CoSeR: Bridging Image and Language for Cognitive Super-Resolution

Supplementary Material

Sec. A provides an extensive elucidation of our method, including details of the cognitive encoder supervision method, the one-hot reference attention mechanism, and an in-depth analysis of the network architectures governing the denoising U-Net and ControlNet. In Sec. B, we present comprehensive quantitative comparisons between the proposed method and established models, including the recently introduced DiffBIR [11]. Additionally, our investigation delves into the impact of introducing multiple generated reference images. We provide the results of a user study in the form of voting results and assessments of image quality at the pixel level. Sec C showcases more visualization examples, extensively demonstrating the effectiveness of our method. Finally, we talk about the future work in D.

A. Detailed Illustration of our Method

A.1. Cognitive Encoder Supervision

As aforementioned in the main paper, we use T_e ($T_e \leq T_l$) tokens, preceding the class token $\boldsymbol{L}\left[t_{cls}\right]$ (inclusive), extracted from the CLIP language embedding $\boldsymbol{L} \in \mathbb{R}^{B \times T_l \times C_l}$ for supervision. B, T_l, and C_l denote batch size, token number, and channel number, respectively. If there are insufficient supervision tokens, we use the class token for end-filling. The loss function for training the cognitive encoder is expressed as:

$$\mathcal{L}_{CE} = \left\|\boldsymbol{E} - \boldsymbol{L}'\right\|_2^2, \tag{1}$$

where

$$\boldsymbol{L}' = \begin{cases} \text{Padding}\left(\boldsymbol{L}, \boldsymbol{L}\left[t_{cls}\right]\right), & \text{if } t_{cls} < T_e; \\ \boldsymbol{L}\left[\left(t_{cls} - T_e\right) : t_{cls}\right], & \text{if } t_{cls} \geqslant T_e. \end{cases} \tag{2}$$

We observe that employing $\boldsymbol{L}$ directly as supervision for the cognitive embedding $\boldsymbol{E}$ (setting $T_e = T_l$) hinders the acquisition of cognitive information, as depicted in Figure A.1. In this scenario, the generated reference images might prove irrelevant to low-resolution (LR) images. This limitation stems from the variability in caption length, which leads to Q-Former's learnable queries inadequately capturing semantic information at corresponding positions. To mitigate this issue, we propose using the last T_e tokens for supervision for two reasons. Firstly, the last T_e tokens in the CLIP text embedding inherently encapsulate an overarching representation of all preceding words [14], facilitated by the causal attention mechanism. This mitigates the requirement for a strict one-to-one correspondence between query ordering and semantic representation, thus enabling more effective learning by the queries. Secondly, within the supervision target $\boldsymbol{L}'$, the last query consistently aligns with the class token, thereby preserving the full representational capacity of the class token. Compared to the direct utilization of $\boldsymbol{L}\left[t_{cls}\right]$ or single class token, our approach of employing the last T_e tokens for supervision presents a more accurate understanding of LR images, which is supported by both Figure. 5 in the main paper and Figure A.1.

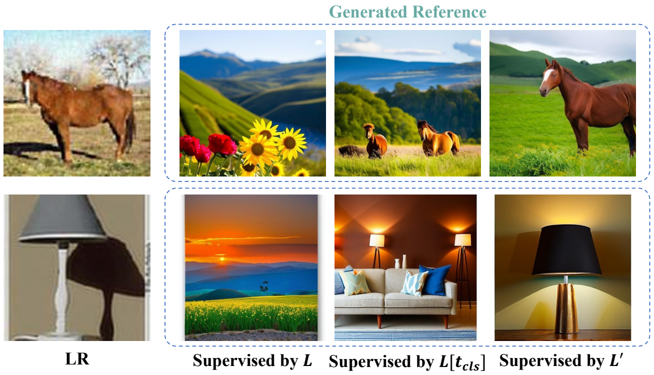


Figure A.1. Reference images generated by cognitive encoders with different supervision methods.

Number of Queries	Gen-score↑
$T_e = 30$	0.5983
$T_e = 40$	0.6048
$T_e = 50$	**0.6147**
$T_e = 60$	0.6110
$T_e = 77$	0.6082

Table A.1. Reference image quality assessment using different numbers of learnable queries.

We investigate the impact of the number of learnable queries, denoted as T_e, in our cognitive encoder on the generation of high-quality reference images. This analysis involve the examination of 200 randomly selected low-resolution test images by varying the query number from 30 to 77. It is noted that the setting of $T_e = 77$ in $\boldsymbol{L}'$ differs from using $\boldsymbol{L}$ for supervision. This distinction arises from the fact that the final tokens of $\boldsymbol{L}'$ are expanded with class tokens when the caption is not sufficiently lengthy. The results presented in Table A.1 demonstrate that our cognitive encoder achieves optimal performance when $T_e = 50$ (where "Gen-score" is defined in the main paper). Hence, we establish $T_e = 50$ as the default value. Notably, the quality of the generated images begins to decline for $T_e > 50$, as also evidenced in Figure A.2. This decline might be attributed to increased learning complexity associated with a higher number of tokens.

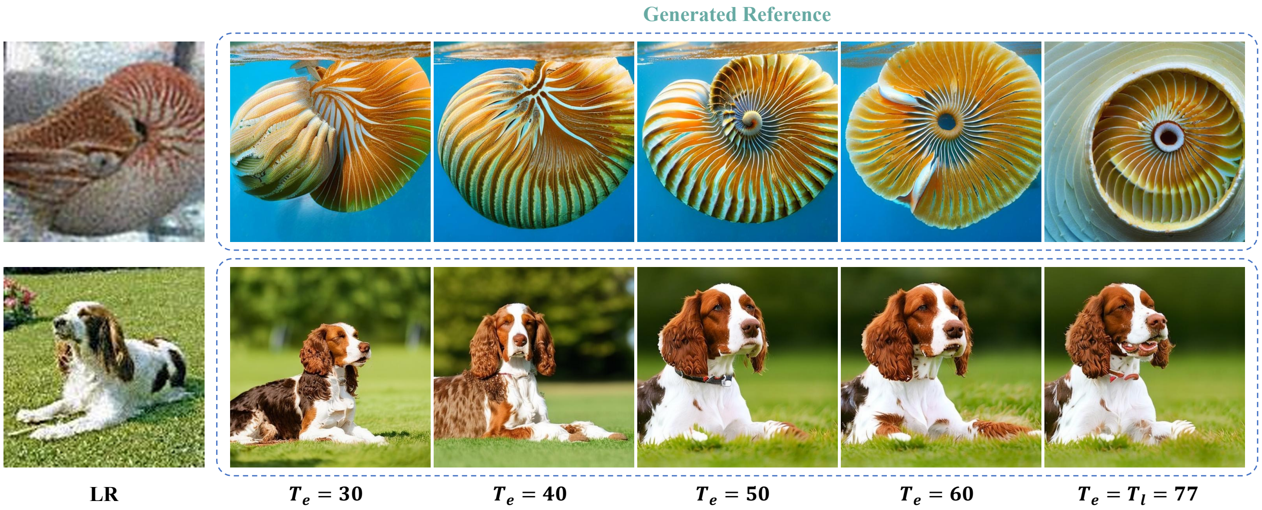


Figure A.2. Reference images generated by cognitive encoders with different numbers of learnable queries.

