SeD: Semantic-Aware Discriminator for Image Super-Resolution Supplementary Materials

Section 1 includes two recently proposed large-scale high-quality benchmarks here for classical image SR evaluation: LSDIR [6] and HQ-50K [16].

Section 2 specifies the details of our real-world SeD implementations.

Section 3 provides the network structures of pixel-wise SeD (U+SeD), image-wise SeD (V+SeD) and CLIP semantic extractor. Additionally, we demonstrate the quantitative comparison between ESRGAN [14] and our implemented "RRDB+V+SeD".

Section 4 demonstrates the details of different fusion methods of SeD.

Section 5 presents the ablation studies of introducing semantic guidance into the generator.

Section 6 visualizes more results of classical image SR and real-world image SR.

1. Evaluation on large-scale benchmarks

With the development of image restoration, researchers have begun showing interest in larger-scale training and testing datasets, in addition to deeper model design. This interest is to align the field with other visual tasks, such as image recognition and image detection. To demonstrate the effectiveness of our SeD, we evaluate our methods on two recently proposed large-scale benchmarks, LSDIR [6] and HQ-50K [16]. There are 250 available images from LSDIR and 1250 images from HQ-50K. Different from commonly used benchmarks (*e.g.*, Set5 [1], Set14 [17], Urban100 [3], *etc.*), these testing images encompass a wide range of natural scenes, along with high resolution and complex textures. The results are shown in Table. 1. Notice that, we *do not* train our model on training datasets of these two benchmarks. Instead, we directly use weights introduced in the main paper.

As demonstrated in the table, SeD outperforms in both objective and subjective metrics, indicating that semantic guidance is capable of not only better reconstructing simple textures in commonly used benchmarks [1, 3, 9, 10, 17], but also effectively handling complex textures in large-scale evaluation datasets. Qualitative comparisons are given in Sec. 6.

Table 1. Evaluation results of $\times$ 4 classical SR on large-scale benchmarks. Metrics are LPIPS↓/PSNR↑/SSIM↑.

Datasets	LSDIR [6]	HQ-50K [16]
ESRGAN [14]	0.138/23.88/0.686	0.176/23.67/0.677
RRDB+LDL [8]	0.118/24.66/0.712	0.171/24.33/0.701
RRDB+SeD	**0.116/25.20/0.727**	**0.157/24.66/0.710**

2. Implementation details for real-world image SR

We perform experiments on real-world image SR. We compare the performance with the original discriminator (*i.e.*, without semantics) on three state-of-the-art methods: Real-ESRGAN [15], LDL [8] and SwinIR [7]. Following them, we evaluate the performance on several commonly-used real-world low-resolution datasets, including DPED [4], OST300 [13] and RealSRSet [19]. The training strategy and dataset of the three methods are slightly different. To keep fairness, we follow their original training settings to train the generator with our SeD, and compare the visual quality with their original GAN-based results. Furthermore, we adopt the well-known no-reference metric NIQE [11] for the quantitative comparison, since the ground-truth images are not available in the real world.

3. Implementation details of image-wise SeD

We incorporate our proposed semantic-aware discriminator to a VGG-like discriminator [12], dubbed V+SeD, which is shown in Fig. 1 (b).

In particular, the Image-wise discriminator has been explored in a series of GAN-based image SR networks [5, 14], since it is simple and effective.

The quantitative results are demonstrated in Table 2. Our V+SeD significantly outperforms the vanilla VGG-like discriminator (which is used by ESRGAN [14]) on both objective and subjective metrics. These further demonstrate the effectiveness and generalization capability of our suggested SeD with respect to various discriminator backbones.

