Supplementary Material to "Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution"

Jie Liang[1]*, **Hui Zeng**[2]* and **Lei Zhang**[1]†
[1]The HongKong Polytechnic University, [2]OPPO Research
{*liang27jie, cshzeng*}*@gmail.com*; *cslzhang@comp.polyu.edu.hk*

In this supplementary file, we provide the following materials:

- More visual comparisons on SISR with synthetic degradation under scaling factor $4\times$ (referring to Section 4.2 in the main paper);
- More visual comparisons on real-world SISR under scaling factor $4\times$ (referring to Section 4.3 in the main paper);
- Quantitative comparisons on SISR with synthetic degradation under scaling factor $2\times$ (referring to Section 4.2 in the main paper).

1. More Visual Comparisons of SISR on Images with Synthetic Degradation

We first show the visual comparisons of $4\times$ SISR on images with bicubic degradation by using three backbones. In specific, Figure 1 compares the methods with light-weight backbone. Figure 2 compares the RRDB-based methods (see also Figure 7 in the main paper), and Figure 3 compares the SwinIR-based ones. Consistent observations with the main paper can be made from these visual comparisons, where the proposed LDL inhibits the visual artifacts and simultaneously recovers richer and realistic details. This validates the generalization capability of the proposed LDL to different types of backbones.

2. More Visual Comparisons of SISR on Real-World Images

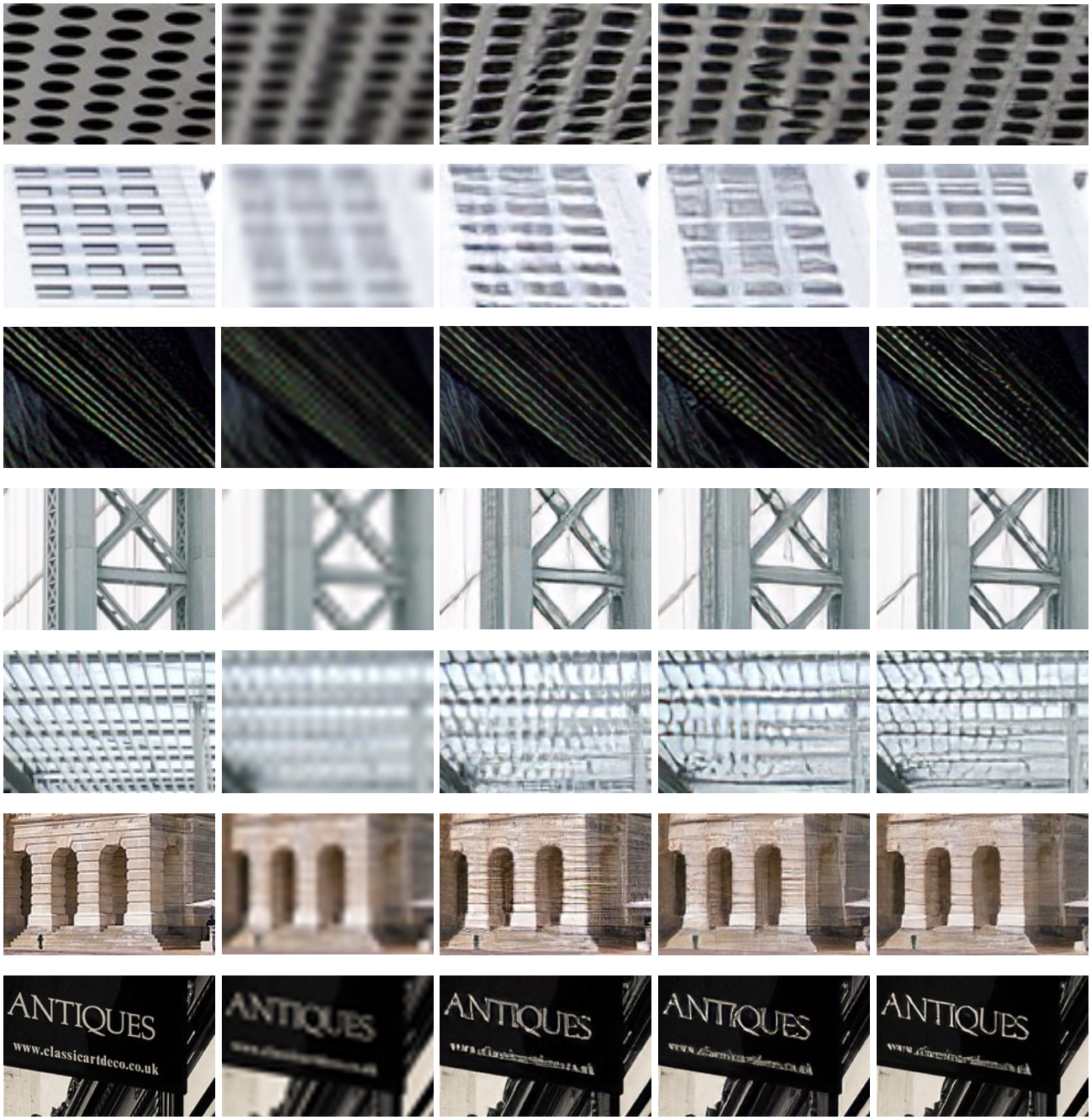
We then show the visual comparisons of $4\times$ SISR on real-world low-resolution images. Following the same comparison strategy as in the main paper, we first train a baseline model by using a specific backbone network (*i.e.*, SRResNet, RRDB or SwinIR) with the degradation model of RealESRGAN [4], and then apply the proposed $\mathcal{L}_{\text{artif}}$ loss to the baseline to train the LDL model. The BSRGAN [8] method is also compared if the corresponding official model is released.

