Implicit Diffusion Models for Continuous Super-Resolution Supplementary Material

Sicheng Gao[1]*, Xuhui Liu[1]*, Bohan Zeng[1]*, Sheng Xu[1], Yanjing Li[1], Xiaoyan Luo[1]
Jianzhuang Liu[2], Xiantong Zhen[3], Baochang Zhang[1,4]†
[1]Beihang University [2]Shenzhen Institute of Advanced Technology, Shenzhen, China
[3]United Imaging [4]Zhongguancun Laboratory, Beijing, China

In this supplementary material, we first provide more details about the evaluation metrics in Section A. Then we show more quantitative comparisons and visualizations on various categories and magnifications and further demonstrate the resolution-continuous results in Section B. Finally, we state the ethical impact in Section C.

A. Metrics

We provide more detailed description about the metrics used in the main paper below:

Peak Signal-to-Noise Ratio (PSNR). PSNR reflects the image reconstruction quality. Because the PSNR value (dB) has limitations in super-resolution evaluation [8], it is only a reference value of image quality between the maximum signal and the background noise. The larger the PSNR value, the less image distortion.

Structure Similarity Index Measure (SSIM). From the perspective of image distortion modeling discussed in [11], SSIM implements the related theory of structural similarity by imitating the human visual system, and is sensitive to the perception of local structural changes of the image. SSIM quantifies image properties from brightness, contrast, and structure, using mean to estimate brightness, variance to estimate contrast, and covariance to estimate structural similarity.

Consistency. Consistency calculates the MSE $(\times 10^{-5})$ between the low-resolution inputs and the corresponding downsampled SR results.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS aims at measuring the perceptual similarity between the generated and real images by using their deep features.

Frechet Inception Distance score (FID) [5] evaluates the image quality by imitating human perception of image similarity. We leverage a pre-trained Inception-V3 [10] to compare the distributions of the generated images with those of the ground-truth images.

Cosine Similarity (CSIM) measures the difference between two individuals by using the cosine value of the angle between two vectors in a vector space. We leverage the pre-trained MoCo [4] to compute the cosine similarity to assess the quality of identity preservation.

*These authors contributed equally.
†Corresponding Author: bczhang@buaa.edu.cn.

B. More Results

B.1. Quantitative Results

Quantitative Comparison with GLEAN [2]. We provide more experiment results in Table 1, including LSUN-Cars [12] and face SR. It needs to be explained that the results of GLEAN on LSUN-Cars are reproduced with the official codes[1] by testing on the same 100 images as IDM.

Table 1. PSNR comparison on $8\times$ and $16\times$ SR on CelebA-HQ [6] and $16\times$ SR on LSUN-Cars [12].

Dataset	GLEAN [2]	IDM (ours)
Cars ($16\times$)	20.22	**20.27**
Face ($8\times$)	23.45	**24.01**
Face ($16\times$)	22.58	**23.00**

Quantitative Comparison of the Ablation Study. We give quantitative results of the ablation study in this section. As shown in Table 2, encoded features by EDSR [7] achieves 0.41dB improvement in PSNR. Moreover, our LR conditioning mechanism exhibits the best 21.52dB PSNR result.

Table 2. PSNR comparison on $16\times$ SR on LSUN-Cats [12] with different conditioning structures.

Method	Concat	Concat w/ encoder	IDM (ours)
PSNR	20.68	21.09	**21.52**

Further analysis. We first analyze the results in terms of PSNR in Table 4 in the main paper. Regression-based methods (LIIF) directly minimize the L1/L2 loss, and they produce less high-frequency details, leading to higher PSNR at the cost of over-smoothing details. As shown in Table 3, LIIF [3] reports poor results in terms of FID. Therefore, recent generative SR methods often do not use PSNR to compare their performance with regression-based methods. In Table 4 in the main paper, we use the double line to divide regression-based and generative methods.

Moreover, to validate semantic identity information consistency, we compare the $8\times$ SR outputs of face images in terms of cosine similarity (CSIM) of MoCo features. Results of Table 3 show that IDM surpasses prior arts in preserving identity information.