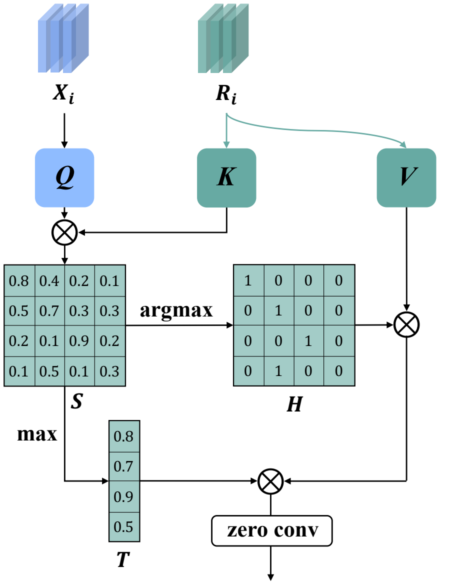


Figure A.3. The architecture of one-hot reference attention in the All-in-Attention (AiA) module.

A.2. One-Hot Reference Attention

The reference image contains high-definition textures that maintain consistent semantics with the corresponding LR image. However, not all features from the reference image are useful for LR recovery. The conventional attention mechanism calculates the weighted sum of all queries in value features, potentially leading to a blurring effect [19]. To address this issue, we introduce one-hot attention in the reference module to enhance the LR image with the most pertinent reference feature.

The one-hot attention mechanism is depicted in Figure A.3, where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the query, key, and value features, respectively. We represent the LR control and reference image control at the i-th scale as $\boldsymbol{X}_i$ and $\boldsymbol{R}_i$. $\boldsymbol{Q} \in \mathbb{R}^{B\times T_x \times C}$ and $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{B\times T_r \times C}$ are derived from $\boldsymbol{X}_i, \boldsymbol{R}_i$. The similarity $\boldsymbol{S} \in \mathbb{R}^{B\times T_x \times T_r}$ between $\boldsymbol{Q}$ and $\boldsymbol{K}$ is computed with normalized inner product:

$$\boldsymbol{S} = \langle \boldsymbol{Q}, \boldsymbol{K} \rangle \,. \tag{3}$$

We derive the one-hot map $\boldsymbol{H} \in \mathbb{R}^{B\times T_x \times T_r}$ along the T_r dimension of $\boldsymbol{S}$ and record the maximum values as $\boldsymbol{T} \in \mathbb{R}^{B\times T_x}$. The final output of the one-hot attention is then expressed as:

$$\boldsymbol{Z}_{\text{out}} = \text{ZeroConv}\left[(\boldsymbol{H}\boldsymbol{V}) \odot \boldsymbol{T}\right], \tag{4}$$

where $\odot$ denotes element-wise multiplication. It is noteworthy that we opt not to use *softmax* and, instead, employ the correlation matrix $\boldsymbol{T}$ to diminish less similar features while amplifying those that are potentially valuable. Additionally, to prevent the newly introduced attention components from influencing the well-established representation of Stable Diffusion [15] during early training, we integrate zero convolutions [22] at the end.

A.3. Network Structure

Denoising U-Net. The denoising U-Net in the proposed Cognitive Super-Resolution (CoSeR) network is depicted in Figure A.4. In our architecture, we adopt the All-in-Attention (AiA) module, replacing all original attention modules present in both the middle and decoder components of the Stable Diffusion denoising U-Net. It is crucial to highlight that cognitive embedding is utilized across all attention modules in the denoising U-Net, extending beyond solely the AiA modules.

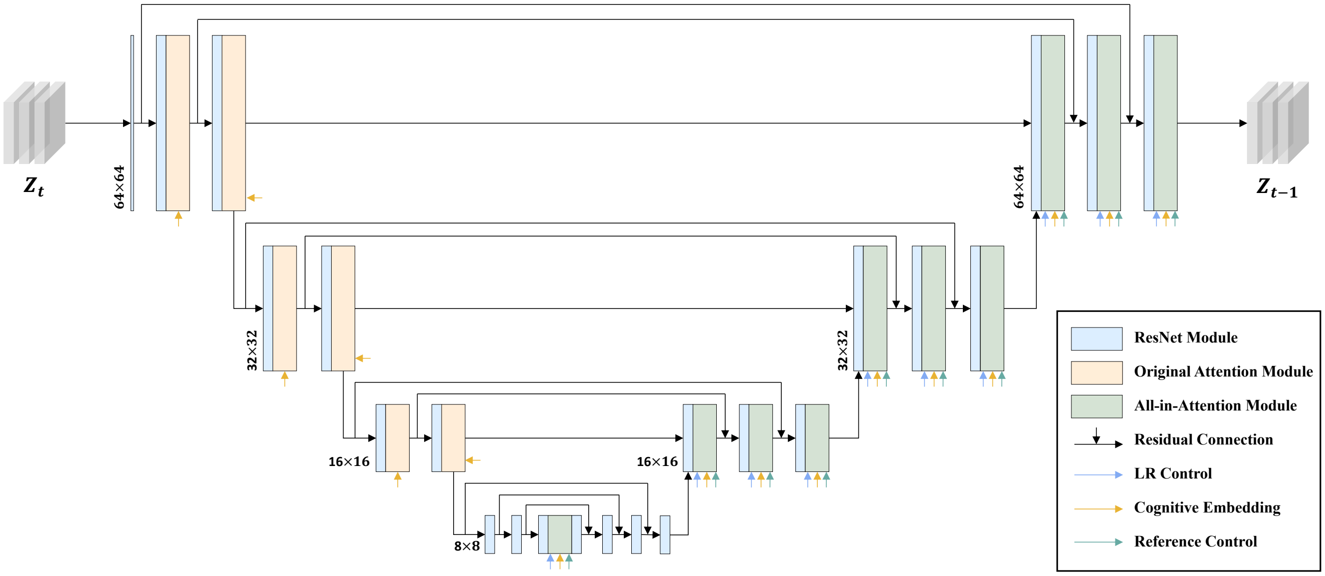
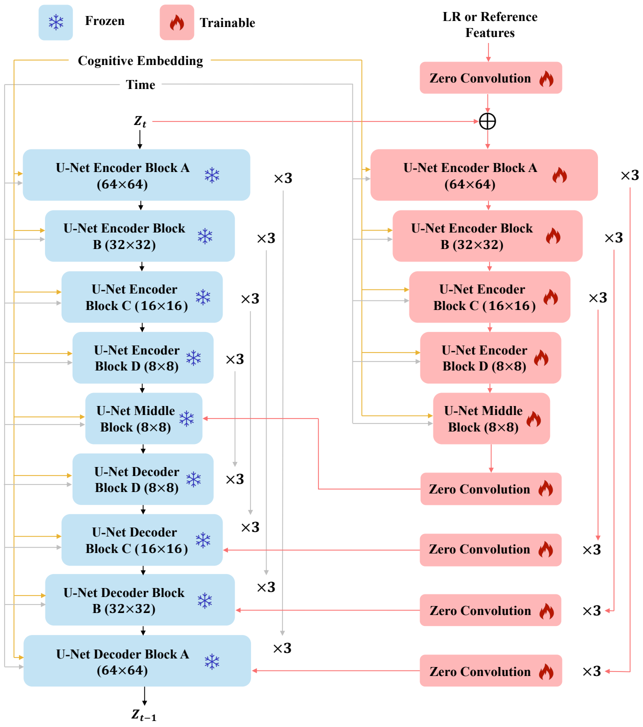


Figure A.4. Network structure of the denoising U-Net in the proposed CoSeR framework.