Figure 4 compares the GAN-SR methods with SRResNet backbone, Figure 5 compares the RRDB-based methods and Figure 6 compares the SwinIR-based ones. As can be seen, the proposed LDL improves the SISR image quality over competing methods. It suppresses significantly the artifacts caused by the complicated and unknown degradation in real-world SISR tasks. This also demonstrates the generalization performance of our method to real-world SISR tasks.

3. More Quantitative Comparisons for $2\times$ SISR

To validate the effectiveness of our method in generalizing to different scaling factors, we further conduct experiments on $2\times$ SISR using the three backbones (SRResNet, RRDB and SwinIR), and report the results in Table 1. Since most of the existing GAN-SR methods [1,3,4,6] do not train the $2\times$ SISR models, we train three baseline models using SRResNet, RRDB and SwinIR as backbone for comparison. As can be seen from Table 1, the proposed LDL achieves consistent improvement on most benchmarks in terms of both perceptual quality (LPIPS, DISTS and FID) and reconstruction accuracy (PSNR and SSIM). This demonstrates the effectiveness of our LDL in generalizing to different scaling factors.

*Equal contribution.
†This work is supported by the Hong Kong RGC RIF grant (R5001-18).



(a) HR (b) Bicubic (c) SFTGAN (d) SRGAN (e) SRResNet+LDL

Figure 1. Visual comparison among light-weight GAN-SR methods, including SFTGAN [5], SRGAN [1] and our SRResNet+LDL, under the scaling factor of $4\times$ with bicubic degradation. Here, both SRGAN and our SRResNet+LDL use SRResNet as the backbone network, and SFTGAN has similar magnitude of parameters in its restoration module. As can be seen, our method achieves clear improvement in reconstructing realistic details. For example, the regular patterns of buildings in the 1^{st}, 2^{nd} and 4^{th} rows are better restored by our method compared to SFTGAN and SRGAN. The textures in the 3^{rd} (feathers) and the last (text) row are more continuous and sharp. Besides, our method has clear advantages in inhibiting artifacts. For example, in the second last row, the surface of the building in our result is cleaner than SFTGAN.