Table 3. Quantitative Comparison on $8\times$ face SR in FID and CSIM.

	LIIF [3]	GLEAN [2]	SR3 [9]	IDM (ours)
FID	105.0	81.38	72.60	**56.06**
CSIM	0.8051	0.9330	0.9524	**0.9587**

[1] https://github.com/ckkelvinchan/GLEAN

Running time. We further give the comparison result of the inference speed in terms of FPS (Frames Per Second): SR3: 0.0171, IDM: **0.0202**. The inference speed of IDM is affected by the iterative denoising process. Thus IDM is slower than other regression and GAN methods, but 18.13% faster than SR3 due to the simplification of network structure.

B.2. Visualization

General Scene Super-Resolution. To demonstrate the generic effectiveness of our IDM, we select eight challenging examples from the general scene dataset DIV2K [1], such as natural perspectives, buildings, animals and plants. And we compare IDM with LIIF [3] on $4\times$ SR in Fig. 1, Fig.2, Fig. 3, Fig. 4, and Fig. 5, which clearly show that our IDM achieves much better results.

Face Super-Resolution. We compare IDM with LIIF [1] and SR3 [9] for the $8\times$ face SR in Fig. 6, where the ground-truth images have a resolution of 128×128. The images are randomly-sampled from CelebA-HQ [6]. We find that LIIF suffers from a critical over-smoothing problem, and SR3 loses some realistic facial details. In contrast, IDM achieves photo-realistic face generation that is highly consistent with the ground-truth.

Natural Image Super-Resolution. We show multiple results of randomly-sampled examples from the testing datasets of Cats, Bedrooms, and Towers in LSUN [12] in Fig. 7, Fig. 8 and Fig. 9. These results again validate the remarkable ability of IDM in synthesizing high-fidelity SR images.

B.3. Visualization of Continuous Super-Resolution

To demonstrate the effectiveness of IDM in generating resolution-continuous images, we provide more comparison results of $16\times$ face SR with LIIF [1] on the testing dataset of CelebA-HQ [6] in Fig. 10, where the resolution of corresponding ground-truth is 256×256. Although LIIF can generate relatively stable resolution-continuous images, it performs poorly in fine detail synthesis. IDM exhibits very good performance not only in producing resolution-continuous images, but also in synthesizing high-quality details. Especially, even if the magnification is out of the training range, *i.e.* $17\times$ and $18\times$, IDM is still effective to generate high-fidelity SR images.

C. Ethic Impact

This work can be used for the human face super-resolution task which is common in mobile phone photographing. It does not have a direct negative social impact. Because of personal security, we should prevent it from being abused for malicious purposes.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 3, 5, 6, 7, 8, 9

[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021. 2, 3

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[5] 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. 1

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017. 2, 3, 10, 14

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2

[8] 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. 1

[9] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 2, 3

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 1

[12] 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:1506.03365*, 2015. 2, 3, 11, 12, 13

Figure 1. Comparison of 4× SR on the DIV2K dataset. **Upper Part:** LIIF [1]; **Lower Part:** IDM. IDM synthesizes more fine-grained details and achieves remarkable performance. Best viewd by zooming in.

Figure 2. Comparison of 4× SR on the DIV2K dataset. **Left Part:** LIIF [1]; **Right Part:** IDM. IDM synthesizes more fine-grained details and achieves remarkable performance. Best viewd by zooming in.

Figure 3. Comparison of 4× SR on the DIV2K dataset. **Left Part:** LIIF [1]; **Right Part:** IDM. IDM synthesizes more fine-grained details and achieves remarkable performance. Best viewd by zooming in.

Figure 4. Comparison of 4× SR on the DIV2K dataset. **Left Part:** LIIF [1]; **Right Part:** IDM. IDM synthesizes more fine-grained details and achieves remarkable performance. Best viewd by zooming in.

Figure 5. Comparison of 4× SR on the DIV2K dataset. **Left Part:** LIIF [1]; **Right Part:** IDM. IDM synthesizes more fine-grained details and achieves remarkable performance. Best viewd by zooming in.

