Supplementary Material
ENTED: Enhanced Neural Texture Extraction and Distribution for Reference-based Blind Face Restoration

Yuen-Fui Lau
HKUST
yflauad@connect.ust.hk

Tianjia Zhang
HKUST
tzhangbl@connect.ust.hk

Zhefan Rao
HKUST
zraoac@connect.ust.hk

Qifeng Chen
HKUST
cqf@ust.hk

In the supplementary document, we would articulate the specific details of the proposed method, *ENTED*. In Section 1, we illustrate the basic experimental setup for our study. In Section 2 we explain the effect of how residual connection affects the final restored image. In Section 3, we discuss the impact of the reference image with different similarity levels. in addition, we present visual comparisons with state-of-the-art methods and more visual results on real-world data to illustrate the performance and generalization ability in Section 4. The limitation and future work are discussed in Section 5.

1. Implementation Setup

We adopt the well-known face dataset, FFHQ [3] for training and CelebA-HQ [5] for evaluation. We categorized face images in CelebA-HQ into groups according to their identity. We randomly choose five images from each group, one as the input image and the other four as the HQ reference image. We get a total of 2398 groups of images. The input images and the reference images are down-sampled to 512×512 via bilinear interpolation. The FFHQ dataset is used to train *ENTED*.

The comparison baseline methods are PSFRGAN [1], DFDnet [4], GFP-GAN [7], GPEN [8] and VQFR [2], the same as those used in the main paper. We use the pre-trained models and implementations provided by their official repositories.

2. Effects of Residual Connections

Without the establishment of residual connection, the fidelity of face features decreased dramatically, as seen in Figure 1. In the neural texture extraction and distribution framework, we found that residual connections have an essential function in preserving fidelity. Observed that the degraded input image doesn't lose all of the fidelity details, some of the fidelity information is still remaining in the degraded input image. The residual connection permits communication between the content encoder and decoder, facilitating the exchange of fidelity information between them. However, the information flow from the content encoder, on the other hand, contains noise. We add weight between the preceding layer output and the residual connection to minimize the influence of noise on the decoding process while keeping the fidelity information flow. The whole residual connection is

$$F_D^l = F_o^l + \lambda \cdot F_c^l, \tag{1}$$

where λ is the weight factor for residual mapping and we choose $\lambda = 0.2$ through all our experiments.

3. Effects in Different Similarity Levels of Reference Images

We also conduct experiments on evaluating how the similarity level affects the quality of the final restored image. To evaluate how similar the target image and the reference image are, we divide reference images into 4 levels of similarity and assign them to the corresponding level by evaluating the **LPIPS** value between the reference image and target image. Table 1 demonstrates the results of the experiments. We found that the performance in **LPIPS** improves as the similarity level increase. However, it can be observed that the performance in **FID** and **NIQE** don't increase with the similarity level. This could be due to the fact that we categorize the similarity level based on the **LPIPS** value, and images at the corresponding similarity levels encourage improving the model performance in terms of **LPIPS**.

Figure 1. Using the same reference image and experimental setting for each face restoration, the output without residual connection exhibits degradation on the facial identity, but the result with residual connection has more high-quality facial details that are consistent with the ground truth image.

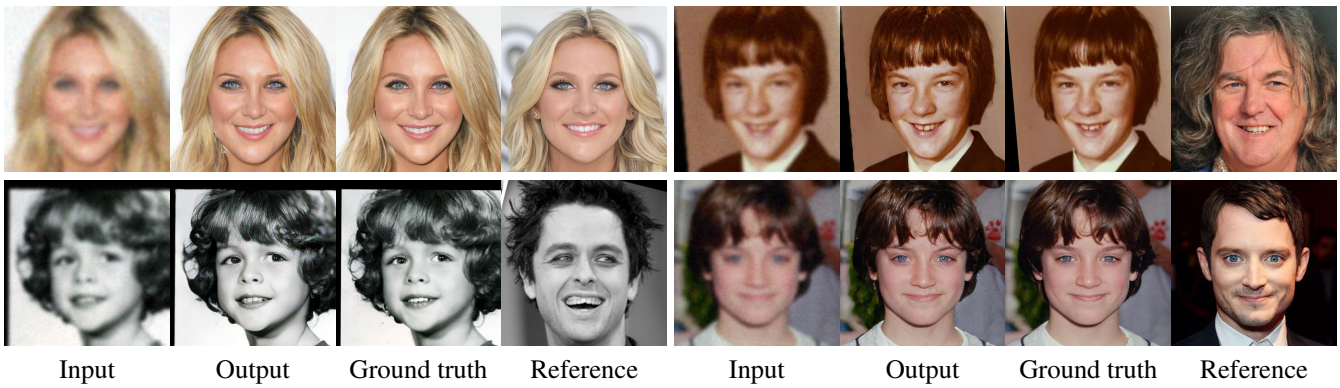

Figure 2. Our face restoration results with reference images. The top left-hand corner demonstrated the results on synthesized data and the remaining three examples illustrate the results on real-world old photos.