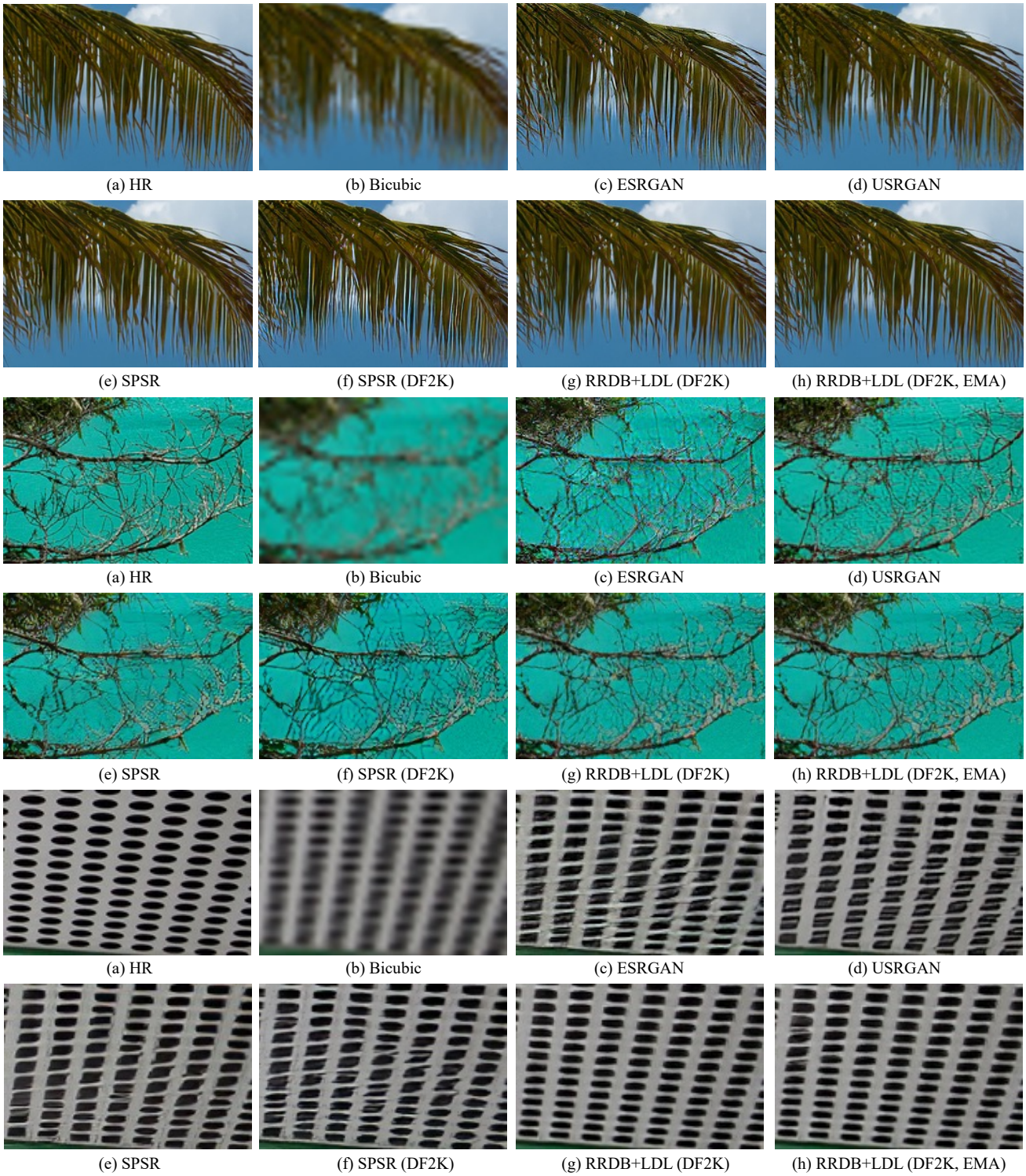