Figure 6. Visual comparison of 8× SR on CelebA-HQ [6].

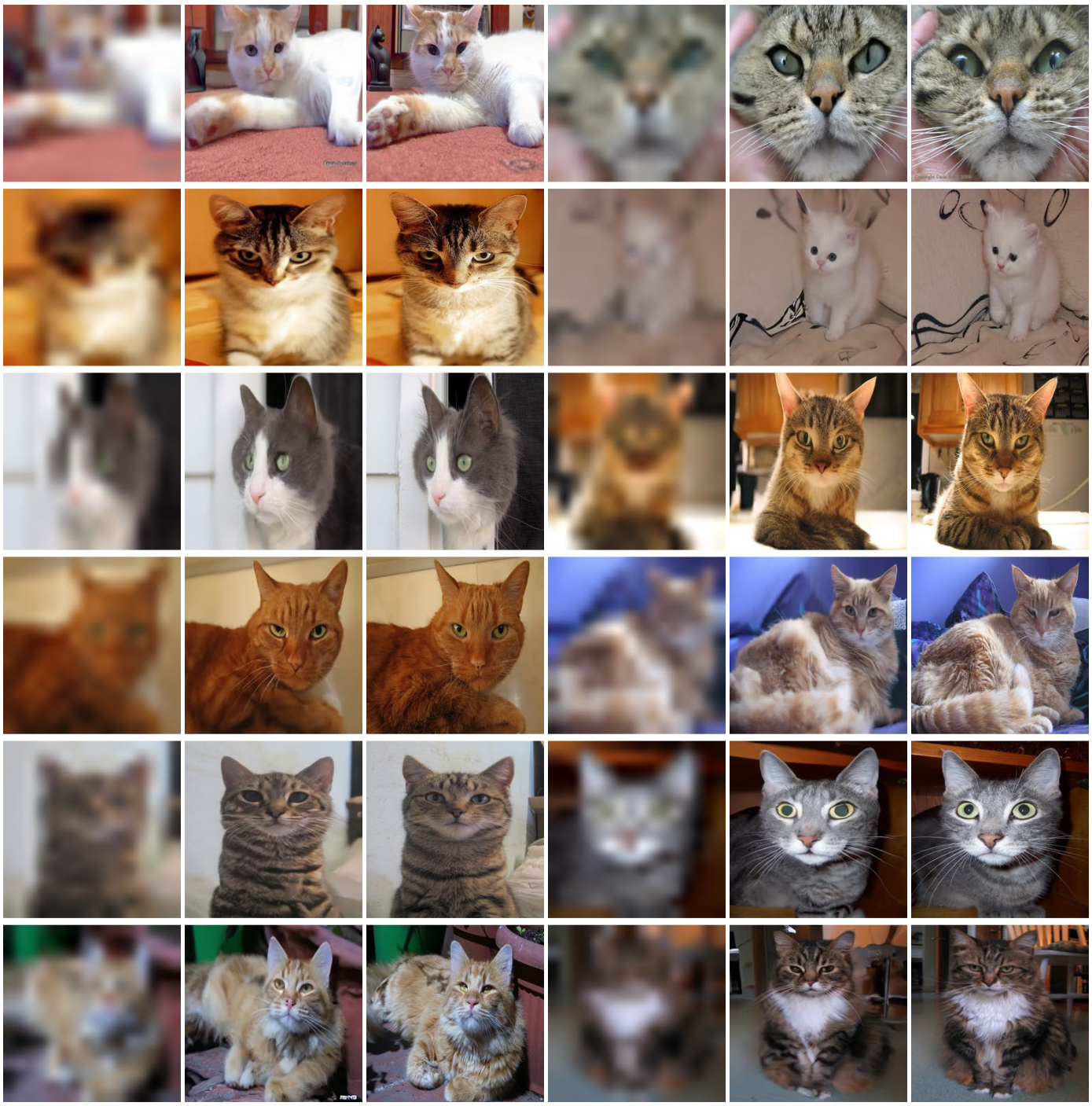

Figure 7. Results of 16× SR on the Cats dataset of LSUN [12]. IDM achieves consistent textures and details with the ground-truth.