Frozen
Trainable
LR or Reference Features
Cognitive Embedding
Time
Z_t
Zero Convolution
U-Net Encoder Block A (64×64)
U-Net Encoder Block B (32×32)
U-Net Encoder Block C (16×16)
U-Net Encoder Block D (8×8)
U-Net Middle Block (8×8)
U-Net Decoder Block D (8×8)
U-Net Decoder Block C (16×16)
U-Net Decoder Block B (32×32)
U-Net Decoder Block A (64×64)
×3
Z_{t-1}

(a) Denoising U-Net

(b) ControlNet

Figure A.5. Network structure of ControlNet in the proposed CoSeR framework.

ControlNet. We utilize ControlNet [22] to generate multi-scale control features for both LR and reference images. As illustrated in Figure A.5, we mirror the weights and structure of the denoising U-Net in the ControlNet. Following [22], zero convolutions are incorporated at the beginning and end of the ControlNet module. Subsequently, the resulting control features are directed to the All-in-Attention (AiA) modules situated within the middle and decoder components of the denoising U-Net, excluding U-Net Decoder Block D, which lacks attention modules. Importantly, cognitive embedding is also employed in the ControlNet module.

B. Additional Experiments

B.1. Quantitative Comparisons to Official Models

For fair comparisons, we conduct a re-training of real-world super-resolution (SR) models using the ImageNet [4] dataset in the main paper. Remarkably, our CoSeR model achieves the highest performance. To provide a comprehensive analysis, we compare CoSeR against the officially released models: RealSR [8], Real-ESRGAN+ [17], SwinIR-GAN [10], BSRGAN [21], FeMaSR [3], DiffBIR [11], and StableSR [16]. It is noted that we exclude the comparison with DiffBIR on the ImageNet Test2000 dataset due to potential data overlap with its official training set. Additionally, all diffusion-based models, including LDM, DiffBIR, StableSR, and CoSeR, employ 200 sampling steps. As outlined in Table B.2, across various evaluation metrics such as FID [6], DISTS [5], LPIPS [23], CLIP-Score [14], and MUSIQ [9], our method consistently excels, positioning CoSeR as the superior and more robust approach.

B.2. Comparisons to Re-trained DiffBIR

As an extension to the quantitative comparisons provided above, we further conduct a comparative analysis with DiffBIR [11], specifically re-trained using our ImageNet train-

Datasets	Metrics	RealSR	Real-ESRGAN+	SwinIR-GAN	BSRGAN	FeMaSR	DiffBIR	StableSR	CoSeR
ImageNet Test2000	FID↓	86.36	39.37	44.86	49.94	45.19	–	24.70	**19.41**
	DISTS↓	0.2649	0.1915	0.2000	0.2043	0.1995	–	0.1608	**0.1482**
	LPIPS↓	0.4519	0.3122	0.3327	0.3401	0.3403	–	0.2979	**0.2863**
	CLIP-Score↑	0.6242	0.7642	0.7325	0.7126	0.7272	–	0.8459	**0.8755**
	MUSIQ↑	50.18	61.92	57.60	64.37	60.27	–	63.20	**65.51**
RealSR [2]	FID↓	157.85	106.24	105.99	111.25	106.08	90.30	96.39	**80.82**
	DISTS↓	0.2529	0.2021	0.1969	0.2081	0.2125	0.1932	0.1899	**0.1826**
	LPIPS↓	0.3672	0.2805	0.2755	0.2801	0.2688	0.2967	0.2639	**0.2438**
	CLIP-Score↑	0.7458	0.8425	0.8425	0.8304	0.8473	0.8414	0.8531	**0.8545**
	MUSIQ↑	60.40	66.68	65.93	68.35	67.51	69.20	69.25	**70.29**
DRealSR [18]	FID↓	148.58	97.60	98.94	110.53	95.71	86.49	83.36	**71.22**
	DISTS↓	0.2673	0.2121	0.2056	0.2033	0.2016	**0.1959**	0.2034	0.1977
	LPIPS↓	0.4212	0.2973	0.2946	0.3062	0.2777	0.3075	0.2960	**0.2702**
	CLIP-Score↑	0.7360	0.8623	0.8571	0.8498	0.8680	0.8630	0.8729	**0.8766**
	MUSIQ↑	54.28	66.30	66.74	67.64	67.60	68.64	69.57	**70.18**

Table B.2. Quantitative comparisons to officially released models on both ImageNet Test2000 and real-world benchmarks RealSR and DRealSR. The best results are highlighted in **bold** and the second best results are in underlined.

Datasets	Methods	DISTS↓	LPIPS↓	CLIP-Score↑	PSNR↑	SSIM↑	FID↓	MUSIQ↑
ImageNet Test2000	DiffBIR	0.1523	0.3156	0.8683	21.12	0.5366	21.30	**67.40**
	CoSeR	**0.1482**	**0.2863**	**0.8755**	**22.28**	**0.5998**	**19.41**	65.51
RealSR	DiffBIR	0.1907	0.2727	0.8379	20.49	0.5511	**78.31**	68.63
	CoSeR	**0.1826**	**0.2438**	**0.8545**	**21.24**	**0.6109**	80.82	**70.29**
DRealSR	DiffBIR	0.2008	0.2980	0.8581	19.85	0.4934	**68.21**	68.60
	CoSeR	**0.1977**	**0.2702**	**0.8766**	**19.95**	**0.5350**	71.22	**70.18**

Table B.3. Quantitative comparisons between CoSeR and re-trained DiffBIR. The better results are highlighted in **bold**.

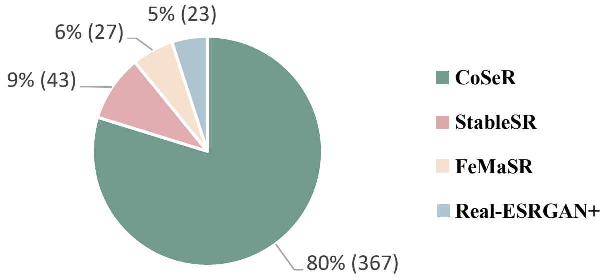


Figure B.6. The voting results obtained from 23 users. The percentage of votes chosen along with the corresponding numerical count are adjacent to the pie chart.

ing set. The results displayed in Table B.3 underscore the superiority of CoSeR across three benchmarks, demonstrating better performance across nearly all metrics.

B.3. Quantitative Comparisons on Widely Used Benchmark Datasets

To further validate the generalization ability of our method, we validated it on some common benchmark datasets: Set5 [1], Set14 [20], BSD100 [12], Urban100 [7], and Manga109 [13]. Unlike traditional methods that utilize official LR images created by bicubic downsampling, we generated LR-HR pairs via the Real-ESRGAN pipeline to better replicate real-world scenarios. As illustrated in Table B.4, our method outperforms others in all benchmarks, with the exception of Set5. The deviation in Set5 results can be attributed to its limited number of test images, introducing variability.

B.4. Voting Results of User Study

As detailed in the main paper, we invite 23 subjects to discern the visually superior result among the four SR candidates generated by Real-ESRGAN+, FeMaSR, StableSR, and CoSeR. This user study encompasses 20 real-world low-resolution images sourced from the Internet or captured via mobile phones, resulting in a total of 20×23 votes gathered. The depicted voting results in Figure B.6 unequivocally illustrate the superior performance of our CoSeR.