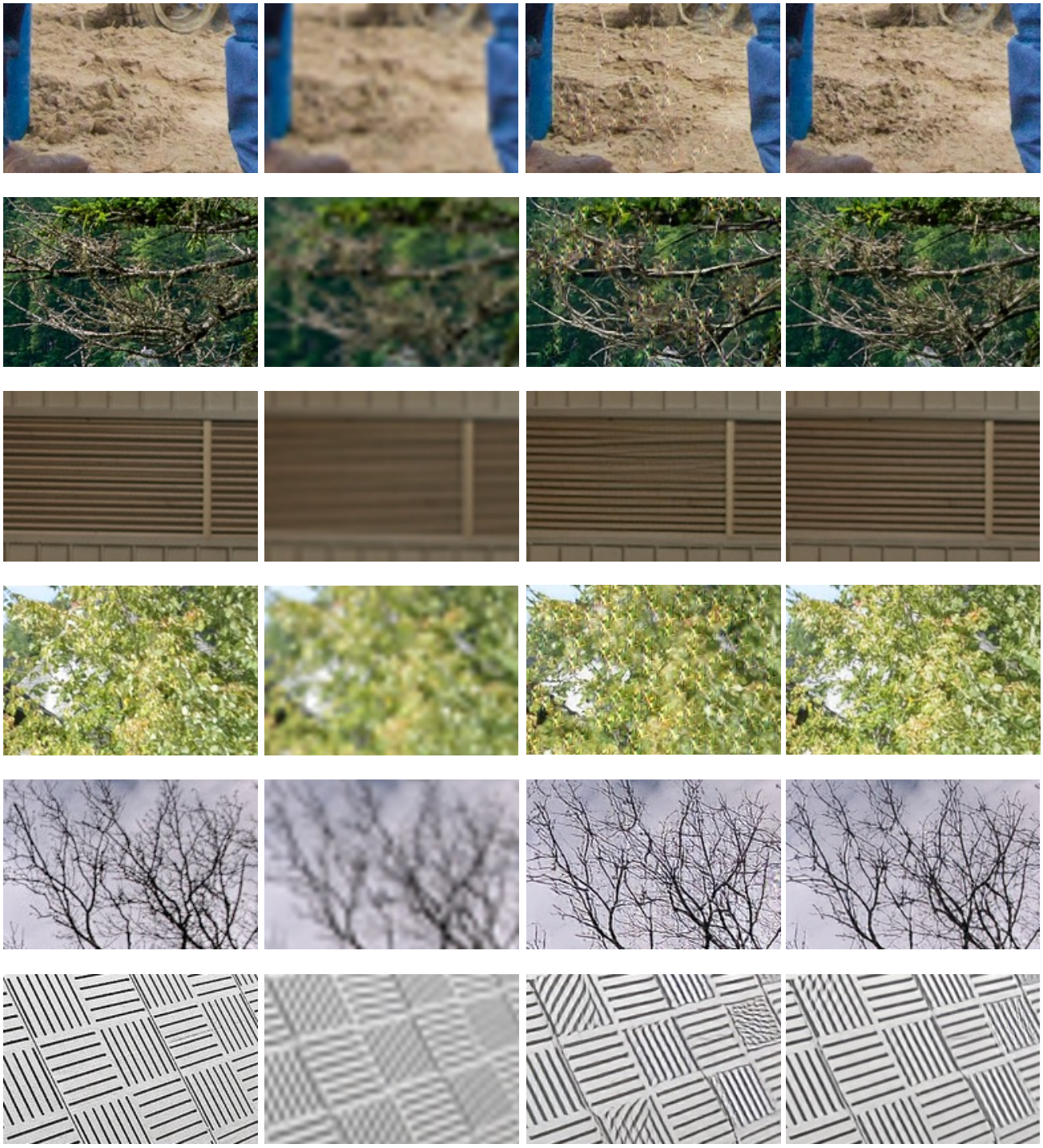
