Better "CMOS" Produces Clearer Images: Learning Space-Variant Blur Estimation for Blind Image Super-Resolution Supplementary Materials

Xuhai Chen[1] Jiangning Zhang[2*] Chao Xu[1] Yabiao Wang[2] Chengjie Wang[2] Yong Liu[1†]
[1] APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent
{22232044, 21832066}@zju.edu.cn, yongliu@iipc.zju.edu.cn
{vtzhang, caseywang, jasoncjwang}@tencent.com

Overview

In this supplementary material, we provide additional details which we could not include in the main paper due to space limitations. We first performed some additional experiments on the CMOS model to further prove the effectiveness of it. Then we conducted more ablation studies for the module GIA and provided more discussions. Finally, we presented additional visual comparison results on the proposed datasets (NYUv2-BSR and Cityscapes-BSR), as well as on several real-world images.

A. Additional Experiments on CMOS

Impact of Other Degradations. Traditional degradation model [3, 9] is usually used to synthesize the LR input for SR as:

$$\boldsymbol{y} = [(\boldsymbol{x} \bigotimes \boldsymbol{k}) \downarrow_s + \boldsymbol{n}]_{JPEG}. \quad (1)$$

It assumes that the LR image $\boldsymbol{y}$ is obtained by first convolving the HR image $\boldsymbol{x}$ with a blur kernel $\boldsymbol{k}$, followed by a downsampling operation with scale factor s and an addition of noise $\boldsymbol{n}$. JPEG compression is also adopted here, as it is widely-used in real-world images. Our proposed datasets, *i.e.*, NYUv2-BSR and Cityscapes-BSR, are also designed based on the traditional model. We extended the blur kernels to space-variant domains, but do not consider noise and JPEG compression in our settings.

Therefore, in order to explore the performance of the proposed CMOS under more severe image degradations, we add additive Gaussian noise and JPEG compression to the NYUv2-BSR dataset and retrained the model. In the testing phase, the covariance σ of the additive Gaussian noise is set to 5, 10 and 15, respectively. The quality factor q of JPEG compression is set to 70 and 50. As shown in Tab. 1, CMOS can maintain its good performance when the noise or JPEG compression is not too severe. However, when the images suffer from both degradations, or when one of the degradations is too severe, it will not perform well. This indicates that while the proposed CMOS can be directly applied in scenarios with relatively mild noise or other forms of distortion, designing specialized estimation or recovery methods for other degradations, apart from blurring, is necessary when the degradations are more severe and diverse.

σ	q	PSNR ↑	SSIM ↑	mIoU ↑
5	✗	22.07	0.8157	32.32
10	✗	21.56	0.8101	30.70
15	✗	20.42	0.8025	28.70
✗	70	21.39	0.8110	30.92
✗	50	20.52	0.8021	29.12
5	70	20.92	0.8044	29.63
10	70	20.72	0.8042	28.61
15	70	19.94	0.7977	26.51
✗	✗	**24.52**	**0.8340**	**35.61**

Table 1. Impacts of additive Gaussian noise and JPEG compression on the blur estimation and semantic segmentation performance of CMOS. σ is the covariance of the additive Gaussian noise and q is the quality factor q of JPEG compression.

Space-variant vs. Space-invariant. To further demonstrate that CMOS can handle different blur types, we present the results of different methods on space-invariant and space-variant blur respectively. The space-invariant results represent the average PSNR and SSIM values for the space-invariant blurred images within each test group, while the space-variant results correspond to the average values for the space-variant ones within each test group.

As shown in the Tab. 2 and Tab. 3, although the previous advanced SR methods which only focus on space-invariant

*Equal contribution.
†Corresponding author.

