, it achieves superior image quality with a much smaller model size.

*Corresponding author (hesfe@scut.edu.cn).

Table 1. Model size and training/inference time.

Methods	#Param (M)	Train (days)	Inference (ms)
GPEN [6] (32×)	71.01	1	101
GLEAN [1] (64×)	189.65	7	140
Ours (64×)	82.27	3	145

2. Qualitative Results

2.1. Ablation Studies

In Fig. 1, we show more visual results of ablation studies, including our invertible LR image, direct invertible rescaling output, StyleGAN output using our predicted codes (compared with pSp [5]), without invertible encoder output, image-level fusion output and final output with all modules. Our final setting with all modules can generate faithful details and maintain a good fidelity, while artifacts can be observed easily in other settings.

2.2. Comparisons with GAN Inversion Methods

We compare our framework with PULSE [4], pSp [5], GPEN [6], and GLEAN [1]. The comparison results are presented in Fig. 2, and our method demonstrates a superior performance in producing faithful and remarkable results. Note that we do not provide results of CNN-based face super-resolution methods and invertible face restoration methods as they are blurry and lack details, which can be found in the main paper.

2.3. Results on Unseen Faces

We perform experiments on face images collected from the Web which are unseen in the training to verify the generalization capacity of our method in real world. Fig. 3 shows the qualitative results and we can see that our method still

generates results close to the ground truth, while artifacts can be found easily in GPEN and GLEAN.

2.4. Results in Different Domains

Besides human face domain [3], we also extend our method to various domains such as Cat [8] and Church [7]. As illustrated in Fig. 4 and Fig. 5, our method is able to generate realistic results with plausible details in different domains, demonstrating the powerful generalization capacity.

2.5. Comparisons with JPEG Compression

Fig. 6 shows a visual comparison between our method (with different LR resolutions) and JPEG (with different image quality value). Our method achieves a good balance between image quality and storage size.

References

[1] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 1, 4

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1, 2

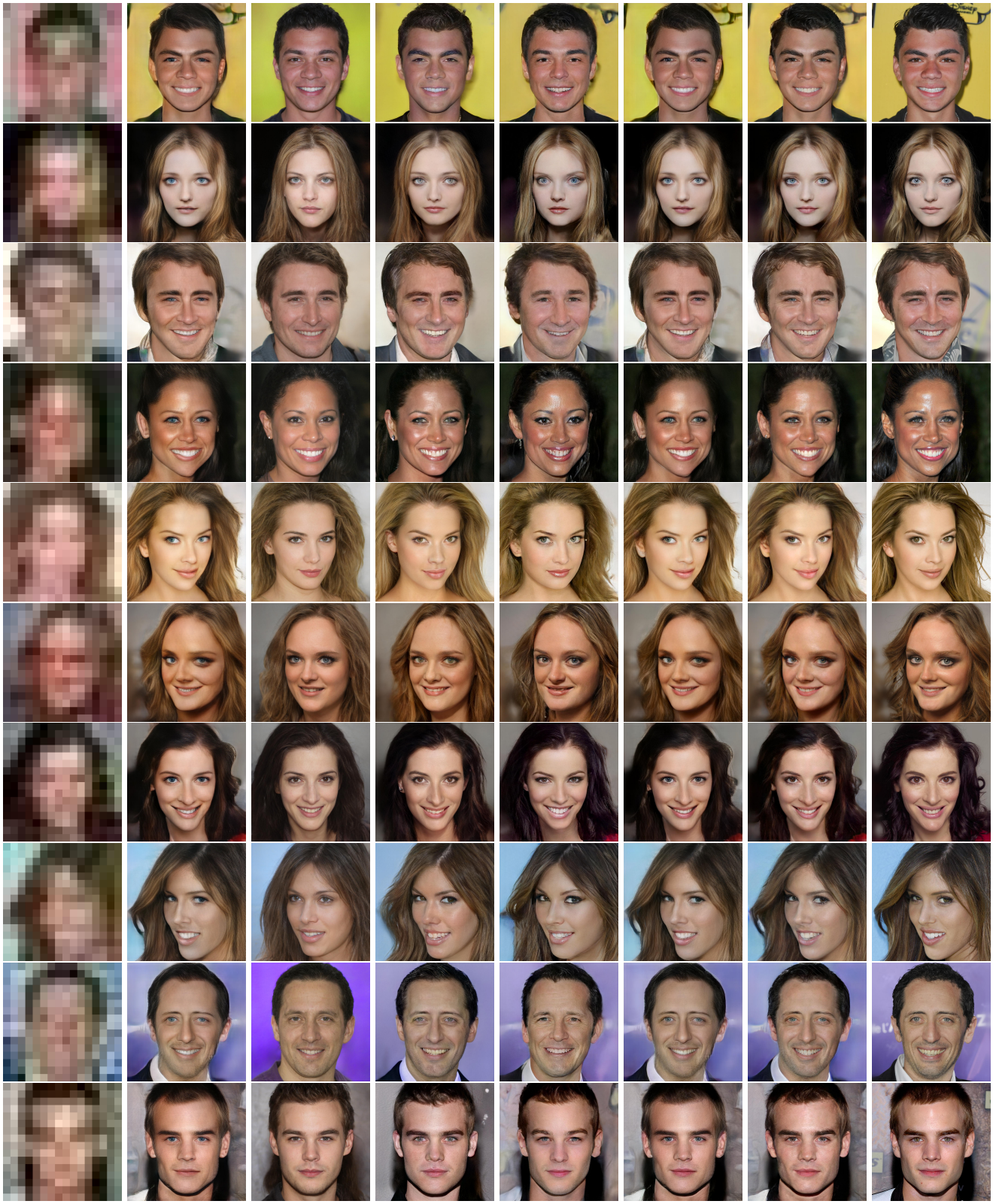
[4] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 1, 4