4. Detailed Results

We illustrate a quantitative comparison with state-of-the-art approaches on real-world datasets in the main paper, and here we present a visual comparison with state-of-the-art methods. Figure 3 depicts how *ENTED* produces images with finer-grained details in various facial characteristics such as skin texture, ears, teeth, hair pattern, etc. Moreover, we present more visual restoration results on old photos. Figure 2 shows visual results on both black and white images and browned images. Our results show that *ENTED* can restore images with consistent color range, the least domain gap, and the highest perceptual similarity to the ground truth.

5. Limitation

Unlike conventional single-image super-resolution, *ENTED* is a reference-based image restoration approach that requires additional reference previously. This complicates the training process because most public datasets are not supposed to provide additional reference images. On the other hand, it is not always easy to obtain a high-quality image as a reference for real-world applications. Also, the quality of the reference image influences the quality of the restoration results. In the future, we will try to overcome the limitation of the extra reference image by testing the performance of different connection modes and exploring more structures to transfer prior information from the pretraining dataset, such as using a more powerful generator presented in the work of stable diffusion [6].

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*, pages 11896–11905, 2021. 1

[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. *arXiv preprint arXiv:2205.06803*, 2022. 1

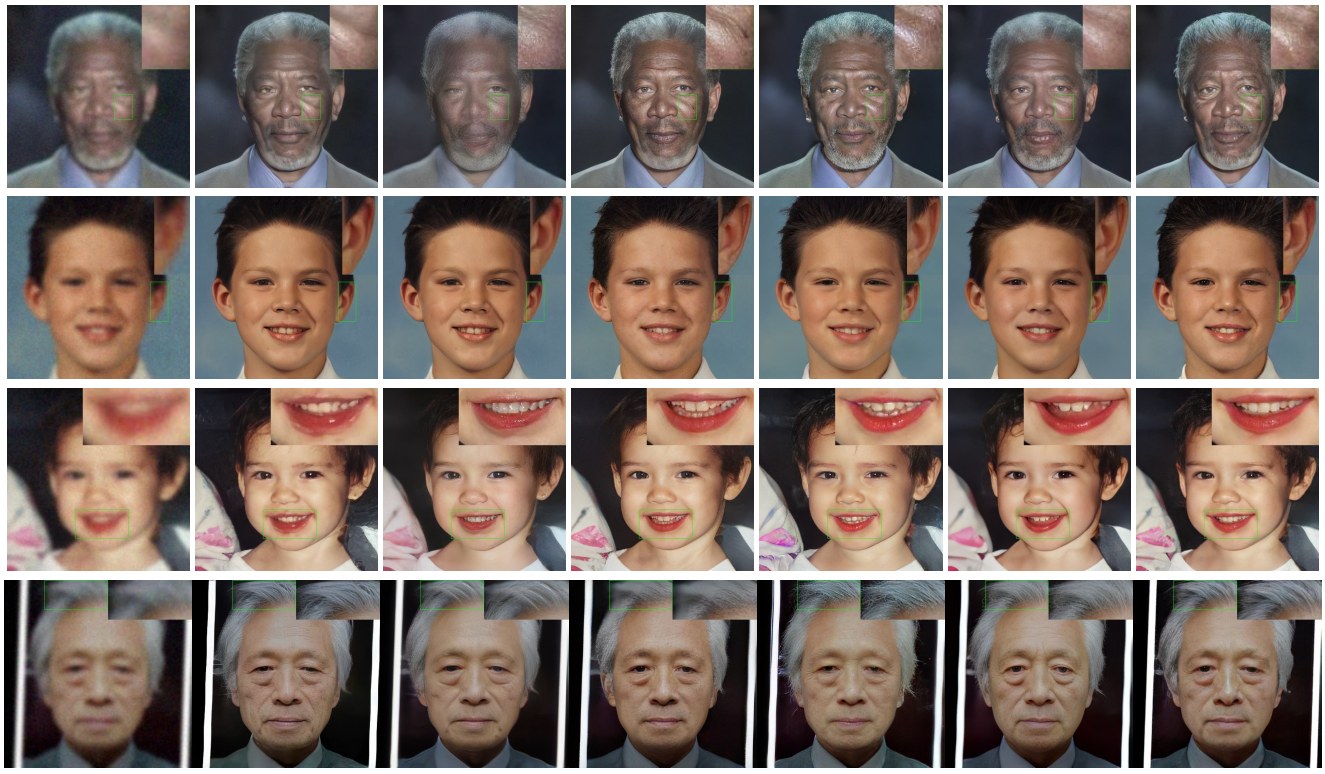

Figure 3. Visual comparison on real-world datasets.

Table 1. Quantitative results with different levels of the reference image. We compare reference images with the corresponding ground truth image based on the LPIPS value between them. "L1" denotes the most relevant reference image, which has the lowest LPIPS with the corresponding ground truth image, and "L4" denotes the least relevant reference image, which has the highest LPIPS with the ground truth image.

Similarity level	**LPIPS** ↓	**FID** ↓	**NIQE** ↓
L1	0.2156	12.80	3.60
L2	0.2160	12.82	3.61
L3	0.2166	12.82	3.60
L4	0.2168	12.80	3.59

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

[4] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, pages 399–415. Springer, 2020. 1

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 1

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[7] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021. 1

[8] 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