Method	Group1	Group2	Group3	Group4	Group5	Avg.
KernelGAN [2]	23.22/23.08	24.19/22.86	23.27/23.15	23.45/23.08	23.35/23.14	23.39/23.06
KOALAnet [6]	27.64/27.70	28.58/27.52	27.93/27.52	29.25/27.35	29.87/27.20	28.65/27.46
DCLS [11]	28.16/27.82	29.14/27.64	28.67/27.60	30.14/27.35	31.04/27.08	29.43/27.50
DAN [5]	32.86/26.67	32.73/26.80	32.90/26.57	33.53/26.52	32.50/26.69	32.90/26.65
MANet [8]	29.83/30.24	30.91/30.03	30.09/30.06	30.92/29.86	31.42/29.76	30.63/29.99
CMOS (ours)	31.68/32.19	32.99/31.85	31.69/32.07	32.58/31.80	32.83/31.80	32.35/31.94
Upper Bound	33.31/33.92	34.79/33.53	32.97/33.87	33.76/33.72	33.75/33.74	33.72/33.76

Table 2. Average PSNR of different methods for **Space-invariant/Space-variant** blind SR on NYUv2-BSR. Avg. represents the average results on the 5 test groups. The best and second best results are highlighted in red and blue colors, respectively.

Method	Group1	Group2	Group3	Group4	Group5	Avg.
KernelGAN [2]	0.7508/0.7410	0.7812/0.7346	0.7454/0.7448	0.7537/0.7427	0.7426/0.7470	0.7547/0.7420
KOALAnet [6]	0.8722/0.8786	0.8896/0.8736	0.8702/0.8741	0.8862/0.8727	0.8940/0.8714	0.8824/0.8741
DCLS [11]	0.8795/0.8800	0.8975/0.8754	0.8833/0.8741	0.8940/0.8725	0.9098/0.8699	0.8928/0.8744
DAN [5]	0.9214/0.8709	0.9232/0.8703	0.9225/0.8658	0.9242/0.8659	0.9175/0.8696	0.9218/0.8685
MANet [8]	0.9095/0.9122	0.9216/0.9085	0.9037/0.9110	0.9116/0.9095	0.9176/0.9089	0.9128/0.9100
CMOS (ours)	0.9143/0.9174	0.9254/0.9135	0.9079/0.9161	0.9154/0.9146	0.9205/0.9140	0.9167/0.9151
Upper Bound	0.9283/0.9316	0.9377/0.9285	0.9190/0.9315	0.9266/0.9309	0.9283/0.9302	0.9280/0.9305

Table 3. Average SSIM of different methods for **Space-invariant/Space-variant** blind SR on NYUv2-BSR. Avg. represents the average results on the 5 test groups. The best and second best results are highlighted in red and blue colors, respectively.

Metrics	MANet [8]	DCLS [11]	Ours
PSNR ↑	30.97	**32.04**	31.99
SSIM ↑	0.8650	0.8907	**0.8971**

Table 4. Experiments on the benchmark evaluation dataset BSD100 [12]. The SR scale factor is set to 2.

Methods	Avg. PSNR ↑	Avg. SSIM ↑
Real-ESRGAN [13]	27.61	0.8825
BSRGAN [14]	29.38	0.8977
SwinIR [7]	29.42	0.9042
CMOS (ours)	**32.03**	**0.9154**

Table 5. Comparisons with three implicit modeling methods. We fine-tune the models on the NYUv2-BSR dataset.

blur, *i.e.*, KernelGAN [2], DCLS [11] and DAN [5], can deal with the situation of space-invariant blur well, the restoration performance of space-variant blur is poor. DAN, in particular, has an average gap of 6.25 dB and 0.0533 in PSNR and SSIM. By contrast, the performance of SR methods aiming to solve space-variant problems, *i.e.*, KOALAnet [6] and MANet [8], is similar in both cases. Notably, the proposed method CMOS can achieve the best results in space-variant blur, and also maintains its performance in space-invariant blur.

