Supplemental Material:
DiffIR: Efficient Diffusion Model for Image Restoration

Bin Xia [1,3], Yulun Zhang [2], Shiyin Wang [3], Yitong Wang [3],
Xinglong Wu [3], Yapeng Tian [4], Wenming Yang [1] *, and Luc Van Gool [2]
[1] Tsinghua University, [2] ETH Zürich, [3] ByteDance Inc, [4] University of Texas at Dallas

1. Evaluation on Real-world SR

We train and validate our DiffIR$_{S2}$ on real-world SR using the same settings of Real-ESRGAN [15]. Specifically, we adopt the same loss functions of Real-ESRGAN [16], which further introduce perceptual loss and adversarial loss to the basic $\mathcal{L}_1$ loss. We set the learning rate of the KDSR$_T$ to 2×10^{-4}. We further validate the effectiveness of KDSR on Real-World datasets. For optimization, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.99$. In both two stages of training, we set the batch size to 64, with the input patch size being 64. We evaluate all methods on the dataset provided in the challenge of Real-World Super-Resolution: NTIRE2020 Track1 and Tracks [9]. In addition, we also validate our DiffIR on RealSRSet [2]. Since NTIRE2020 Track1 and RealSRSet datasets provide a paired validation set, we use the LPIPS [22], DISTS [4], and PSNR for the evaluation.

The quantitative results are shown in Tab. 1. We can see that DiffIR$_{S2}$ outperforms SOTA real-world SR method KDSR$_S$-GAN on LPIPS, DISTS, and PSNR, consuming fewer computational costs. In addition, we can see that DiffIR$_{S2}$ outperforms classic real-world SR method Real-ESRGAN on LPIPS, DISTS, and PSNR, only consuming its 63% Mult-Adds. Furthermore, compared with DM-based LDM [11], DiffIR$_{S2}$ achieve much better performance consuming only 2% Mult-Adds.

We also visualize the results on NTIRE2020 Track2, which was captured with smartphones. The qualitative results are shown in Fig. 1. We can see that DiffIR$_{S2}$ achieves the best performance.

2. Algorithm

The algorithm of DiffIR$_2$ training is summarized in Alg. 1. The algorithm of DiffIR$_2$ inference is summarized in Alg. 2.

3. More Training Details on Inpainting

We train our DiffIR for inpainting using the same loss functions of LaMa [12], which further introduce multiple perceptual losses and adversarial loss to the basic $\mathcal{L}_1$ loss.

For our experiments on image-inpainting in the paper Sec. 5.2, we used the code of LaMa [12] to generate synthetic masks. In training, we adopt the Adam optimizer with learning rates 0.0002 and 0.0001 for DiffIR and discriminator networks, respectively. All models are trained for 1M iterations with a batch size of 30. In addition, we use random crops of size 256×256 to train DiffIR on Places and CelebA-HQ. In testing, we use a fixed set of 2k validation and 30k testing samples from CelebA-HQ [5] and Places [23]. Moreover, we validate DiffIR$_{S2}$ on crops of size 512×512 and 256×256 on Places and CelebA-HQ validation datasets, respectively.

4. More Training Details on SR

Compared with DIRformer for other IR tasks, we add a $\times 4$ upsampling network [16] at the end of DIRformer for super-resolution (SR). We train our DiffIR for SR using the same loss functions of ESRGAN [16], which further introduce perceptual loss and adversarial loss to the basic $\mathcal{L}_1$ loss.

We train DiffIR for 1M iterations with a batch size of 64. In addition, we use random crops of size 256×256 to train DiffIR on DIV2K [1] (800 images) and Flickr2K [13] (2650 images) datasets for $4\times$ super-resolution. We train our DiffIR using Adam optimizer with learning rates 0.0002 and 0.0001 for DiffIR and discriminator networks, respectively.

5. More Training Details on deblurring

Following previous works in single image motion deblurring [3, 19, 18], we train our DiffIR only using $\mathcal{L}_1$ loss for fair comparisons. We train DiffIR for 300K iterations with the initial learning rate 2^{-4} gradually reduced to 1^{-6} with the cosine annealing [7]. Following previous work [18], we progressively increase patch size and decrease batch size. Specifically, we start training with patch size 128×128 and batch size 64. The patch size and batch size pairs are updated to $[(160 \times 160, 40), (192 \times 192, 32), (256 \times 256, 16), (320 \times 320, 8), (384 \times 384, 8)]$ at iterations $[92K, 156K, 204K, 240K, 276K]$.