B.5. Number of Reference Images

We investigate the influence of using multiple generated reference images on the quality of SR results. Employing the

Datasets	Metrics	BSRGAN	DASR	StableSR	CoSeR
Set5	DISTS↓	0.2514	**0.2204**	0.2492	0.2356
	FID↓	157.42	**131.22**	188.52	139.53
	CLIP-Score↑	0.6885	0.7521	0.7338	**0.7962**
Set14	DISTS↓	0.2252	0.2319	0.2117	**0.2087**
	FID↓	161.55	168.52	151.77	**134.74**
	CLIP-Score↑	0.7355	0.7527	0.8029	**0.8106**
BSD100	DISTS↓	0.2601	0.2530	0.2363	**0.2200**
	FID↓	169.06	157.22	133.49	**114.78**
	CLIP-Score↑	0.6427	0.6752	0.7119	**0.7429**
Urban100	DISTS↓	0.2460	0.2492	0.1775	**0.1681**
	FID↓	81.85	82.24	50.95	**47.75**
	CLIP-Score↑	0.8360	0.8362	0.9152	**0.9265**
Manga109	DISTS↓	0.1516	0.1390	0.1124	**0.1101**
	FID↓	59.27	51.75	41.64	**40.96**
	CLIP-Score↑	0.8900	0.9128	**0.9311**	0.9283

Table B.4. Quantitative comparisons on widely used benchmark datasets.

Number of Ref. Images	FID↓	MUSIQ↑	MANIQA↑
1	19.80	64.21	0.2107
2	**19.54**	64.85	0.2169
3	19.58	**64.92**	**0.2170**

Table B.5. Results of using multiple reference images.

Datasets	Metrics	Real-ESRGAN+	FeMaSR	StableSR	CoSeR
ImageNet Test2000	PSNR↑	**22.64**	20.95	22.24	22.28
	SSIM↑	**0.6268**	0.5674	0.6093	0.5998
RealSR	PSNR↑	**21.93**	20.45	21.19	21.24
	SSIM↑	**0.6497**	0.6061	0.6247	0.6109
DRealSR	PSNR↑	**20.47**	18.46	19.86	19.95
	SSIM↑	**0.5781**	0.5036	0.5487	0.5350

Table B.6. Pixel-level PSNR and SSIM assessment of SR quality.

same LR input, we randomly sample noise maps to create several reference images utilizing identical cognitive embeddings. The findings presented in Table B.5 demonstrate that introducing a greater number of reference images yields improved performance. However, it's noteworthy that the improvement plateaus when using 2 or 3 reference images, suggesting that these images already contain sufficient high-definition textures to guide the process. As a result, we recommend utilizing 2 reference images as an optimal balance between quality enhancement and computational efficiency.

B.6. Pixel-level Image Quality Assessment

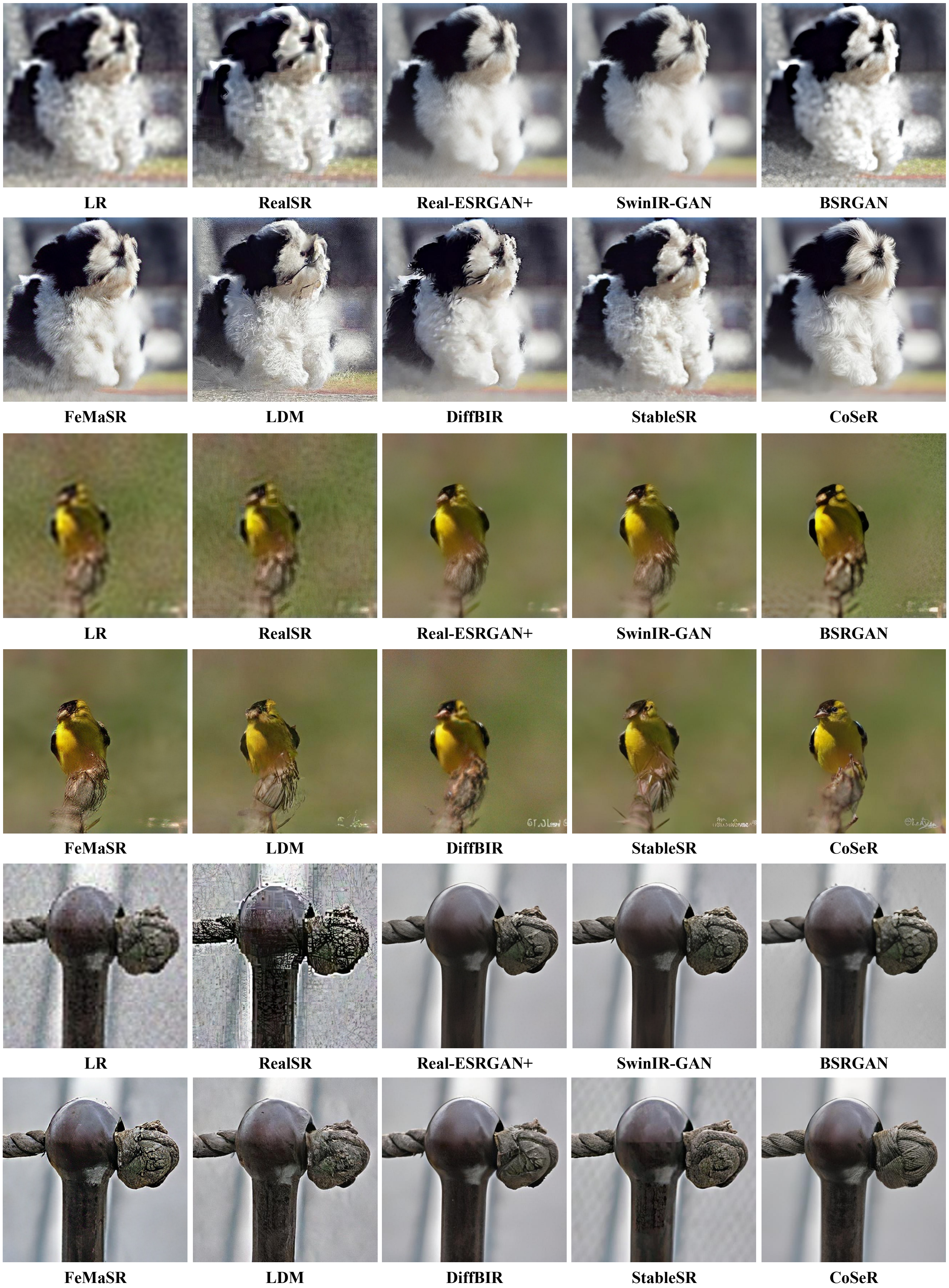
While acknowledging that PSNR and SSIM metrics exhibit a weak correlation with human perception, particularly for large-scale super-resolution tasks, we present the corresponding results in Table B.6 for reference purposes. Our CoSeR achieves favorable results. The All-in-Attention (AiA) module contributes to competitive or superior pixel-level fidelity compared to other diffusion-based models like DiffBIR and StableSR, as shown in Table B.3 and Table B.6.