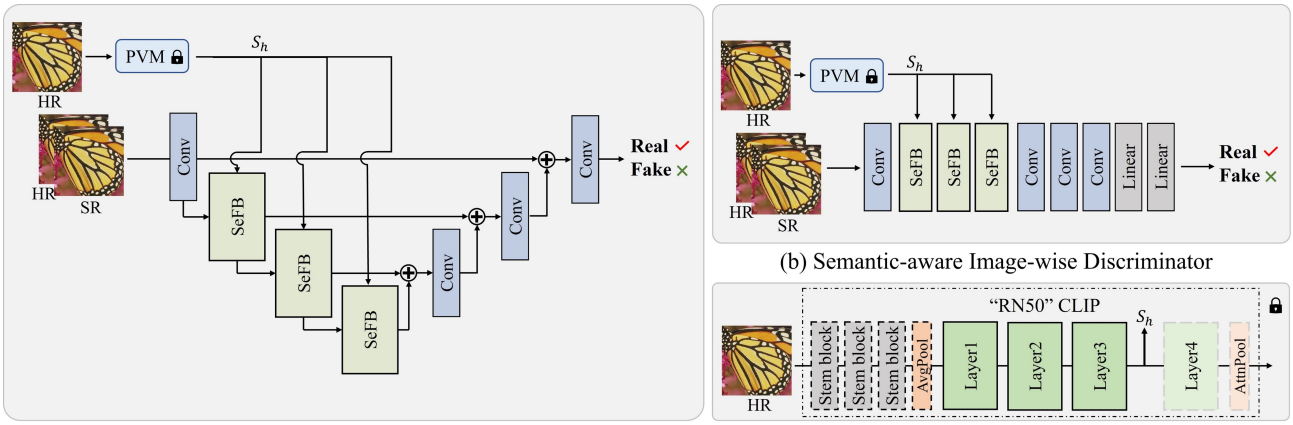


Figure 1. The framework of (a) pixel-wise U+SeD, (b) image-wise V+SeD, (c) "RN50" CLIP semantic extractor.

Datasets	ESRGAN	RRDB+V+SeD
Set5	0.076/30.44/0.852	**0.070**/30.83/0.862
Set14	0.133/26.28/0.699	**0.125**/27.06/0.729
DIV2K	0.115/28.20/0.777	**0.107**/28.83/0.794
Urban100	0.123/24.37/0.734	**0.118**/25.32/0.766
Manga109	0.065/28.41/0.859	**0.057**/29.31/0.878

Table 2. Quantitative comparison between ESRGAN and V+SeD. The best perceptual results of each group are highlighted in bold. Each result is presented in terms of LPIPS↓/PSNR↑/SSIM↑, ↑ and ↓ indicate that a larger or smaller score is better, respectively.

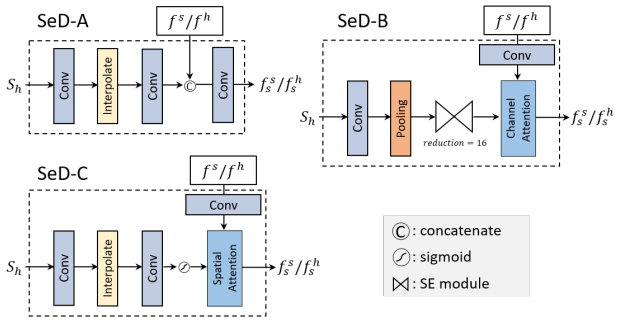


Figure 2. Frameworks of SeD-A, SeD-B, SeD-C, respectively.

4. Implementation details of different fusion strategies

We present the network architectures of our used SeD-A, SeD-B, and SeD-C in our ablation studies in Fig. 2. Among them, **SeD-A** uses concatenation operation as the fusion method. **SeD-B** utilizes a channel-wise attention mechanism to fuse semantic information adaptively. **SeD-C** leverages spatial-wise attention. In contrast, our proposed SeD performs cross-attention between semantic feature and image feature, taking full advantage of abundant semantic information contained in LVMs, and maintaining the spatial information.

5. Ablation studies with Semantic-aware Generator

As described in the main paper, one intuitive method to generate semantic-aware textures is to integrate the semantic guidance of images into the generator. To verify this idea, we conduct experiments of the semantic-aware generator on $\times$ 4 real-world image SR task. We choose RRDB [14] with 11 residual in residual blocks as the baseline generator. For semantic-aware RRDB, namely Se-RRDB, we replace the 5^{th} and the 11^{th} block with the same semantic-aware fusion block we used in the SeD. We synthesize the degraded images by Real-ESRGAN [15] model, and evaluate the perceptual qualities of restored images on RealSR [2] dataset in terms of NIQE [11]. We first train two separate generators without discriminators, then fine-tune them on vanilla discriminator and our SeD, respectively. The results are shown in Table. 3.

As we can see, Se-RRDB performs worse than RRDB across all training paradigms, which reveals that incorporating the semantic information in the generator may not be appropriate for real-world image SR problems. The reason we guess is that it is difficult to extract accurate semantic information from low-quality images since the severe distortions in real world will cause the failure of the semantic extractor. Moreover, introducing the semantics into the generator will cause the catastrophic growth of computation complexity in the inference stage, where semantic extraction is costly and time-consuming.