*Corresponding Author

Table 1. 4× SR quantitative comparison on real-world SR competition benchmarks. The Mult-Adds are computed based on an LR size of 256 × 256. Best and second best performance are marked in bold and underlined, respectively. The bottom two methods marked in gray adopt the diffusion model.

Methods	Mult-Adds (T)	RealSRSet [2]			NTIRE2020 Track1 [9]		
		LPIPS↓	DISTS↓	PSNR↑	LPIPS↓	DISTS↓	PSNR↑
BSRGAN [21]	1.18	0.3648	0.1676	26.90	0.3691	0.1368	26.75
Real-ESRGAN [15]	1.18	0.3629	$\underline{0.1609}$	26.07	0.3471	0.1326	26.40
KDSR$_s$-GAN [17]	0.86	$\underline{0.3610}$	0.1627	$\underline{27.18}$	$\underline{0.3198}$	$\underline{0.1252}$	$\underline{27.12}$
LDM [11]	37.25	0.4369	0.1982	26.37	0.4763	0.1844	25.68
DiffIR$_{S2}$ (Ours)	0.74	**0.3527**	**0.1588**	**27.65**	**0.3088**	**0.1131**	**27.31**

Algorithm 1 DiffIR$_{S2}$ Training

Input: Trained DiffIR$_{S1}$ (including CPEN$_{S1}$ and DIRformer), $\beta_t (t \in [1, T])$.
Output: Trained DiffIR$_{S2}$.

Init: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$.
Init: The DIRformer of DiffIR$_{S2}$ copies the parameters of trained DiffIR$_{S1}$.
for I_{LQ}, I_{GT} **do**
  $\mathbf{Z} = \mathrm{CPEN_{S1}}(\mathrm{PixelUnshuffle}(\mathrm{Concat}(I_{GT}, I_{LQ})))$. (paper Eq. (**5**))
  Diffusion Process:
  We sample $\mathbf{Z}_T$ by $q\left(\mathbf{Z}_T \mid \mathbf{Z}\right) = \mathcal{N}\left(\mathbf{Z}_T; \sqrt{\bar{\alpha}_T}\mathbf{Z}, (1-\bar{\alpha}_T)\,\mathbf{I}\right)$ (*i.e.*, diffusion process. paper Eq. (**10**))
  Reverse Process:
  $\hat{\mathbf{Z}}_T = \mathbf{Z}_T$
  $\mathbf{D} = \mathrm{CPEN_{S2}}(\mathrm{PixelUnshuffle}(I_{LQ}))$ (paper Eq. (**12**))
  for $t = T$ to 1 **do**
    $\hat{\mathbf{Z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{Z}}_t - \epsilon_\theta(\mathrm{Concat}(\hat{\mathbf{Z}}_t, t, \mathbf{D}))\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)$ (paper Eq. (**11**))
  end for
  $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}_0$
  $\hat{I}_{HQ} = \mathrm{DIRformer}(I_{LQ}, \hat{\mathbf{Z}})$
  Calculate $\mathcal{L}_{diff}$ loss (paper Eq. (**13**)).
end for
Output the trained model DiffIR$_{S2}$.

6. More Visual Comparisons on Inpainting

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA inpainting methods (ICT [14], LaMa [12], and RePaint [8]). The results are shown in Fig 2. We can observe that our DiffIR$_{S2}$ can produce more realistic and reasonable structures and details than other competitive inpainting methods.

7. More Visual Comparisons on SR

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA GAN-based SR methods. The results are shown in Figs 3 and 4. Our DiffIR$_{S2}$ achieves the best visual quality containing more realistic details.

8. More Visual Comparisons on Deblurring

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA image motion deblurring methods. The results are shown in Fig 5. Our DiffIR$_{S2}$ has the best visual quality containing more realistic details close to corresponding HQ images.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 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 *ICCV*, 2019. 1, 2

[3] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 2021. 1

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 2020. 1

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

Algorithm 2 DiffIR$_{S2}$ Inference

Input: Trained DiffIR$_{S2}$ (including CPEN$_{S2}$ and DIRformer), $\beta_t (t \in [1, T])$, LQ images I_{LQ}.
Output: Restored HQ images $\hat{I}_{HQ}$.