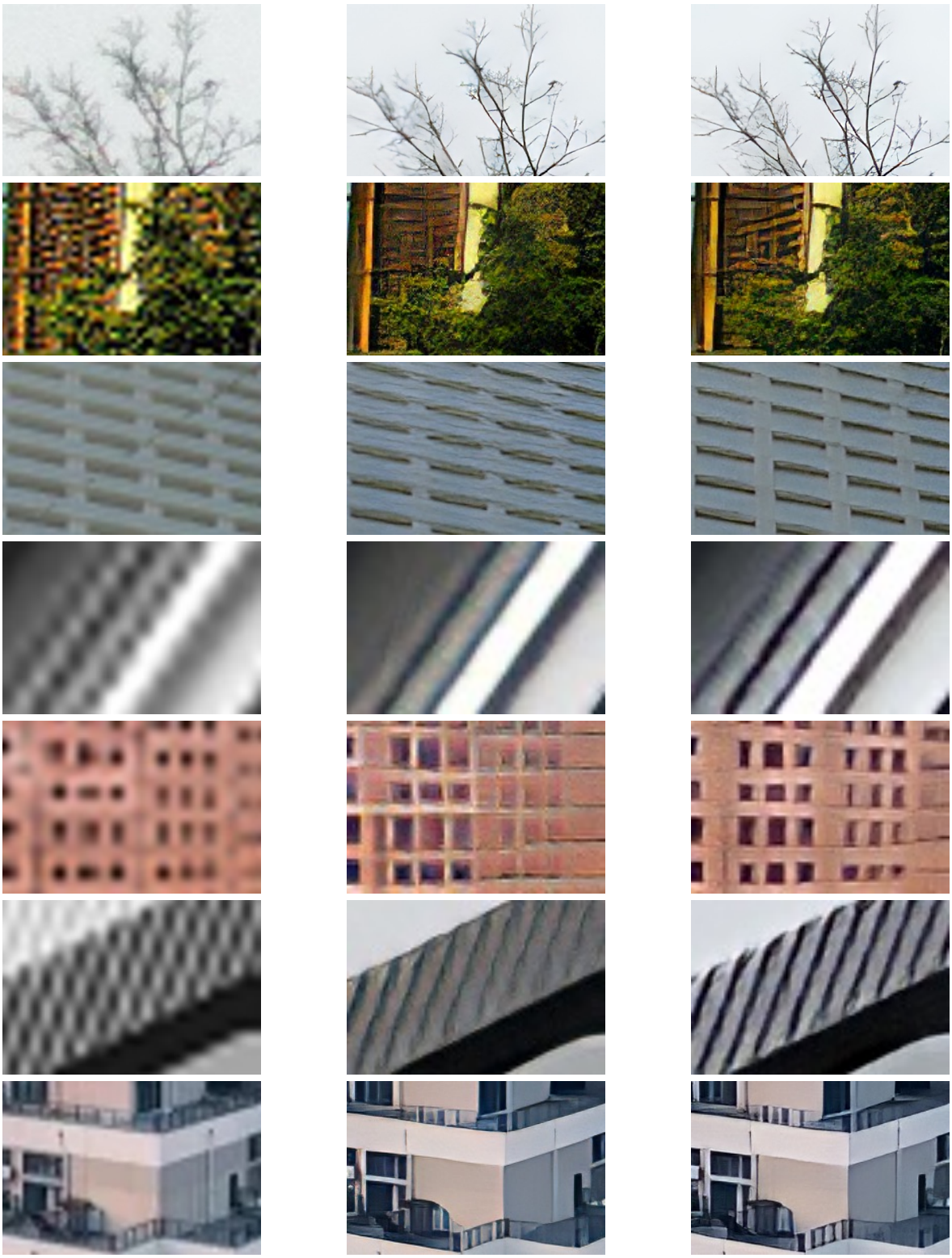