C. Qualitative Comparisons

We provide visual comparisons on ImageNet Test2000 dataset (Figures D.7, D.8, D.9, D.10), real-world or unknown degradation type images (Figures D.11, D.12), RealSR dataset (Figure D.13) and DRealSR dataset (Figure D.14). Our CoSeR obtains outstanding visual performance.

D. Future Work

The cognitive-based recovery process extends beyond super-resolution (SR) tasks; it is beneficial for various visual tasks such as deblurring, denoising, and inpainting. Our future work includes expanding its application to more diverse image restoration tasks. Additionally, prevalent SR algorithms based on diffusion models often require a large number of sampling steps for higher visual quality. Addressing the challenge of accelerating the sampling process without compromising SR performance is also a focal point of our ongoing research.



LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

Figure D.7. Qualitative comparisons on ImageNet Test2000 dataset (part 1/4).

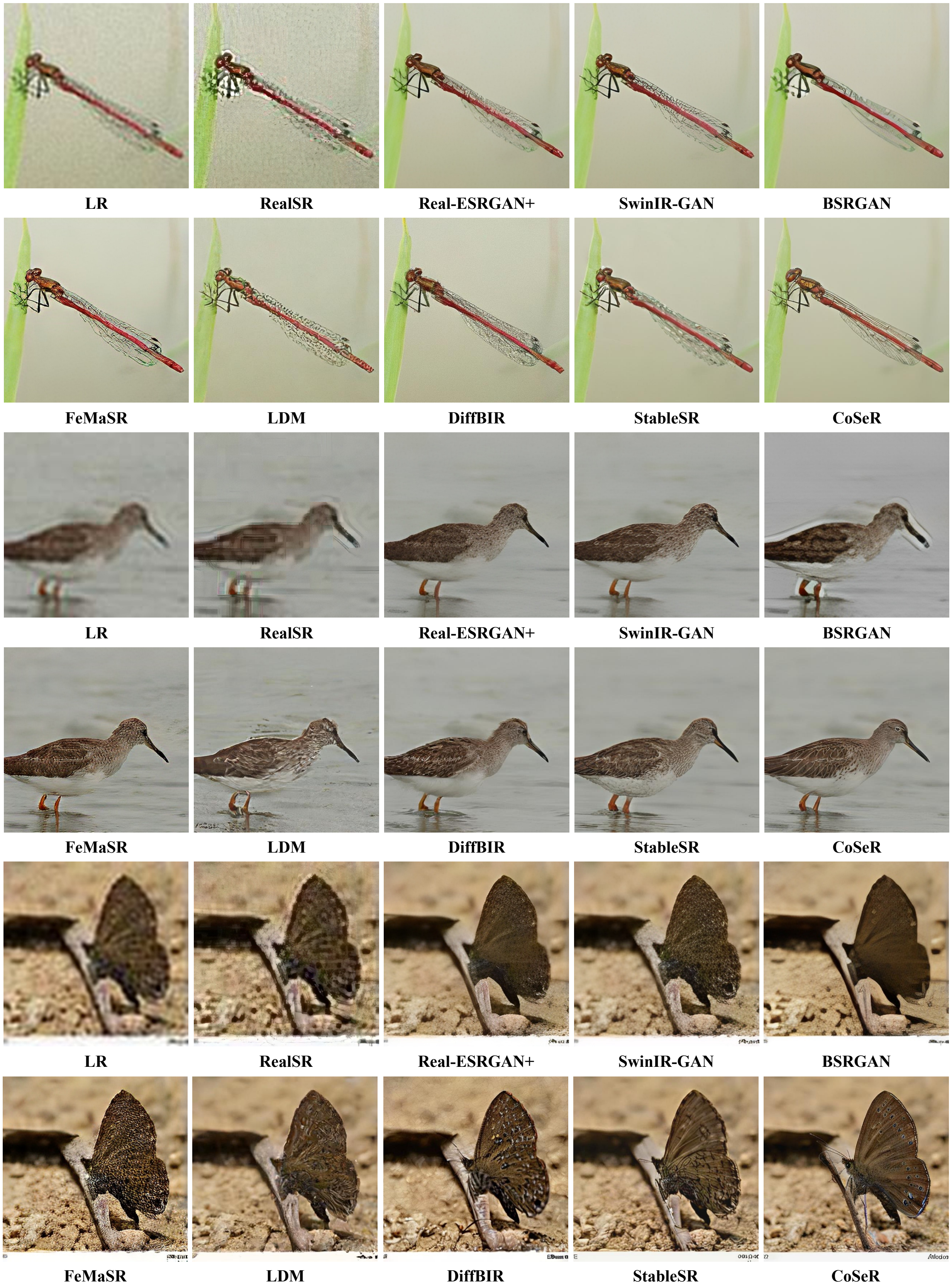


Figure D.8. Qualitative comparisons on ImageNet Test2000 dataset (part 2/4).