Perceptual-based Metric. To further prove the advantage of our method, we report a perceptual-based metric Learned Perceptual Image Patch Similarity (LPIPS) [15], which provides a perceptual distance between the SR image and the ground-truth. The lower the value of this metric, the higher the perceived quality of the restored image. As shown in Tab. 6, the proposed model CMOS produces the best results compared to other methods, which is consistent with the visualization results in the qualitative comparisons.

Other Comparision methods. There are two mainstreams of research on blind image SR, namely explicit modeling methods [2, 5, 8], and implicit modeling methods [7, 13, 14]. The explicit modeling methods first extract the blur kernels from the LR images and then use them to synthesize HR images from the LR images, while the implicit modeling methods restores HR images directly from the LR images without kernel estimation. Although our approach follows the first line of research, we also compare other three implicit modeling approaches, *i.e.*, Real-ESRGAN [13], BSR-

Method	Group1	Group2	Group3	Group4	Group5	Avg.
KernelGAN [2]	0.2826	0.2820	0.2843	0.2799	0.2813	0.2820
KOALAnet [6]	0.2211	0.2227	0.2263	0.2232	0.2231	0.2233
DCLS [11]	0.2366	0.2361	0.2429	0.2423	0.2416	0.2399
DAN [5]	0.2407	0.2399	0.2468	0.2463	0.2457	0.2439
MANet [8]	0.1455	0.1458	0.1483	0.1468	0.1466	0.1466
CMOS (ours)	0.1313	0.1314	0.1337	0.1329	0.1329	0.1324

Table 6. A perceptual-based metric LPIPS [15] of different methods for space-variant blind SR on NYUv2-BSR. Avg. represents the average results on the 5 test groups. The best and second best results are highlighted in red and blue colors, respectively.

Window Size	PSNR ↑	SSIM ↑	mIoU ↑
5×5	24.08	0.8333	35.60
5×10	24.11	0.8332	35.54
15×15	24.13	0.8325	35.39
15×20 (ours)	**24.52**	**0.8340**	**35.61**

Table 7. Influence of different window sizes in the spatial grouping feature interaction part of the proposed module GIA.

Layers	PSNR ↑	SSIM ↑	mIoU ↑
2	23.98	0.8326	35.45
4	24.03	0.8322	35.38
6	23.37	0.8275	35.40
1 (ours)	**24.52**	**0.8340**	**35.61**

Table 8. Influence of different number of layers in the spatial grouping feature interaction part of the proposed module GIA.

GAN [14] and SwinIR [7]. We fine-tune them on the NYUv2-BSR dataset and the average results of the five test groups are reported in Tab. 5. Note that for RRDBNet in Real-ESRGAN, we use the same hyperparameters as our method to ensure a fair comparison.

Other Datasets. To fully illustrate the effectiveness of our method, we carried out additional experiments on a benchmark evaluation dataset BSD100 [12]. Since the window size in our default model setting is 15×20, the SR scale factor is set to 2 in this experiment. We use the isotropic gaussian kernels to blur the images and the kenel sizes are fixed to 21×21. The kernel width of the train set is uniformly sampled from range [0.2, 2.0], and we fine-tune our model trained on the Cityscapes-BSR on it. For testing, we use the Gaussian8 [4, 11] kernel setting to generate the evaluation dataset and the results are shown in Tab. 4. Although the semantic information in the BSD100 does not quite agree with that in the Cityscapes-BSR, CMOS still achieves comparable results to the SOTA method DCLS [11], *e.g.* similar PSNR and better SSIM by +0.0064↑.

B. Ablation Studies of GIA

Impact of Different Window Sizes in GIA. In the spatial grouping feature interaction part of GIA, we first divide the feature maps into windows of size $H \times W$, and then carried out subsequent operations for each window. The windows are set to 15×20 in our default settings, which is the resolution of the minimum feature maps in the hierarchical structure of CMOS. We also tried other windows of different sizes, and the results are shown in Tab. 7. For blur estimation, the larger the window is, the better the performance is. With the increase of the window size, PSNR increases from 24.08 to 24.52 dB. Although this is not the case with SSIM, the best result is also achieved when the window size is 15×20. For semantic segmentation, the window size does not seem to matter much and similar results are obtained for all four sizes of windows.