Therefore, in our paper, we explore the more simple and

#	RRDB	Se-RRDB	Vanilla D	SeD	Canon	Nikon
1	✓				7.56	7.83
2		✓			8.11	8.27
3	✓		✓		4.71	5.14
4		✓	✓		5.39	5.66
5	✓			✓	4.51	4.96
6		✓		✓	4.78	5.32

Table 3. Quantitative comparison between RRDB and Semantic-aware RRDB. ✓ means we use this backbone during training. The lower score is better.

effective semantic-aware discriminator (SeD), which improves the perceptual qualities of restored images, *i.e.*, the comparison between 3^{rd} and the 4^{th} lines or the comparison between the 5^{th} and the 6^{th} lines in Table. 3. Moreover, our SeD enables the SR network to restore more photo-realistic textures and does not require any additional computational burden during the inference stage.

6. More visualization results

First, we show the visualization of semantics obtained from PVM in the image as Fig. 3 (brighter region means more semantically important for PVM.) As stated, PVM can bring more fine-grained semantics in one image for guidance. (*e.g.*, covering most semantics from trees, person, *etc.* in the first sample).

Figure 3. Visualization of semantics extracted from PVM.

Then, we show more qualitative comparisons with previous GAN-based SR methods in classical and real-world image SR. As shown in Fig. 4 and Fig. 5, our SeD enables SR networks to correctly restore repeating textures, where previous works typically fail to deal with. Moreover, with semantic guidance, our SeD is capable of recovering more fine-grained textures (*e.g.*, twigs, furs and windows) in natural sceneries, in the meanwhile reducing the artifacts of super-resolved images. These conclusions remain valid for large-scale benchmarks, as illustrated in Fig. 6, Fig. 7 and Fig. 8, which further demonstrates the effectiveness of our SeD.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 1

[2] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 2

[3] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 1

[4] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3277–3285, 2017. 1

[5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[6] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 1, 6

[7] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[8] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 1, 6, 7, 8