Figure 2. Visual comparison among GAN-SR methods that use RRDB as backbone, including ESRGAN [6], USRGAN [7], SPSR [3] and our RRDB+LDL, under the scaling factor of $4\times$ with bicubic degradation. In (e), we employ the officially released model of SPSR that is trained on DIV2K dataset. In (f), we train the SPSR on the DF2K dataset using the officially released code. In (g) and (h), we train our method on DF2K dataset and visualize the results of models Ψ and Ψ_{EMA}, respectively. As can be seen, our method can restore more realistic details than others, *e.g.*, the regular patters in the last example. Besides, our method can inhibit the visual artifacts such as the overshoot pixels in the first and second examples and the structural distortion in the last one. Both models Ψ and Ψ_{EMA} achieve clear improvement over the existing methods.



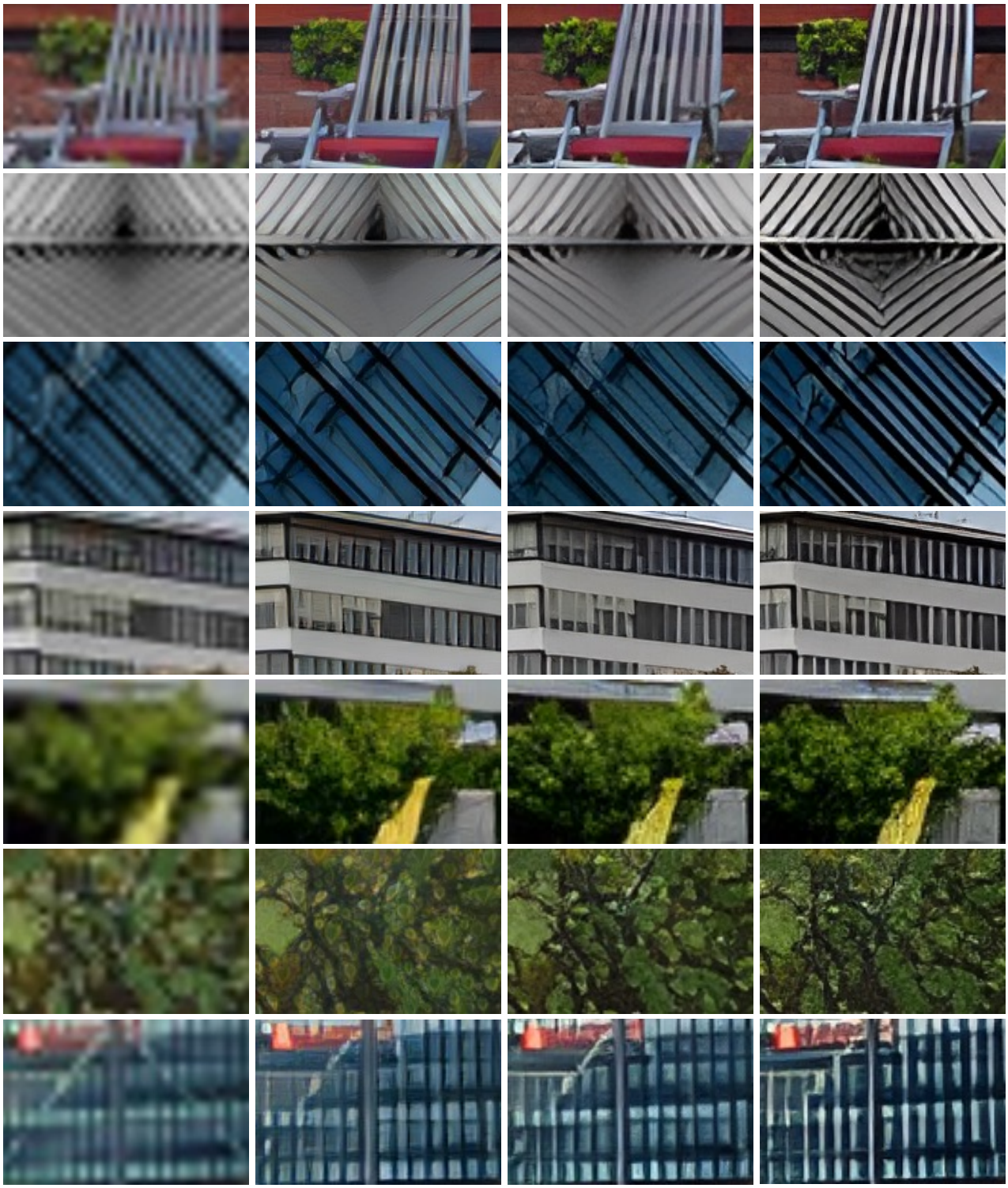
(a) HR (b) Bicubic (c) SwinIR+$\mathcal{L}_{\text{GAN}}$ (d) SwinIR+LDL