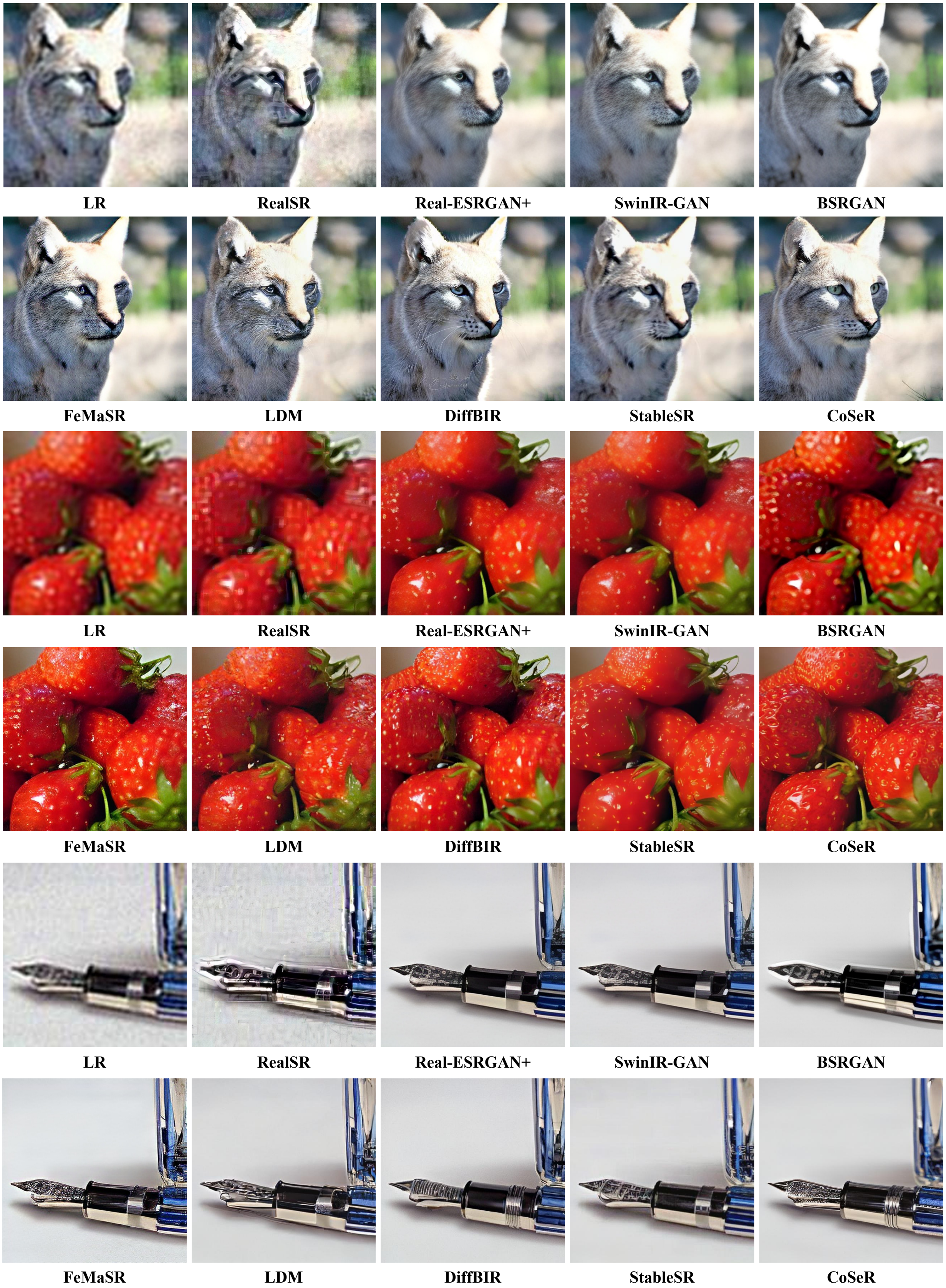
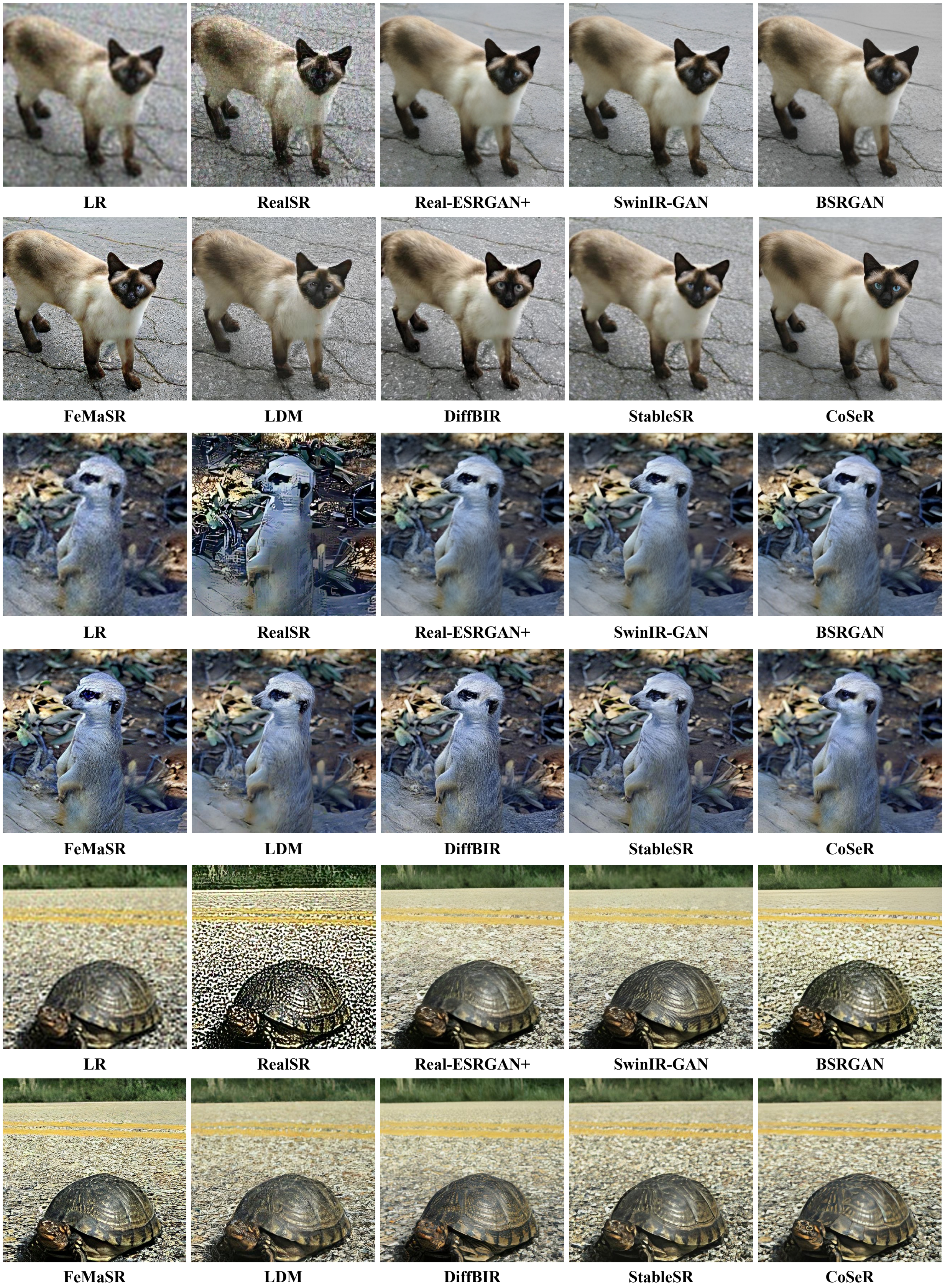


Figure D.9. Qualitative comparisons on ImageNet Test2000 dataset (part 3/4).



LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

LR RealSR Real-ESRGAN+ SwinIR-GAN BSRGAN

FeMaSR LDM DiffBIR StableSR CoSeR

Figure D.10. Qualitative comparisons on ImageNet Test2000 dataset (part 4/4).