[9] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 1

[10] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 1

[11] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1, 2

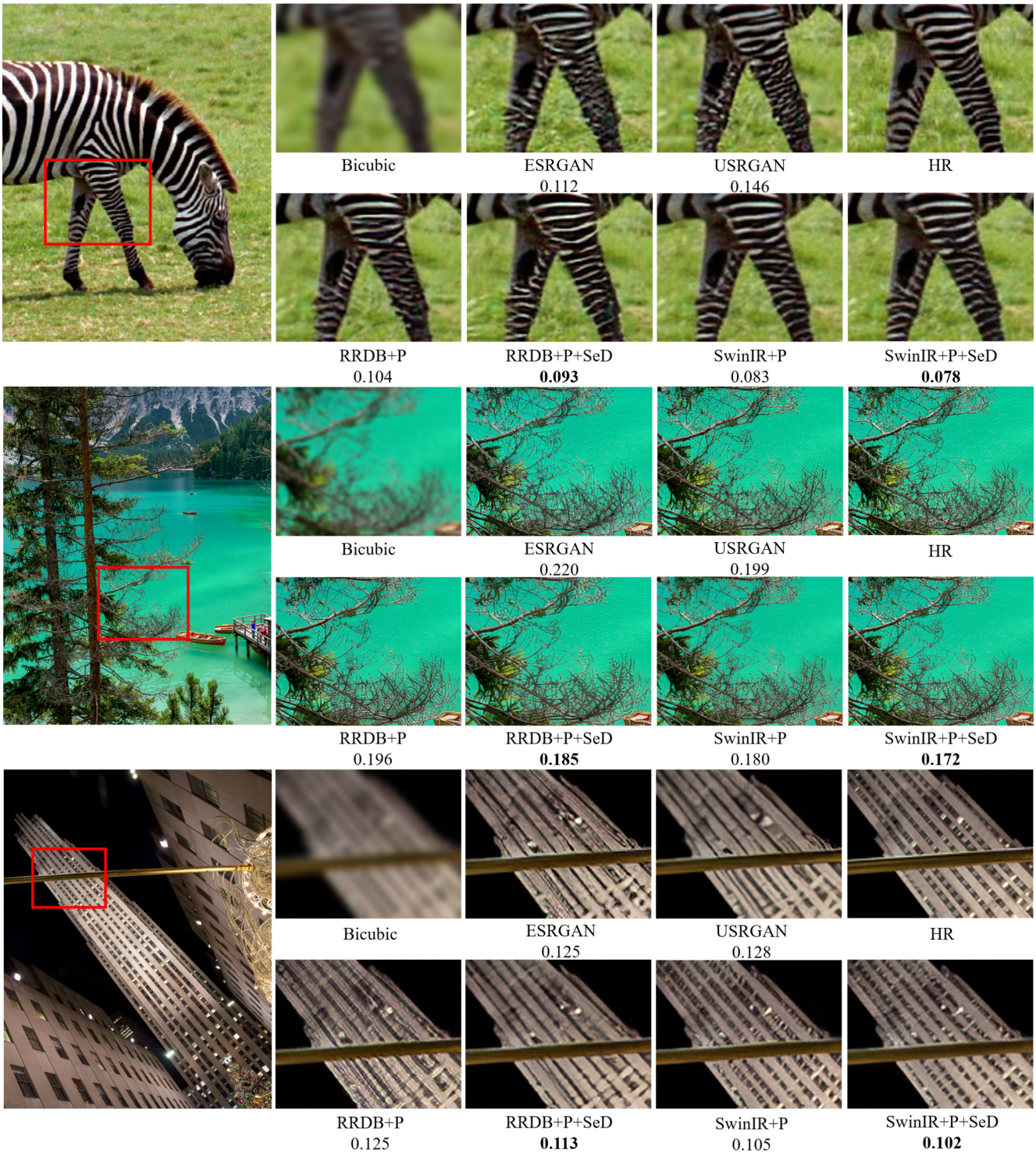


Figure 4. Visual comparisons between SeD and vanilla discriminator on classical image SR. To provide a clearer understanding of the perceptual qualities of images, we present the LPIPS↓ here. It is evident that our SeD further enhances the ability of vanilla GAN to restore more realistic textures.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*