Impact of Different Layers in GIA. Since the features only interact in windows in the spatial grouping feature interaction part of GIA and the global information may be insufficient, we also tried the shifted window scheme in [10] and applied more layers. The window size is set to 15×20. In even layers, the shift size is set to 0, and in odd layers, the shift size is set to half of the window size. As shown in Tab. 8, the shifted window scheme seems not work for GIA and more layers are not beneficial for the results of both blur estimation and semantic segmentation. This may be because the channel grouping feature interaction part of GIA is already sufficient to capture global information, so the spatial part only needs to focus on local information and the global information here may diminish the emphasis on local details.

C. More Discussions

While intentional defocus can be used for artistic effect, it can be frustrating and unnecessary in certain applications.

Ours (MANet)	Test-NYU	Test-City
Train-NYU	32.0/0.9154 (30.1/0.9106)	32.5/0.9199 (27.7/0.8314)
Train-City	27.3/0.8844 (20.2/0.7104)	35.6/0.9383 (34.3/0.9287)

Table 9. Generalization property of the model CMOS. NYU refers to the NYUv2-BSR dataset, while City refers to the Cityscapes-BSR dataset.

As is commonly known, using a wide aperture results in a shallow depth of field, causing defocus in the images. However, there are many real-life examples that require a wide aperture but still want all-in-focus images. An example is the cameras on the self-driving cars, or on cars that map environments, where it uses a fixed shutter speed and the only way to get sufficient light is a wide aperture at the cost of defocus [1]. Therefore, it is very promising to explore the removal of the defocus blur in the SR field.

In order to better eliminate the defocus blur and restore more accurate image details, we designed a semantically relevant model called CMOS and treated the indoor and outdoor scenes differently. We compared our model with MANet cross datasets to evaluate its generalization performance using the PSNR/SSIM metrics, as shown in Tab. 9. The results demonstrate that our method outperforms MANet in terms of generalization.

D. More Qualitative Comparisons

In this section, we present more qualitative comparisons and compare our method against several state-of-the-art SR methods [2,5,6,8,11]. Specifically, we show visual comparisons of various methods on the proposed datasets NYUv2-BSR and Cityscapes-BSR for 4× SR in Fig. 1 and Fig. 2, respectively. It is worth noting that the last rows are two examples of space-invariant blurred images. Furthermore, we also demonstrate several applications of processing real-world images. In these cases, the ground-truth images are not available. As shown in Fig. 3, CMOS can reconstruct sharper and more accurate images than the state-of-the-art SR approaches.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 111–126. Springer, 2020. 4

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 4

[3] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 1

[4] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 3

[5] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33:5632–5643, 2020. 2, 3, 4

[6] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10611–10620, 2021. 2, 3, 4

[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. 2, 3

[8] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4096–4105, 2021. 2, 3, 4

[9] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 1

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[11] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17652, 2022. 2, 3, 4

[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 *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 2, 3

[13] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data supplementary material. 2