Init: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$.
Reverse Process:
Sample $\hat{\mathbf{Z}}_T \sim \mathcal{N}(0, 1)$
$\mathbf{D} = \mathrm{CPEN}_{\mathrm{S2}}(\mathrm{PixelUnshuffle}(I_{LQ}))$ (paper Eq. (12))
for $t = T$ to 1 **do**
  $\hat{\mathbf{Z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{\mathbf{Z}}_t - \epsilon_\theta(\mathrm{Concat}(\hat{\mathbf{Z}}_t, t, \mathbf{D})) \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \right)$ (paper Eq. (11))
end for
$\hat{\mathbf{Z}} = \hat{\mathbf{Z}}_0$
$\hat{I}_{HQ} = \mathrm{DIRformer}(I_{LQ}, \hat{\mathbf{Z}})$
Output restored HQ images $\hat{I}_{HQ}$.

[6] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *AAAI*, 2022. 6, 7

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2017. 1

[8] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2, 5

[9] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *CVPRW*, 2020. 1, 2

[10] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, 2020. 8

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 6, 7

[12] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 1, 2, 5

[13] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 1

[14] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, 2021. 2, 5

[15] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 1, 2, 4

[16] 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*, pages 0–0, 2018. 1

[17] Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. *ICLR*, 2023. 2, 4

[18] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 8

[19] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 8

[20] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020. 6, 7

[21] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. *arXiv preprint arXiv:2103.14006*, 2021. 2, 4

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[23] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 1