Figure 3. Visual comparison between SwinIR [2]+$\mathcal{L}_{\text{GAN}}$ and our SwinIR+LDL, under the scaling factor of $4\times$ with bicubic degradation. As can be seen, SwinIR+$\mathcal{L}_{\text{GAN}}$ introduces artifacts of similar patterns on texture regions like the sands in the 1^{st} row and the twigs and leaves in the 2^{nd} and 4^{th} rows. It may also include overshoot artifacts along with the sharp edges as shown in the second last row. In contrast, our SwinIR+LDL inhibits these artifacts, thanks to the explicit discrimination and penalty on these artifact pixels. Besides, our method demonstrates clear advantages in reconstructing realistic details, especially on regular patterns such as the examples in the 3^{rd} and the last row.



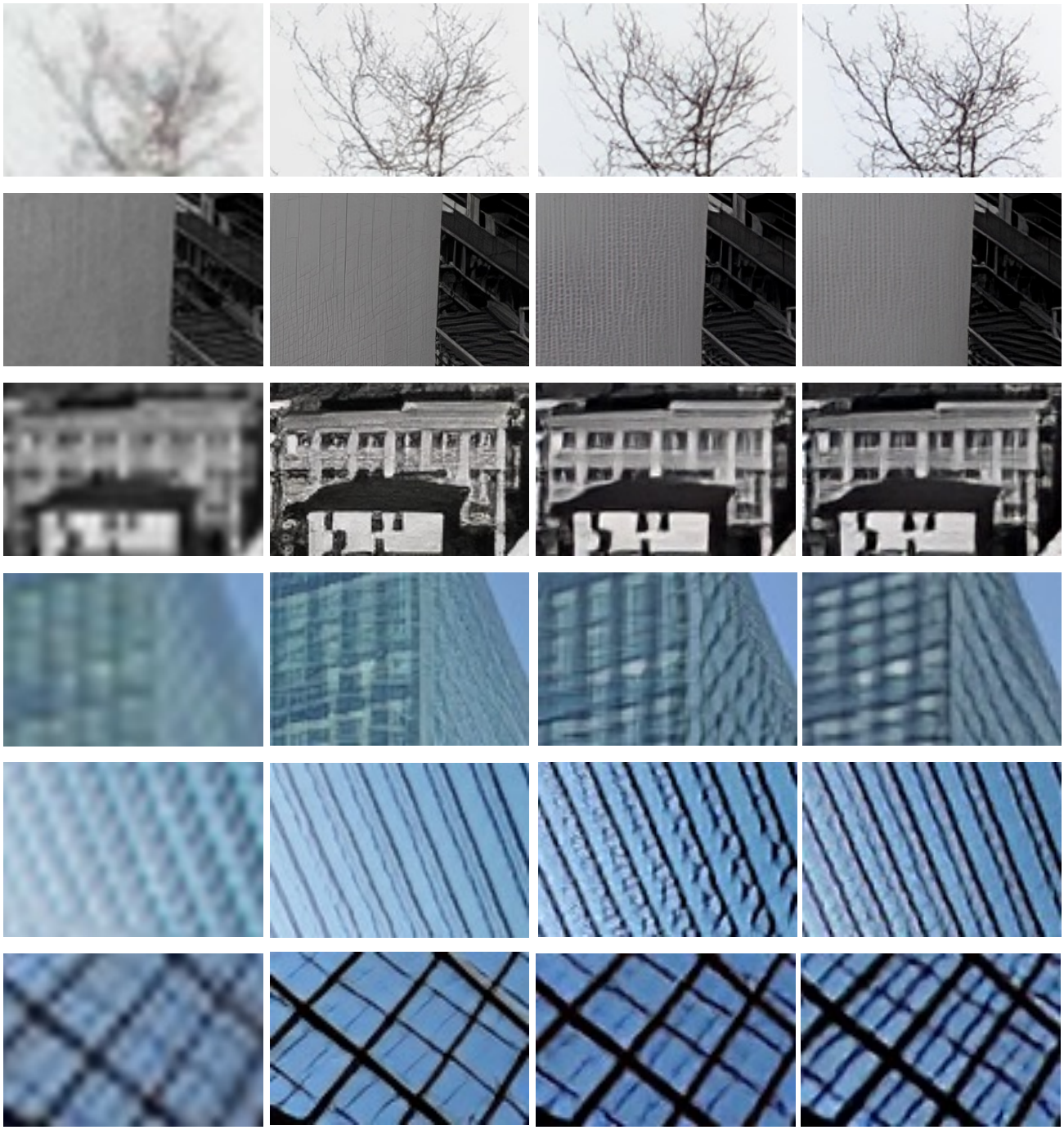
(a) Bicubic (b) SRResNet+RealESRGAN (c) SRResNet+RealESRGAN+LDL

Figure 4. Visual comparisons on $4\times$ real-world image super-resolution using SRResNet as backbone. In (b), we train a baseline by using SRResNet as backbone with the degradation model of RealESRGAN [4]. In (c), we apply our $\mathcal{L}_{\text{artif}}$ loss to the baseline in (b) while keeping other settings unchanged. As can be seen, the proposed LDL method reproduces richer details, such as twigs and the pattern of buildings, compared to the baseline one.



(a) Bicubic (b) BSRGAN (c) RealESRGAN (d) RealESRGAN+LDL

Figure 5. Visual comparison on $4\times$ real-world image super-resolution using RRDB as backbone. The results of BSRGAN [8] are shown in (b) and the results of RealESRGAN [4] are shown in (c). In (d), we apply our $\mathcal{L}_{\text{artif}}$ loss to the baseline (c) while keeping other settings unchanged. As can be seen, our LDL method reconstructs richer and sharper details, such as chairs, trees and the regular pattern of buildings compared to the BSRGAN and RealESRGAN.



(a) Bicubic (b) SwinIR+BSRGAN (c) SwinIR+RealESRGAN (d) SwinIR+RealESRGAN+LDL

Figure 6. Visual comparison on $4\times$ real-world image super-resolution using the SwinIR as backbone. In (c), we train a baseline by using SwinIR [2] as backbone with the degradation model of RealESRGAN [4]. In (d), we apply our $\mathcal{L}_{\text{artif}}$ loss to the baseline (c) while keeping other settings unchanged. As the authors of SwinIR released the model by combining the BSRGAN degradation model with the SwinIR backbone, we also show its results in (b). As can be seen in the first row, our LDL inhibits the artifacts robustly. In the third row, we can see that (b) introduces much artifacts on the surfaces of the building, while (c) is free of artifacts but it is blurry and non-realistic. In contrast, our LDL not only inhibits the artifacts but also reproduces sharp and realistic details.