[5] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 1, 3, 4

[6] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021. 1, 4

[7] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 2

[8] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008. 2

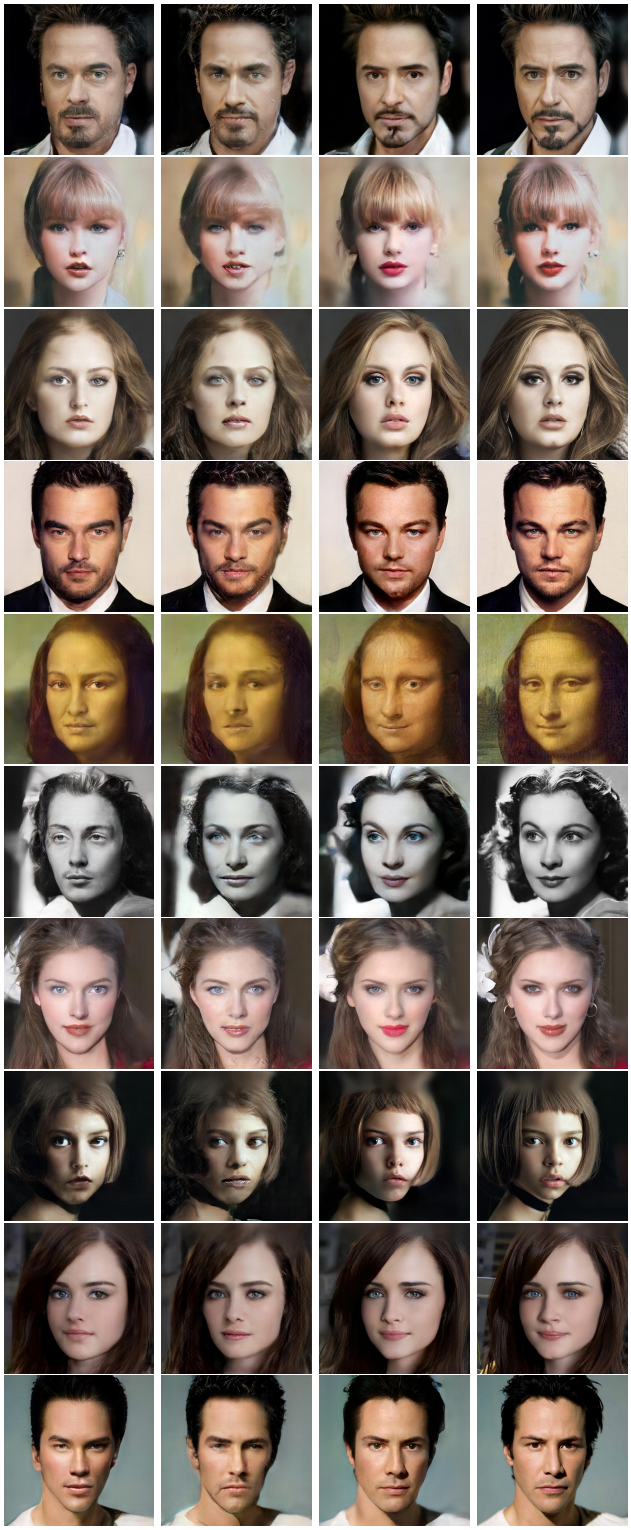


(a) LR (b) Invertible Only (c) pSp [5] (d) w/ Our Codes (e) w/o Invert. Enc. (f) Image-level (g) Ours (h) GT