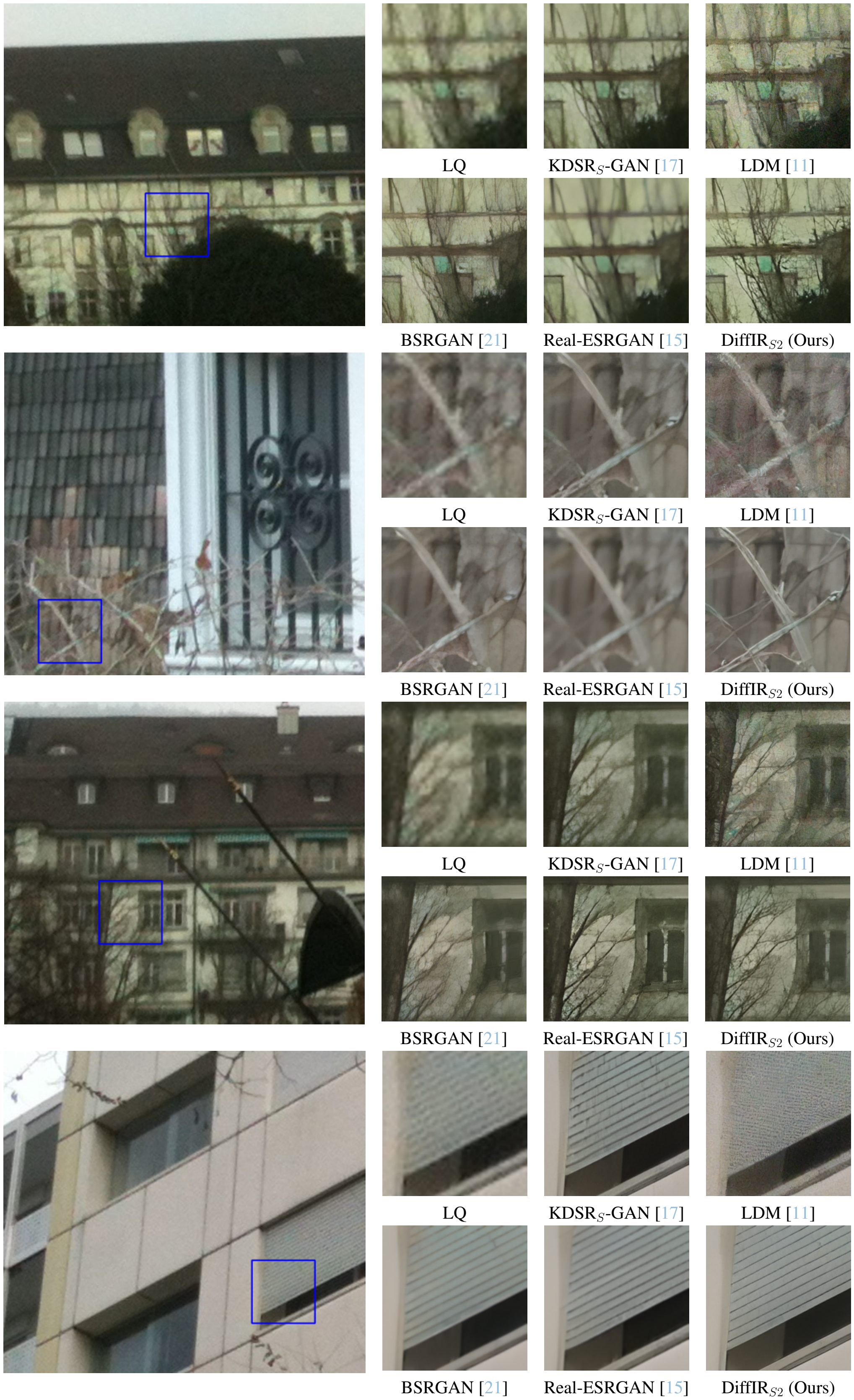
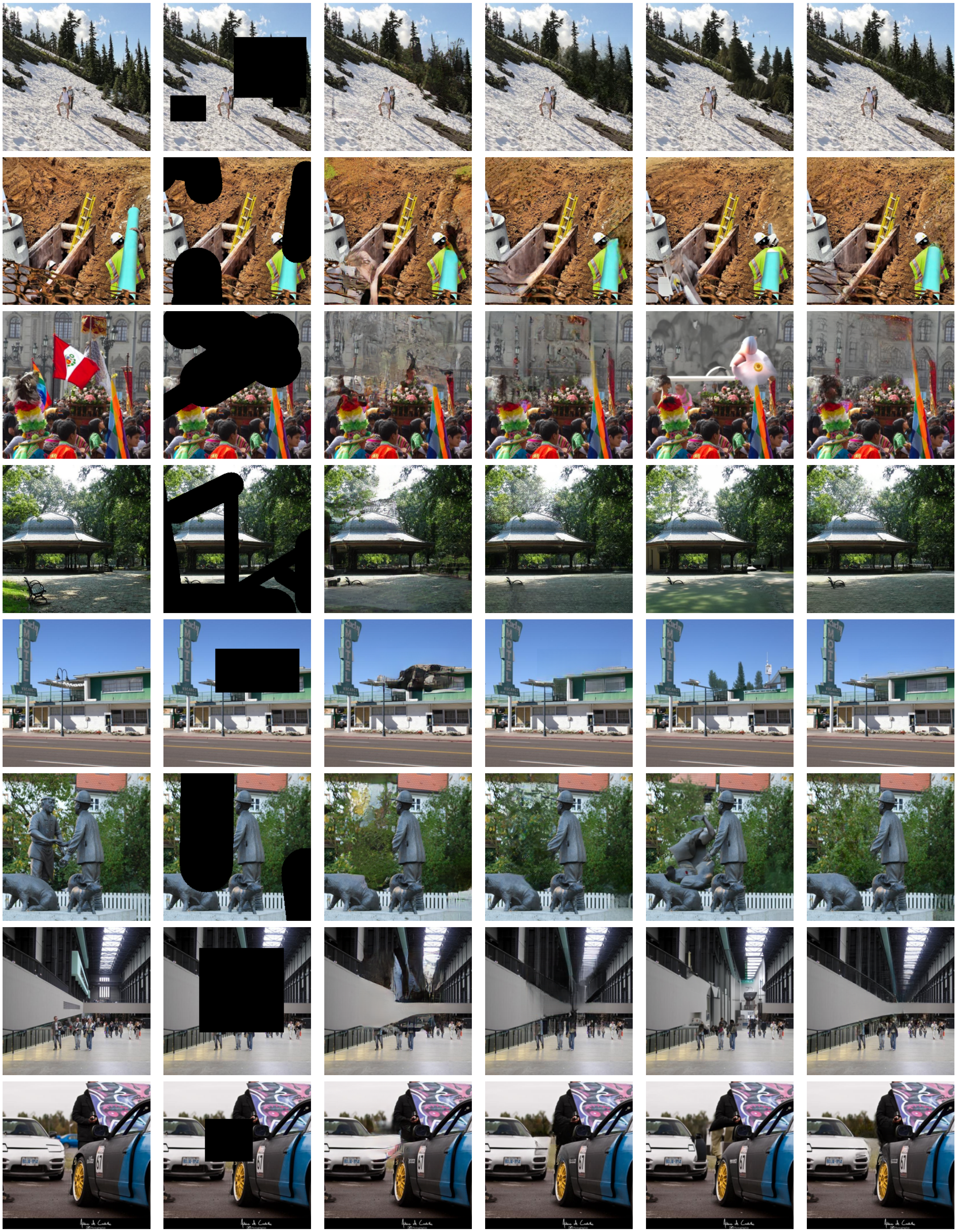