Table 1. Quantitative comparison on $2\times$ SISR with bicubic degradation using three backbones. Since most of the existing GAN-SR methods do not train the $2\times$ SISR models, we train three baseline models (SRResNet, RRDB and SwinIR) using publicly available codes. We then apply the proposed $\mathcal{L}_{\text{artif}}$ loss to the respective baselines and train with the same setting. Three groups of comparisons are made, *i.e.*, SRResNet backbone for the first 2 columns, RRDB for the middle 2, and SwinIR for the last 2. The best results of each group are highlighted in **bold**. $\uparrow$ and $\downarrow$ mean that the larger or smaller score is better, respectively. All methods are trained on DF2K dataset. As can be seen, our proposed LDL scheme improves both the perceptual quality (LPIPS, DISTS, FID) and reconstruction accuracy (PSNR, SSIM) on most benchmarks for all the three backbones. Consistent observations with the experiments of $4\times$ SISR can be made, and this validates the effectiveness of the proposed LDL in improving the SISR performance across different scaling factors.

Metrics	Benchmarks	SRGAN	SRResNet+LDL	ESRGAN	RRDB+LDL	SwinIR+$\mathcal{L}_{\text{GAN}}$	SwinIR+LDL
Training Dataset		DF2K	DF2K	DF2K	DF2K	DF2K	DF2K
LPIPS $\downarrow$	Set5	**0.0168**	0.0181	**0.0155**	0.0157	**0.0124**	0.0139
	Set14	0.0386	**0.0379**	**0.0355**	0.0355	0.0290	**0.0287**
	Manga109	0.0115	**0.0103**	**0.0115**	0.0116	0.0082	**0.0080**
	General100	0.0192	**0.0186**	0.0166	**0.0166**	0.0142	**0.0140**
	Urban100	0.0371	**0.0347**	0.0285	**0.0284**	0.0234	**0.0224**
	DIV2K100	0.0311	**0.0306**	0.0261	**0.0260**	0.0236	**0.0227**
DISTS $\downarrow$	Set5	**0.0383**	0.0391	**0.0352**	0.0366	0.0314	**0.0298**
	Set14	0.0410	**0.0405**	**0.0387**	0.0399	0.0336	**0.0334**
	Manga109	0.0104	**0.0097**	0.0095	**0.0094**	0.0078	**0.0072**
	General100	0.0302	**0.0295**	0.0273	**0.0269**	0.0243	**0.0235**
	Urban100	0.0355	**0.0340**	**0.0295**	0.0297	0.0253	**0.0242**
	DIV2K100	0.0182	**0.0174**	**0.0142**	0.0143	0.0125	**0.0119**
FID $\downarrow$	Set5	8.917	**7.318**	7.143	**6.652**	7.164	**6.533**
	Set14	16.896	**16.359**	17.291	**15.261**	**10.725**	12.618
	Manga109	3.394	**3.088**	3.087	**2.975**	2.647	**2.432**
	General100	6.028	**5.844**	**5.174**	5.421	4.851	**4.711**
	Urban100	**18.032**	18.363	**17.417**	17.429	17.171	**17.166**
	DIV2K100	5.085	**5.084**	4.856	**4.775**	4.387	**4.209**
PSNR $\uparrow$	Set5	36.153	**36.544**	36.242	**36.302**	36.503	**36.749**
	Set14	32.089	**32.536**	32.173	**32.361**	32.791	**33.333**
	Manga109	36.461	**37.237**	37.442	**37.462**	37.932	**38.412**
	General100	36.156	**36.688**	36.703	**36.745**	37.031	**37.376**
	Urban100	30.261	**30.984**	31.069	**31.119**	32.098	**32.567**
	DIV2K100	33.948	**34.510**	34.237	**34.312**	34.682	**35.054**
SSIM $\uparrow$	Set5	0.9419	**0.9444**	0.9417	**0.9423**	0.9418	**0.9448**
	Set14	0.8853	**0.8913**	0.8895	**0.8910**	0.8925	**0.9006**
	Manga109	0.9622	**0.9663**	0.9677	**0.9679**	0.9681	**0.9708**
	General100	0.9405	**0.9444**	0.9442	**0.9450**	0.9460	**0.9495**
	Urban100	0.8993	**0.9073**	0.9094	**0.9104**	0.9198	**0.9247**
	DIV2K100	0.9169	**0.9218**	0.9181	**0.9191**	0.9233	**0.9274**

References

[1] 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. 1, 2

[2] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, 2021. 4, 7

[3] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *CVPR*, 2020. 1, 3

[4] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 1, 5, 6, 7

[5] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 2

[6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 1, 3

[7] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020. 3

[8] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1, 6