[14] 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. 2, 3

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3

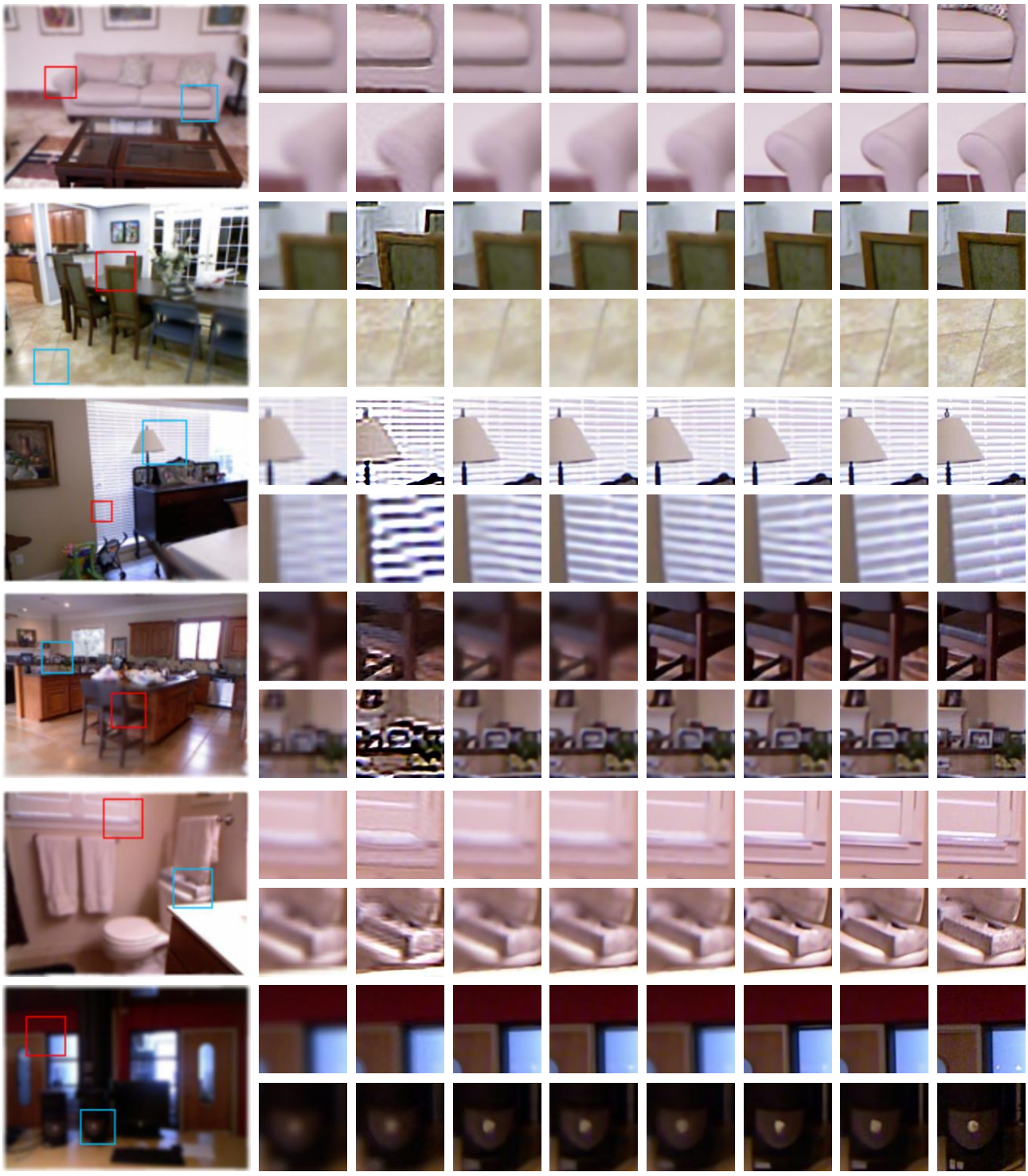

Figure 1. Qualitative comparisons between different SR methods on space-variant (out-of-focus) and space-invariant blur. The last row is an example of space-invariant blur, and the rest are examples of space-variant blur. The images are from the proposed dataset NYUv2-BSR. **(Please zoom in for better view.)**