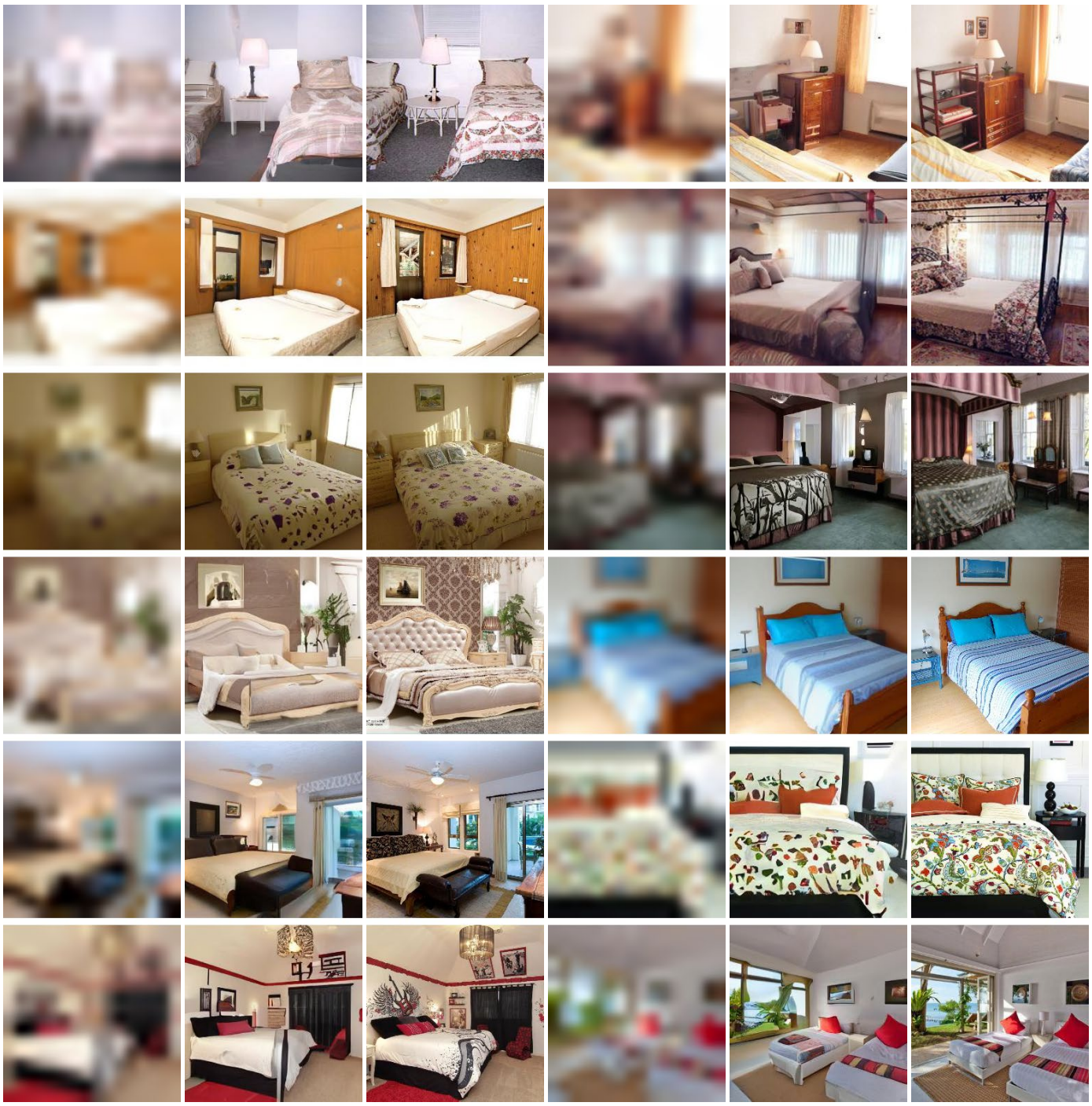

Figure 8. Results of 16× SR on the Bedrooms dataset of LSUN [12]. IDM achieves consistent textures and details with the ground-truth.