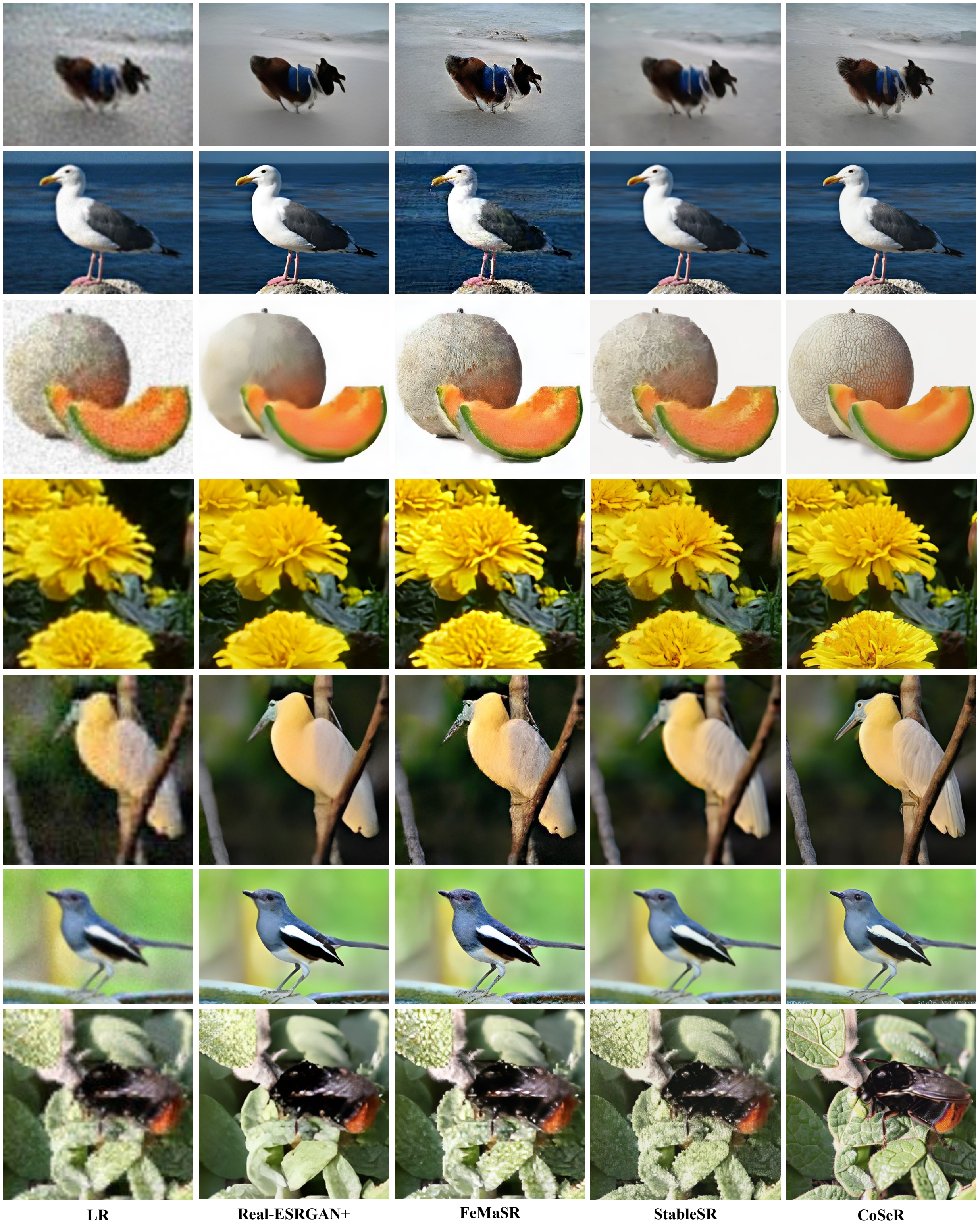


Figure D.11. Qualitative comparisons on real-world or unknown degradation type images (part 1/2).

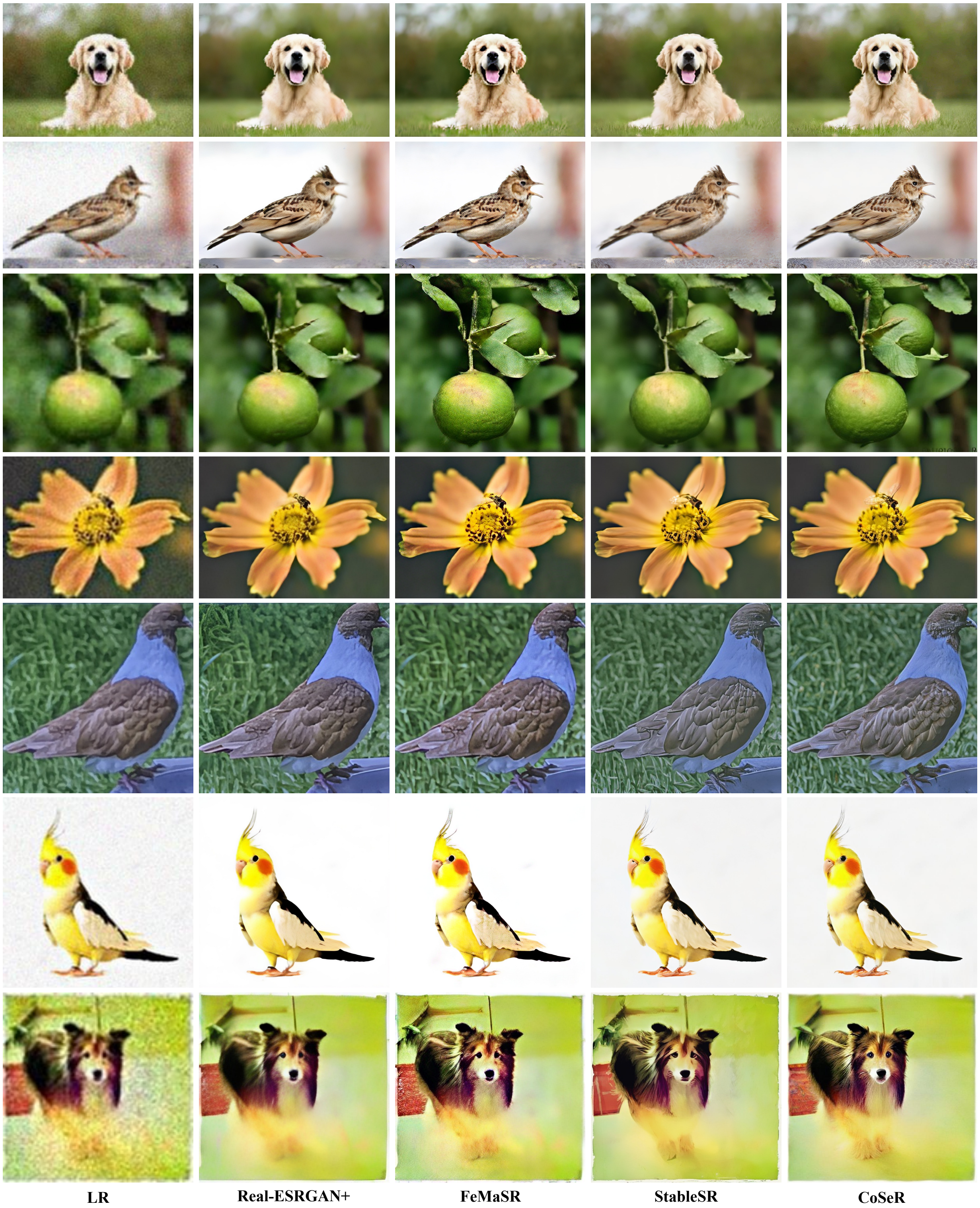


Figure D.12. Qualitative comparisons on real-world or unknown degradation type images (part 2/2).