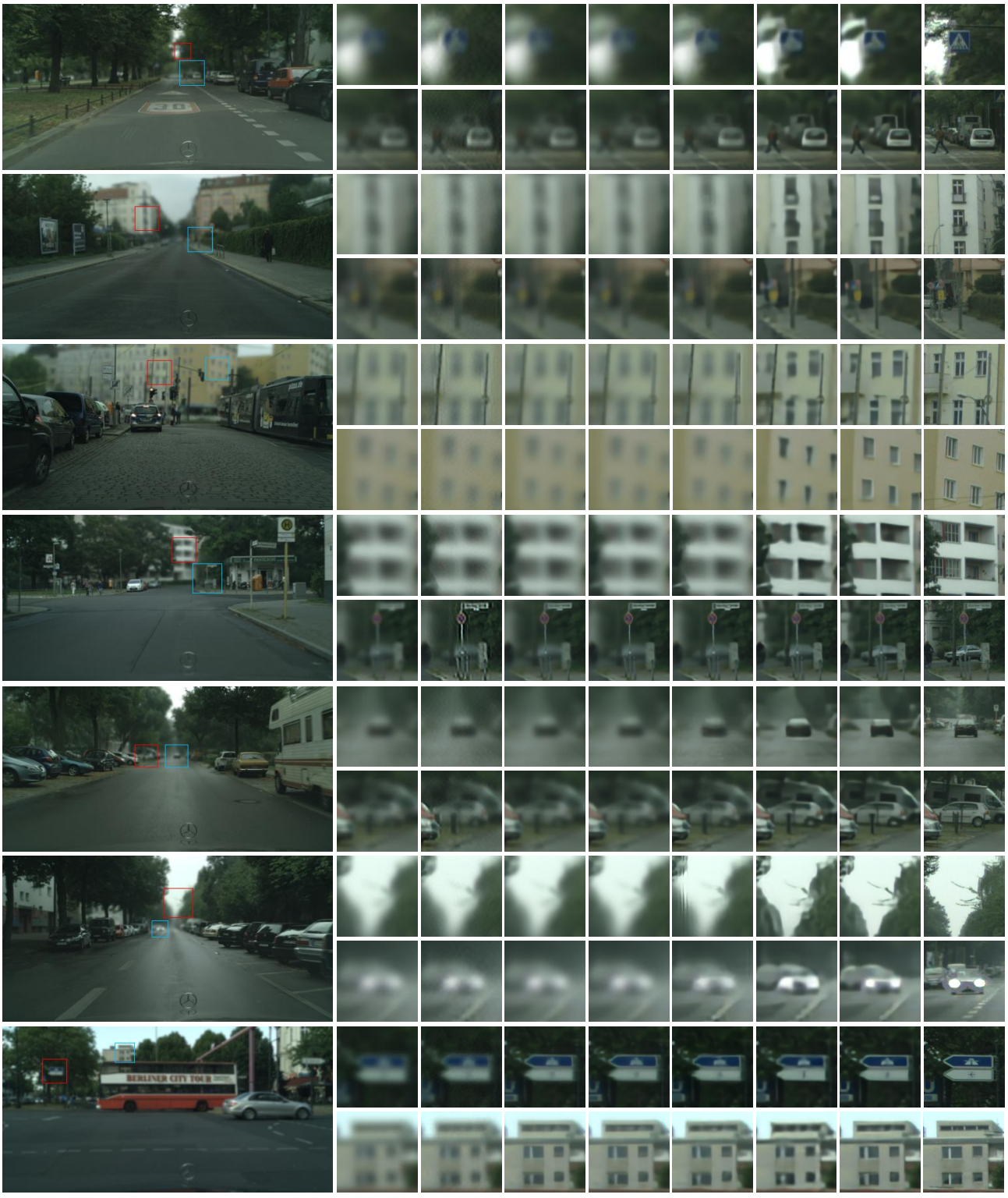

Figure 2. Qualitative comparisons between different SR methods on space-variant (out-of-focus) and space-invariant blur. The last row is an example of space-invariant blur, and the rest are examples of space-variant blur. The images are from the proposed dataset Cityscapes-BSR. **(Please zoom in for better view.)**