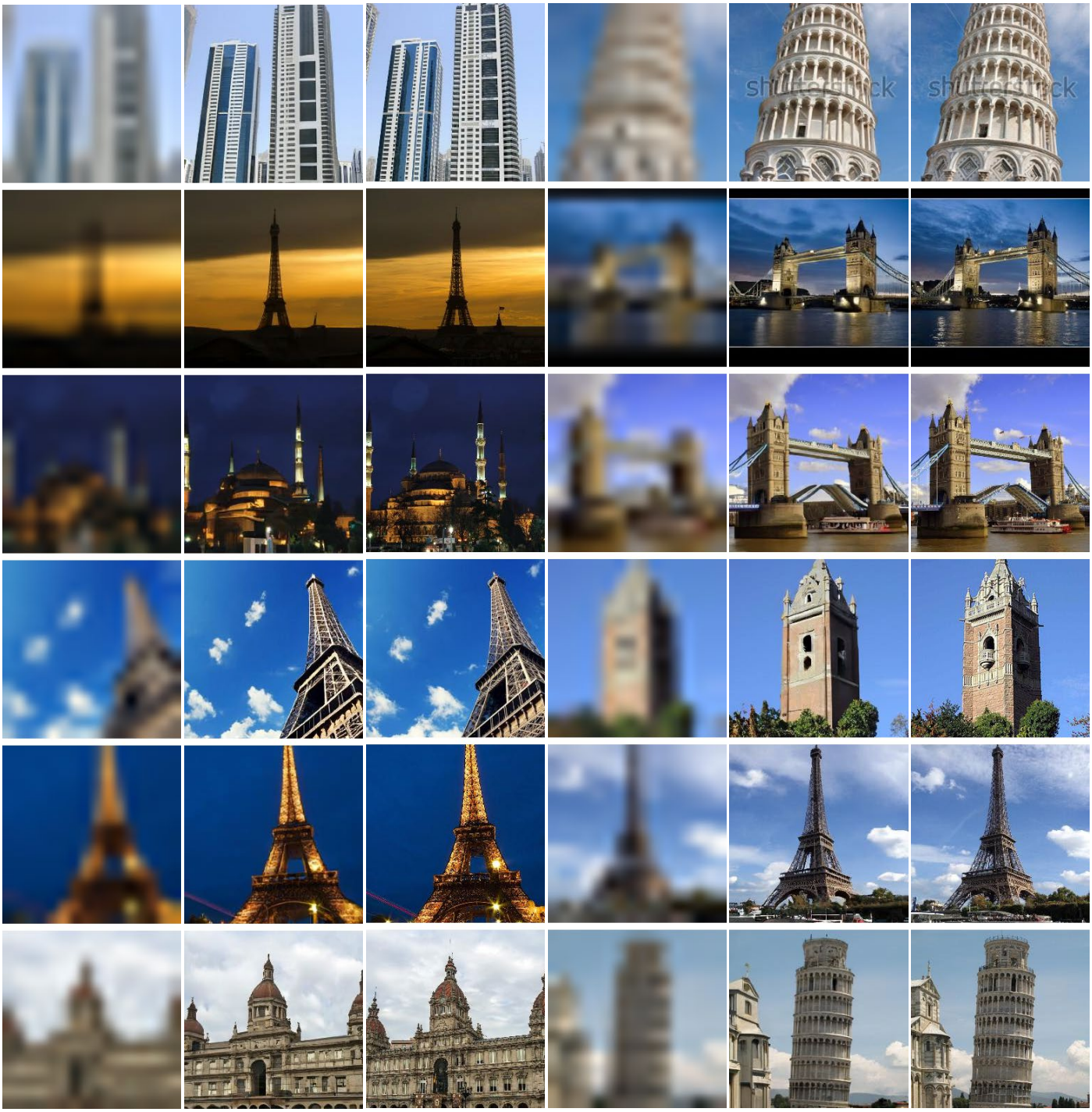

Figure 9. Results of 16× SR on the Towers dataset of LSUN [12]. IDM achieves consistent textures and details with the ground-truth.