Figure 1. Visual comparison of 4× **real-world super-resolution** methods. Zoom-in for better details.



HQ | LQ | ICT [14] | LaMa [12] | RePaint [8] | DiffIR$_{S2}$ (Ours)

Figure 2. More visual comparisons of **inpainting** methods. Zoom-in for better details.

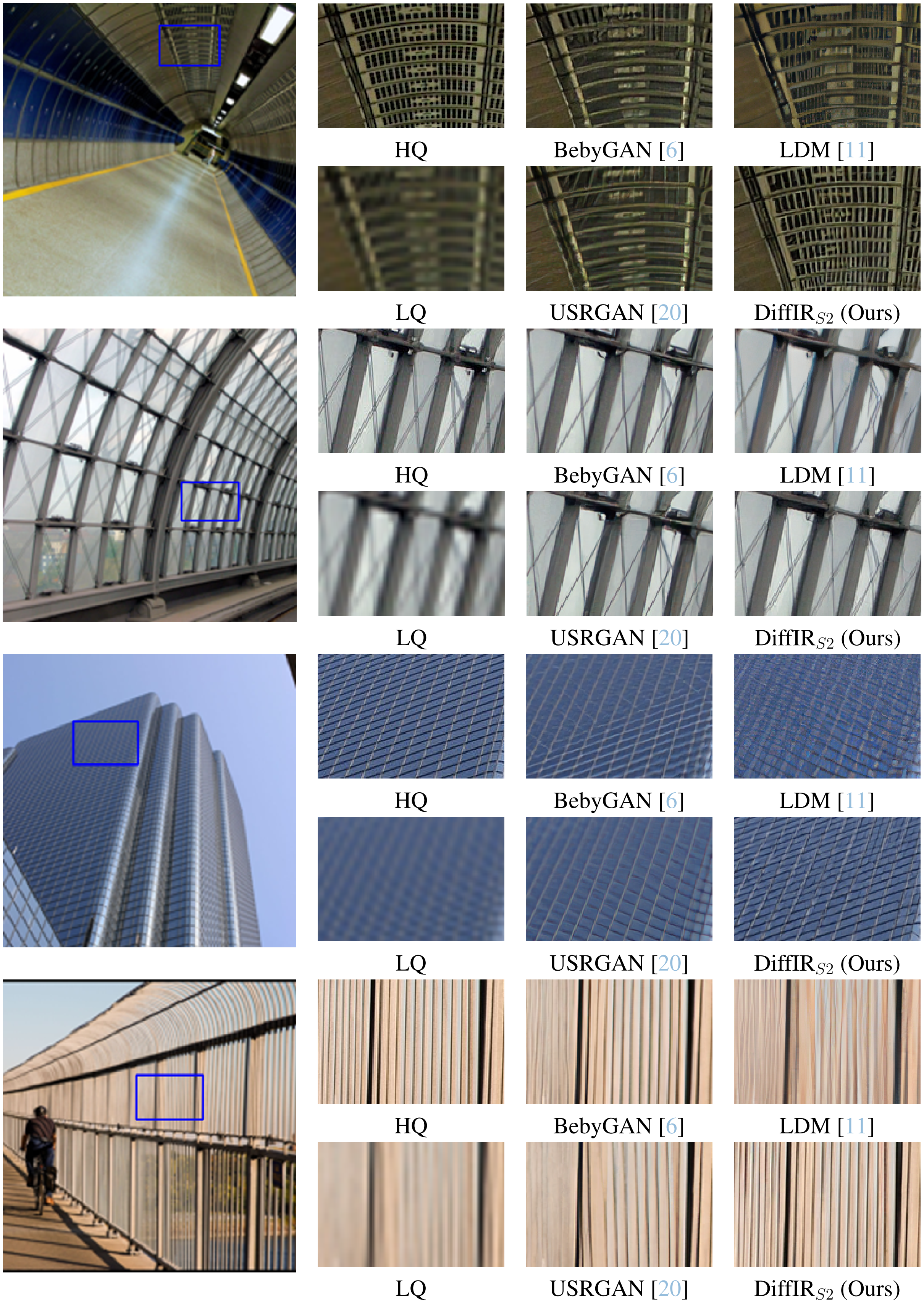


Figure 3. Visual comparison of 4× **image super-resolution** methods. Zoom-in for better details.