Figure 1. Qualitative results on different variants of our method. (Zoom in for better view.)

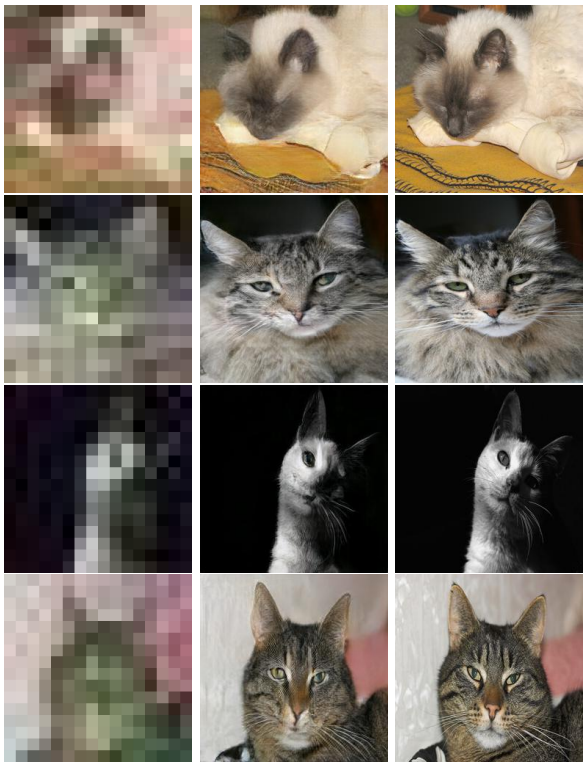
(a) PULSE [4] (64×) (b) pSp [5] (64×) (c) GPEN [6] (32×) (d) GLEAN [1] (64×) (e) Ours (64×) (f) GT

Figure 2. Qualitative comparisons with GAN inversion methods. (Zoom in for better view.)



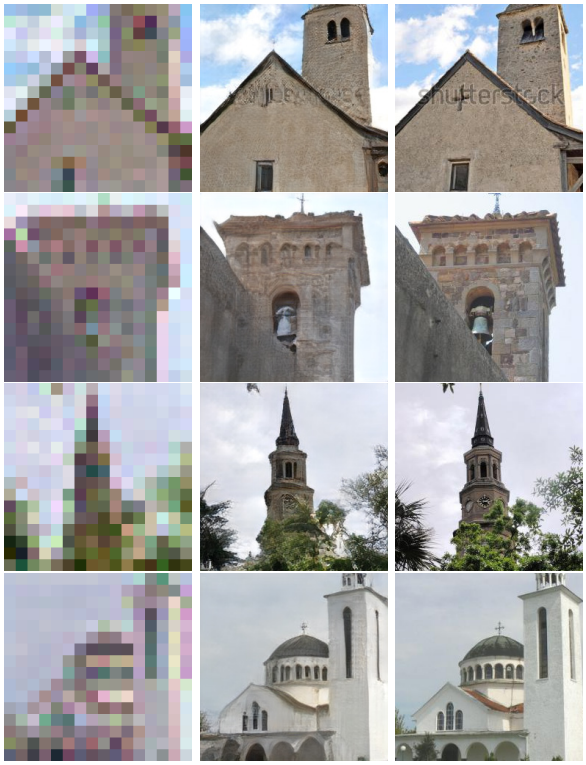
(a) GPEN (32×) (b) GLEAN (64×) (c) Ours (64×) (d) GT

Figure 3. Qualitative results on unseen faces.