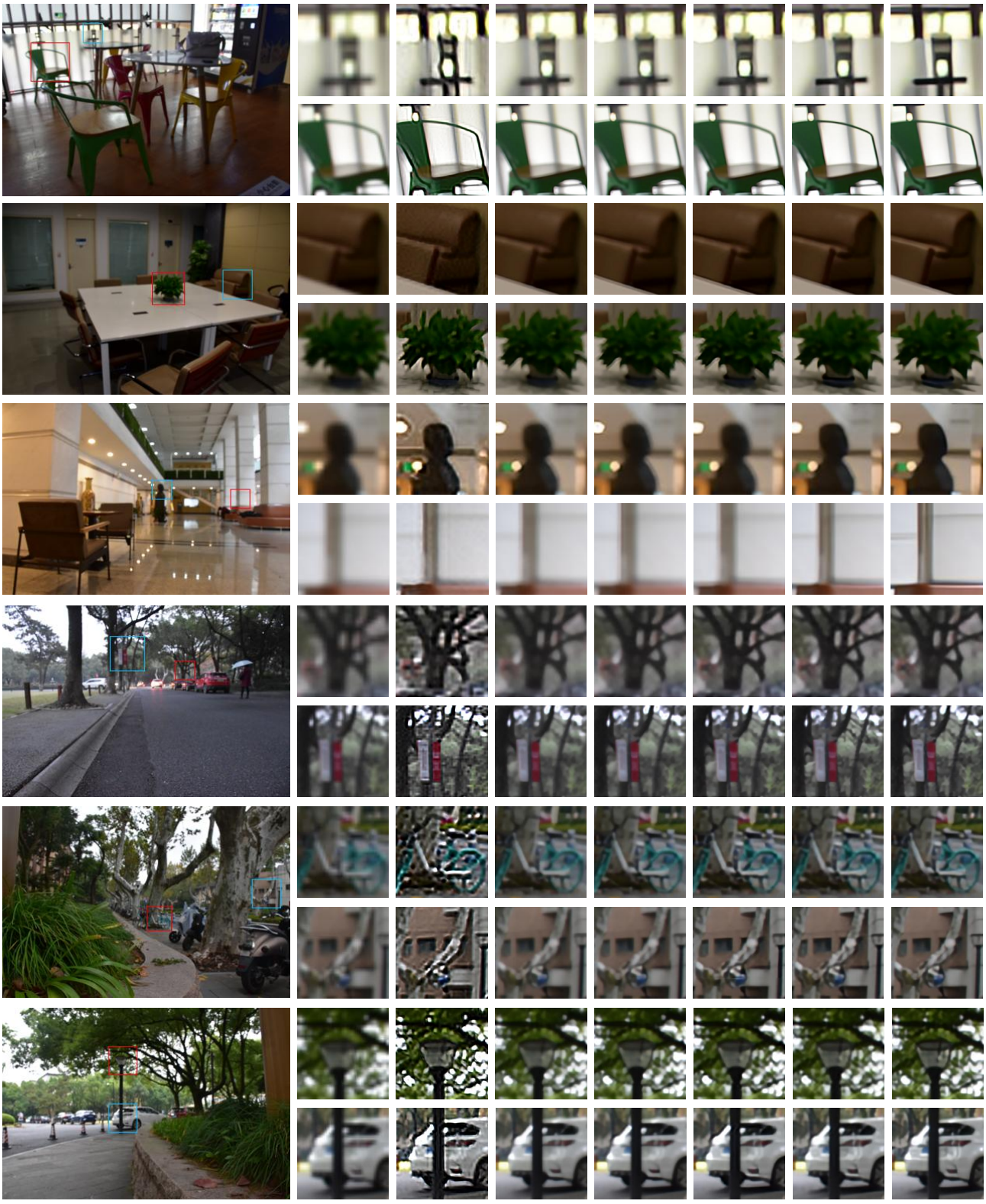

Figure 3. Visual results on real-world images for scale factor 4. The first three images of the indoor scene use the models trained on NYUv2-BSR, and the last three images of the outdoor scene use the models trained on Cityscapes-BSR. **(Please zoom in for better view.)**