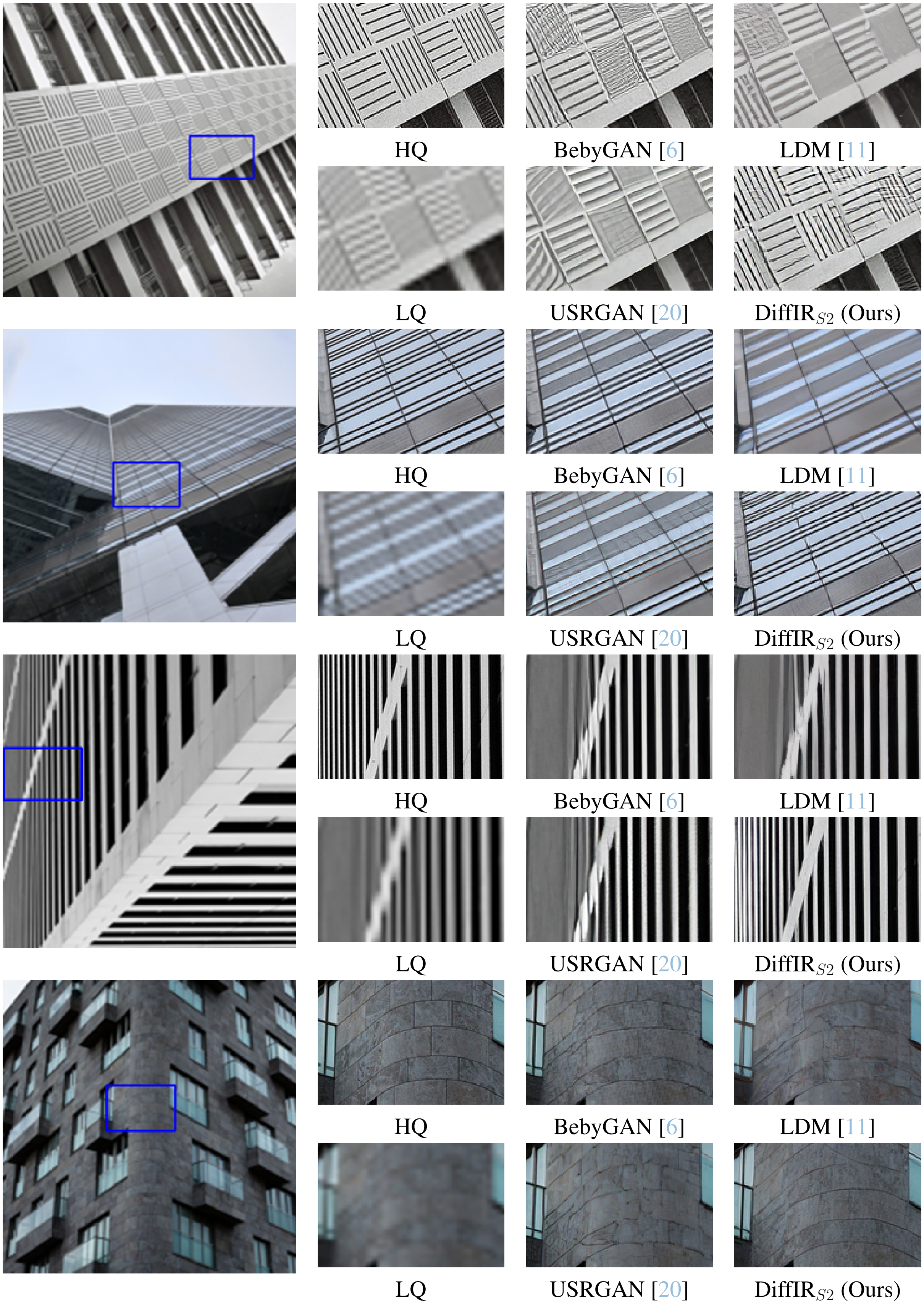


Figure 4. Visual comparison of $4\times$ **image super-resolution** methods. Zoom-in for better details.