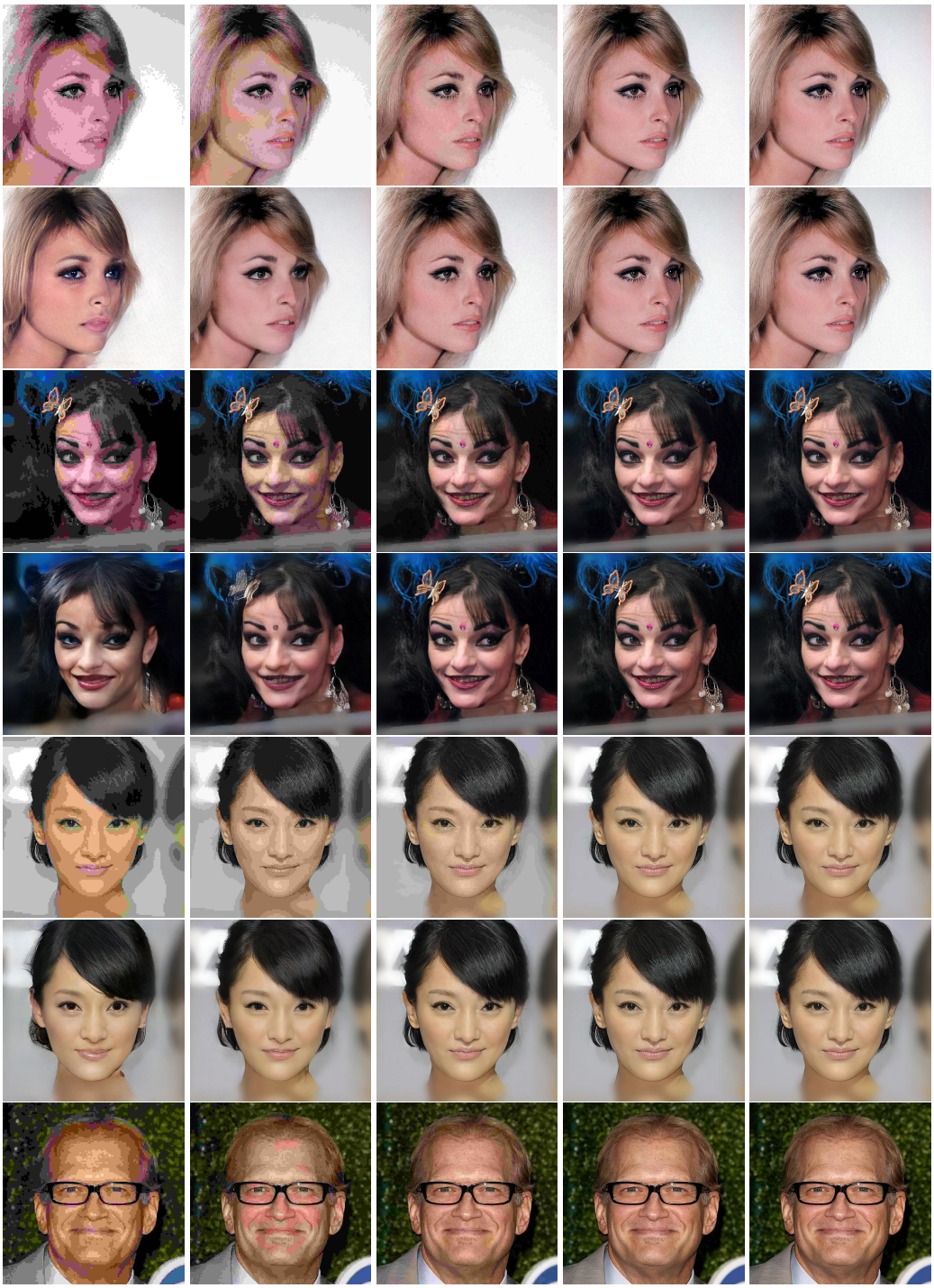
(a) LR (b) Ours (c) GT

Figure 4. More results in the cat domain.



(a) LR (b) Ours (c) GT

Figure 5. More results in the church domain.



(a) quality=1 (b) quality=5 (c) quality=10 (d) quality=40 (e) quality=70

(f) 16×16 (g) 32×32 (h) 64×64 (i) 128×128 (j) GT

Figure 6. Qualitative comparisons with JPEG.