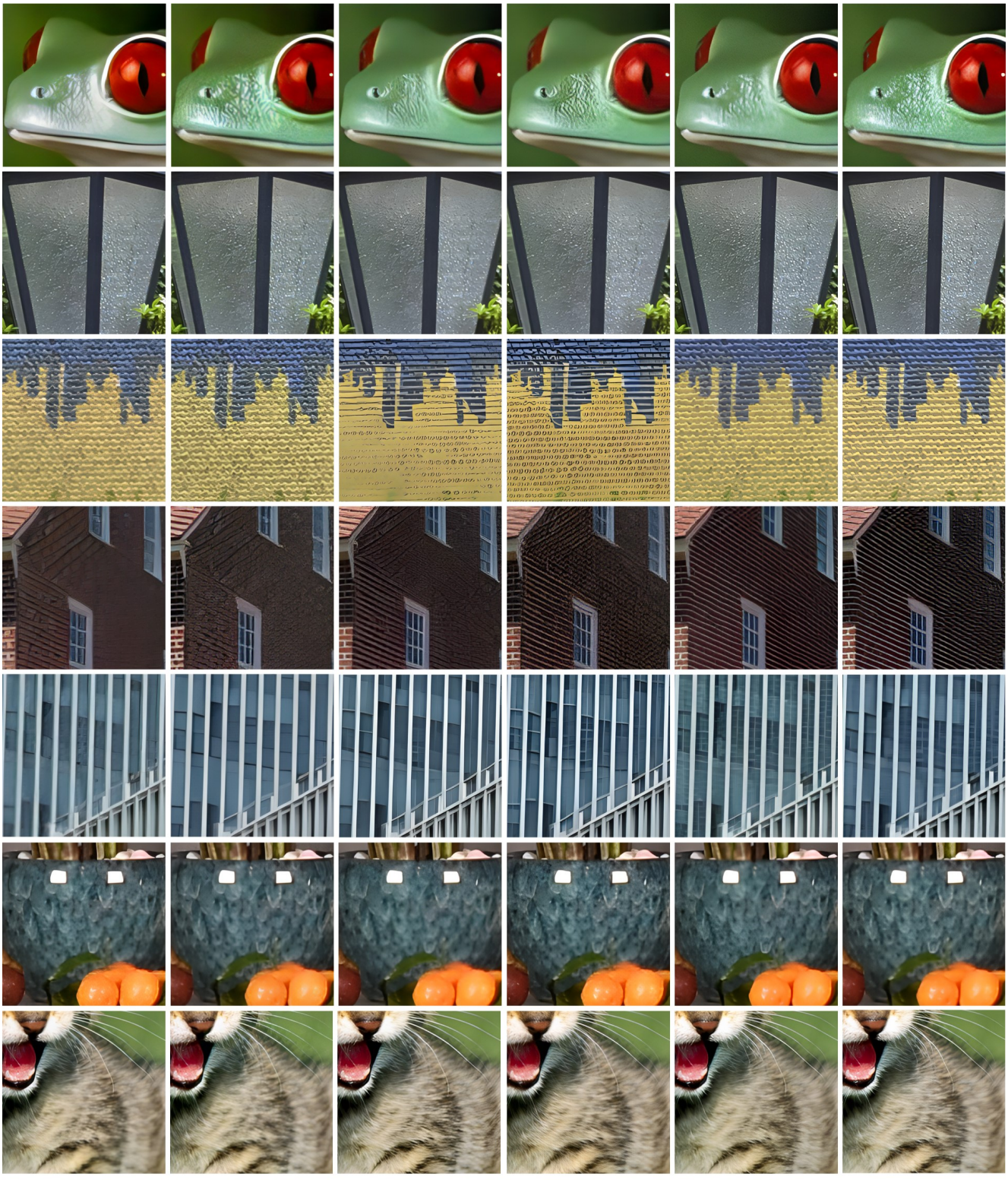


Figure 5. More visual comparisons between SeD and prominent GAN-based methods on real-world image SR of natural sceneries. Zoom in for better view.

Figure 6. Visual Comparisons between SeD and other methods on LSDIR [6] (img_0000127 with resolution 780 × 780).

preprint arXiv:1409.1556, 2014. 1

[13] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 1

[14] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1, 2, 6, 7, 8

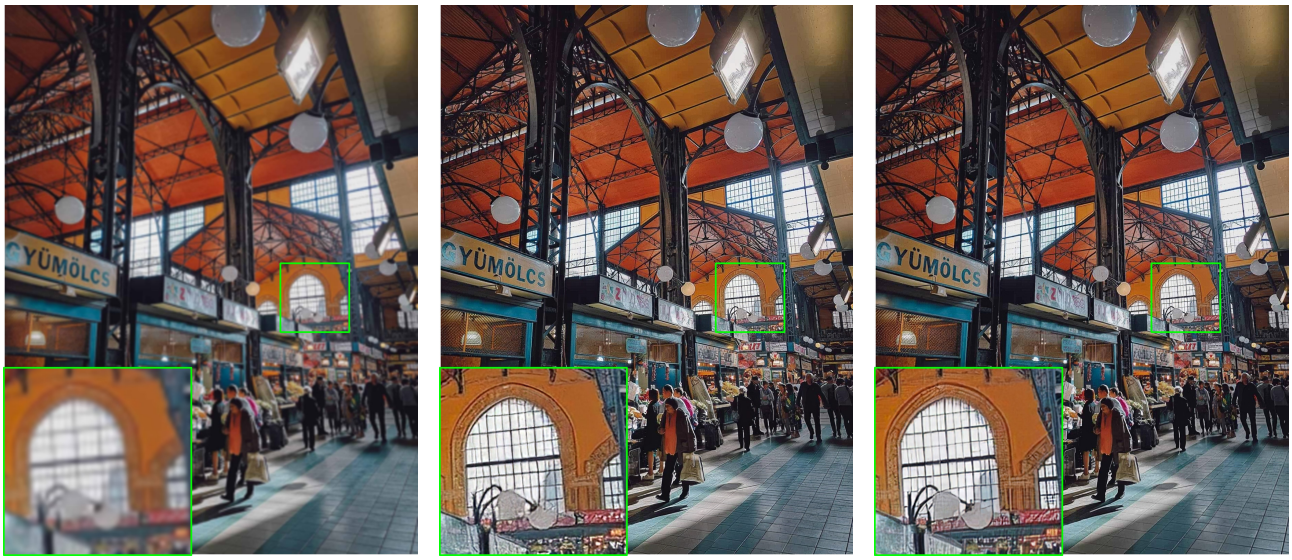
[15] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 2

[16] Qinhong Yang, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Lu Yuan, Gang Hua, and Nenghai Yu. Hq-50k: A large-scale, high-quality dataset for image restoration. *arXiv preprint arXiv:2306.05390*, 2023. 1, 7, 8

[17] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 1

[18] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 6, 7, 8

[19] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1



Bicubic ESRGAN [14] USRGAN [18]

RRDB+LDL [8] RRDB+P+SeD HR

Figure 7. Visual Comparisons between SeD and other methods on HQ-50K [16] (complex/00020 with resolution 1200 $\times$ 1596).

Bicubic | ESRGAN [14]

USRGAN [18] | RRDB+LDL [8]

RRDB+P+SeD | HR

Figure 8. Visual Comparisons between SeD and other methods on HQ-50K [16] (people/00010 with resolution 2040 $\times$ 1536).