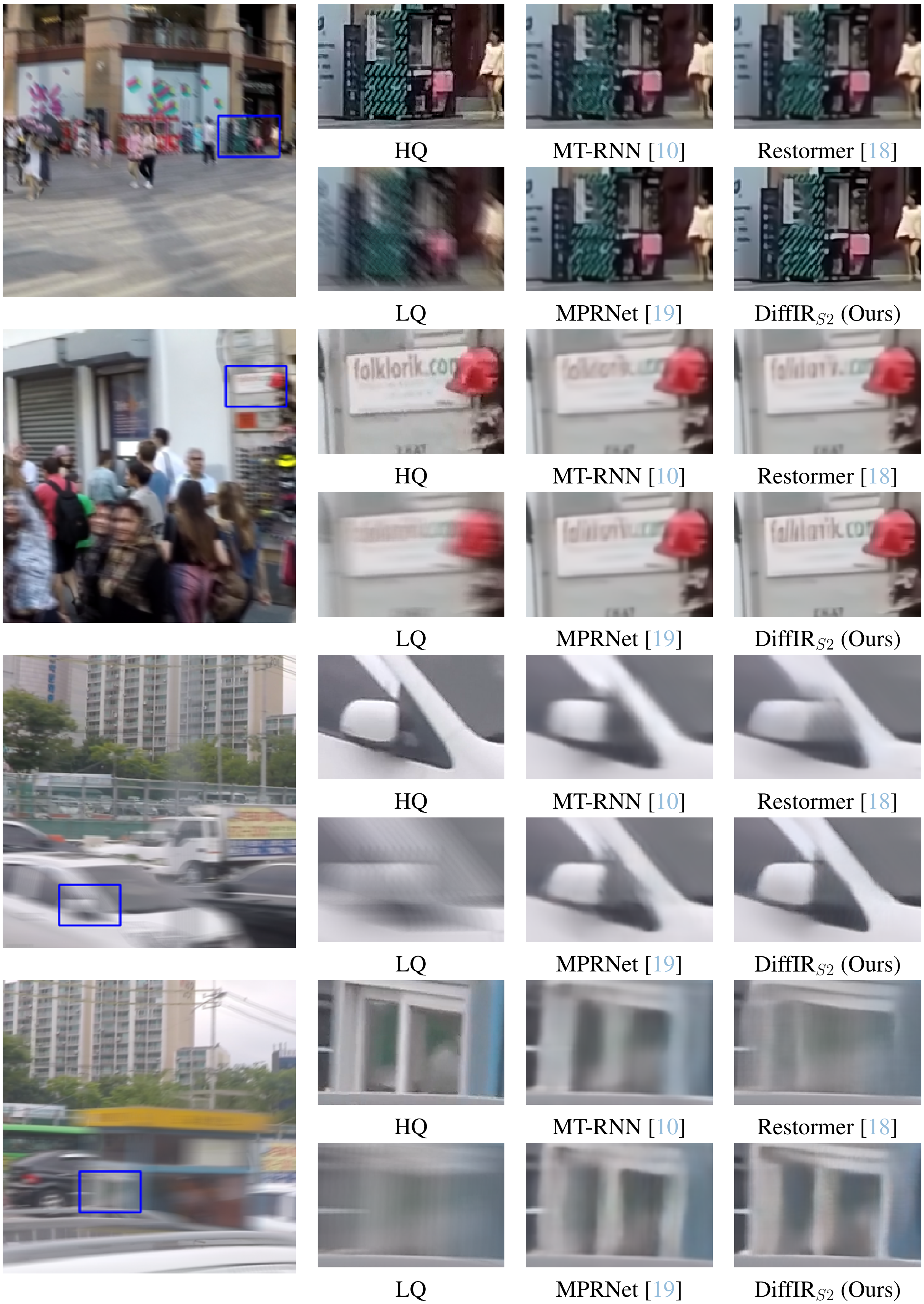


Figure 5. Visual comparison of **single image motion deblurring** methods. Zoom-in for better details.