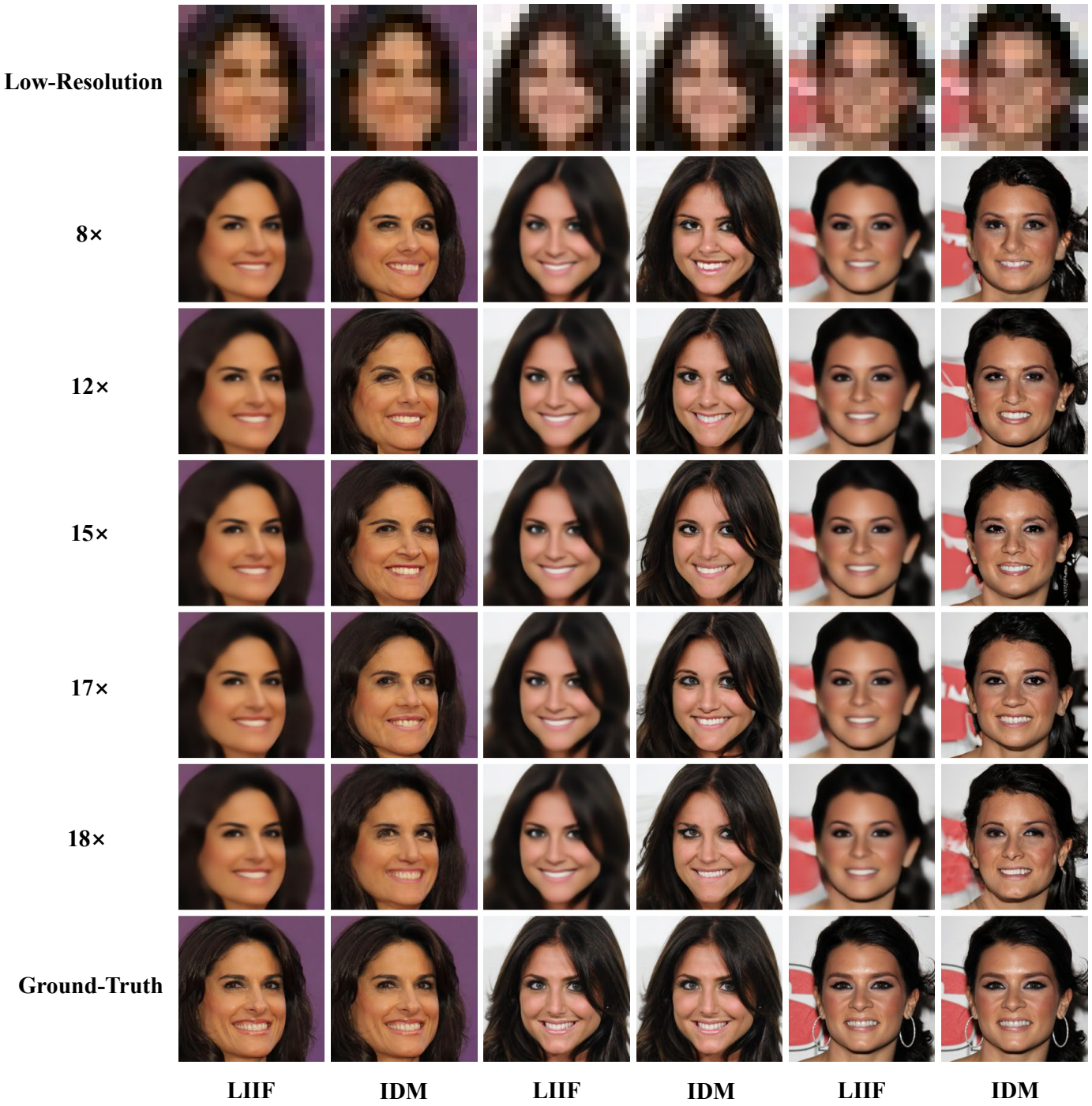

Figure 10. Visualization of continuous SR results on CelebA-HQ [6] when training on 16× SR.