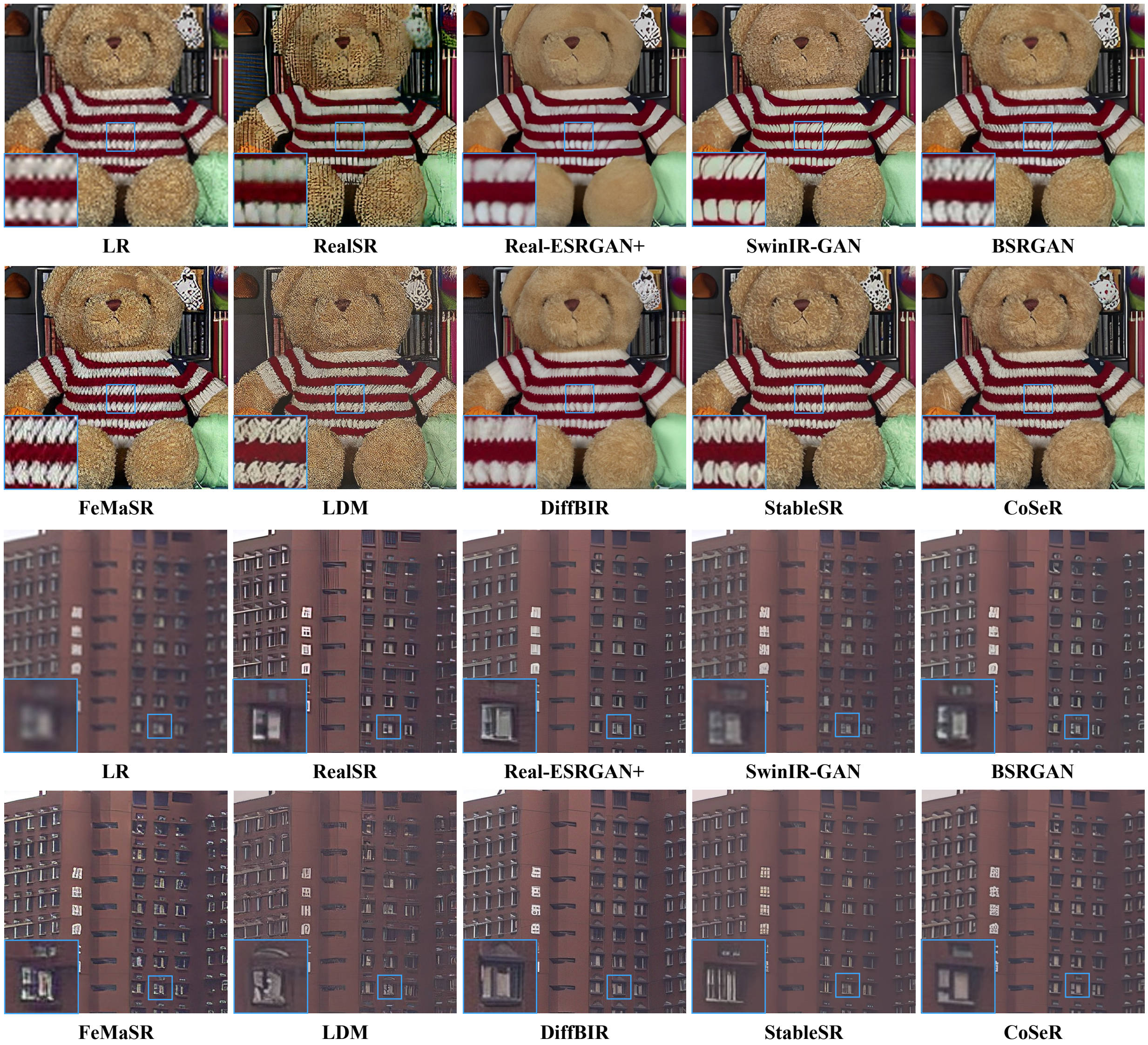


Figure D.13. Qualitative comparisons on RealSR dataset.

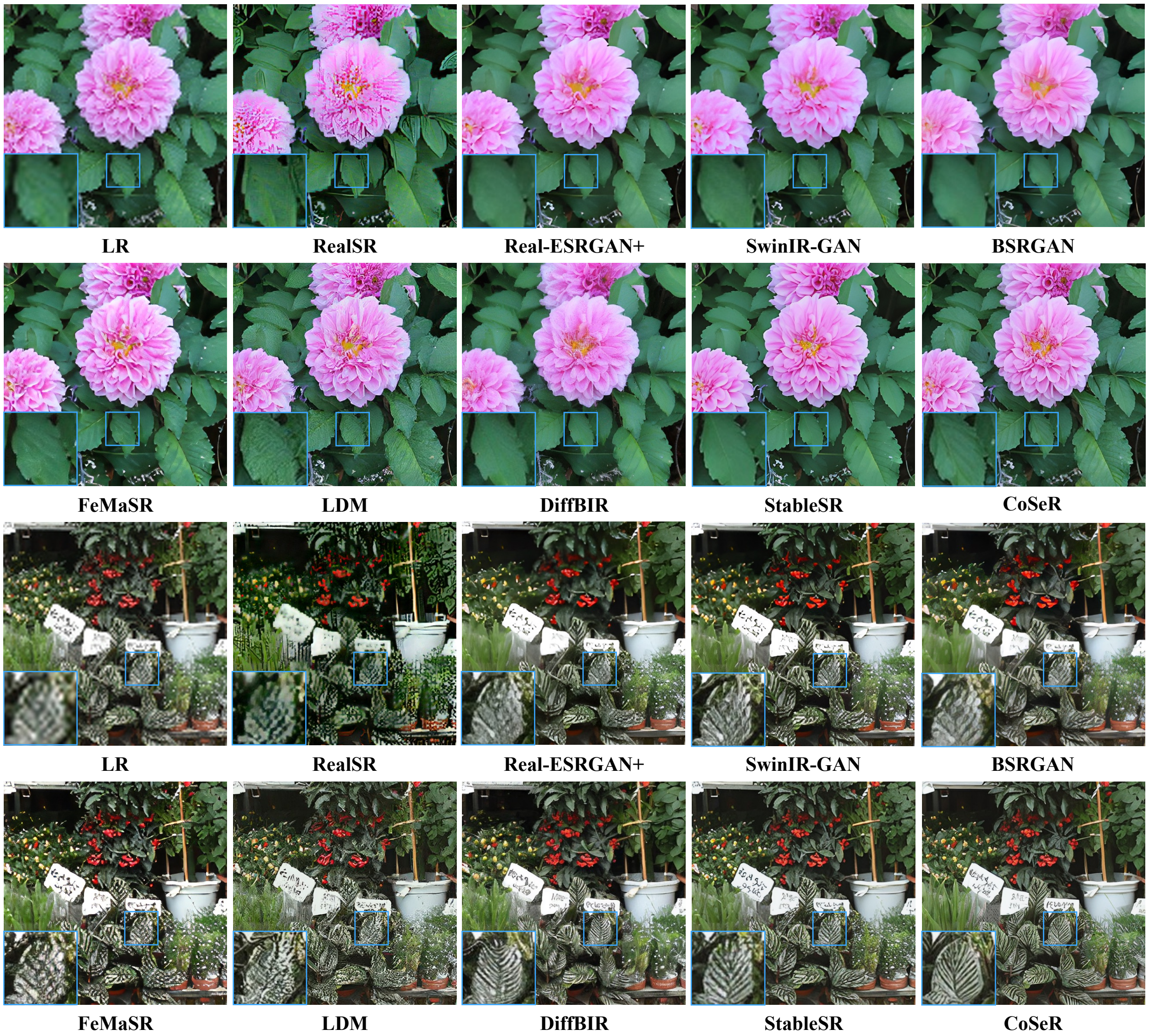


Figure D.14. Qualitative comparisons on DRealSR dataset.

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. 4

[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 *ICCV*, pages 3086–3095, 2019. 4

[3] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACMMM*, pages 1329–1338, 2022. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 3

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 44(5):2567–2581, 2020. 3

[6] 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. *NeurIPS*, 30, 2017. 3

[7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 4

[8] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, pages 466–467, 2020. 3

[9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 3

[10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 3

[11] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1, 3

[12] 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 *ICCV*, pages 416–423. IEEE, 2001. 4

[13] 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*, 76:21811–21838, 2017. 4

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[16] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3

[17] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 3

[18] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, pages 101–117. Springer, 2020. 4

[19] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 2

[20] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference*, pages 711–730. Springer, 2012. 4

[21] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 3

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 3