FoundIR: Unleashing Million-scale Training Data to Advance Foundation Models for Image Restoration - Supplemental Material -

Hao Li[1*] Xiang Chen[1*] Jiangxin Dong[1] Jinhui Tang[2] Jinshan Pan[1†]
[1] Nanjing University of Science and Technology [2] Nanjing Forestry University

Overview

In this document, we first describe more details of the proposed million-scale dataset in Section 1. Next, we present more details of the proposed method in Section 2. Finally, we provide more qualitative comparisons in Section 3.

1. More Details on the Proposed Million-scale Dataset

Devices. Figure 1 shows the hardware and software used in our mechatronic shooting system, including an electronic slider (GVM Slider 120 cm), cameras, multiple tripods, and corresponding power supply equipment. Since the imaging process is closely related to the camera sensor, we employ different shooting devices to ensure data diversity, consisting of three cameras (*i.e.*, SONY ILCE-7M3, SONY DSC-RX10M4, and Canon EOS-R8) and three smartphones (*i.e.*, iPhone 15 Pro, VIVO X100 Pro, and HUAWEI Mate 60). To ensure stability throughout the entire shooting process, we use the GVM Slider app to control the electronic slider and the Imaging Edge Mobile app to control the camera shutter (see Figure 1(b)).

(a) The devices in our mechatronic system

(b) GVM Slider

Figure 1. The hardware and software of our mechatronic shooting system used for capturing paired data.

Data collection. We adjust internal camera settings (Round II) and external imaging conditions (Round III) to capture various degradations. Figure 3 shows some examples from our shooting locations. We spent over 6 months completing the large-scale data collection. Finally, we collect over 8,000 indoor and outdoor scenes, with the relevant statistics shown in Figure 5. Figure 6 summarizes the proportion of degradations captured in different rounds.

*Co-first authorship
†Corresponding author

Table 1. Statistics of training and testing set samples for different degradation types in the proposed million-scale dataset.

Degradation Type	**Training Set**	**Testing Set**	**Total Number**
Blur	109,480	150	109,630
Blur+Noise	29,950	50	30,000
Blur+JPEG Compression	29,940	50	29,990
Blur+Noise+JPEG Compression	29,950	50	30,000
Noise	58,015	100	58,115
JPEG Compression	59,950	50	60,000
Noise+JPEG Compression	29,950	50	30,000
Haze	79,800	200	80,000
Lowlight+Haze	79,800	100	79,900
Rain	39,900	100	40,000
Raindrop	44,828	100	44,928
Lowlight+Rain	40,111	50	40,161
Rain+Haze	79,950	50	79,800
Lowlight	39,962	50	40,012
Lowlight+Blur	85,893	100	85,993
Lowlight+Noise	52,995	50	53,045
Lowlight+JPEG Compression	29,950	50	30,000
Lowlight+Blur+Noise	29,950	50	30,000
Lowlight+Blur+JPEG Compression	29,940	50	29,990
Lowlight+Noise+JPEG Compression	29,950	50	30,000
Total Number	1,010,264	1,500	1,011,614

Table 2. Comparison with training data of existing universal image restoration methods.

Method	**Venue**	**Degradation Tasks**	**Numbers of Training Data**
AirNet [10]	CVPR 2022	Noise, Haze, Rain	77,479
TransWeather [20]	CVPR 2022	Rain, Raindrop, Snow	18,069
IDR [28]	CVPR 2023	Blur, Noise, Haze, Rain, Lowlight	80,067
PromptIR [18]	NIPS 2023	Noise, Haze, Rain	77,479
DiffUIR [31]	CVPR 2024	Blur, Haze, Rain, Lowlight, Snow	138,435
DA-CLIP [15]	ICLR 2024	Blur, Noise, JPEG, Haze, Rain, Raindrop, Lowlight, Snow, Shadow, Inpainting	52,801
InstructIR [4]	ECCV 2024	Blur, Noise, Haze, Rain, Lowlight	10,788
AutoDIR [8]	ECCV 2024	Blur, Noise, Haze, Rain, Raindrop, Super-resolution	114,742
FoundIR (Ours)	-	Blur, Noise, JPEG, Haze, Rain, Raindrop, Lowlight, Blur+Noise, Blur+JPEG, Noise+JPEG, Blur+Noise+JPEG, Rain+Haze, Lowlight+Haze, Lowlight+Rain, Lowlight+Blur, Lowlight+Noise, Lowlight+JPEG, Lowlight+Blur+Noise, Lowlight+Blur+JPEG, Lowlight+Noise+JPEG	1,011,614

Table 3. Data alignment and proportion of different resolution.

Metrics	Mean Flow Magnitude	Flow Standard Deviation	Outlier Ratio
w/o alignment	12.52 pixel	24.71 pixel	38.51%
w/ alignment	1.64 pixel	2.57 pixel	12.59%

(a) Effect of our alignment pipeline

Resolution	$\leq$ 1K	1K$\sim$2K	2K$\sim$4K	$\geq$ 4K
Proportion	0.75%	56.17%	21.96%	21.12%
Average	2514×1516			

(b) Proportion of different data resolution

Data alignment. We present several examples of reference objects as start-marker and end-marker in the data alignment pipeline, as shown in Figure 4. We manually select aligned GT-LQ frames from the uniform moving phase in a frame-wise manner once the start-marker disappears, continuing until the end-marker appears in the deceleration phase. To ensure alignment reliability, we manually inspect each image to exclude anomalous samples, particularly addressing the challenge of reference object identification under extremely dark conditions. To quantitatively evaluate the alignment quality, we calculate the optical flow between the captured GT and LQ data based on the Farneback algorithm (using *cv2.calcOpticalFlowFarneback* in OpenCV), which approximates motion for every pixel via polynomial expansion. Table 3(a) shows that using the alignment pipeline in Section 3.2 can effectively reduce pixel-level misalignment between GT and LQ data.

Data construction. Although we utilize a mechatronic shooting system to collect large-scale real-world paired data, synthetic data generation is employed for haze and JPEG compression degradations. This is because these degradations are difficult to capture as paired data by adjusting internal camera settings and external imaging conditions. Similar to [32], we synthesize hazy images based on the atmospheric scattering model [6]. Since the haze effect is closely related to scene depth, we use a foundation model, Depth Anything [25], to achieve better depth estimation. Following [22], the compressed images are randomly formulated using a quality factor $q \in [30, 90]$, where an image with a lower q has worse quality.

Dataset statistics. Using the proposed shooting system, we capture around 8,500 scenes in total, including 3,800 indoor scenes and 4,700 outdoor scenes. In Table 1, we present the statistics of training and testing set samples for different degradation types in the proposed million-scale dataset. Compared to existing training data (see Table 2), our dataset provides a larger training scale and a greater variety of degradation types for foundation models in image restoration. The average resolution of all images is 2514×1516, and the proportion of different resolutions is reported in Table 3(b).

Dataset samples. We present sample images in Figures 8-10, including 6 isolated and 12 coupled degradation types.

Limitation. Though the proposed unified data collection system can capture large-scale real-world training data for image restoration foundation models, it may not encompass all real-world degradation scenarios. Our future work will enhance the diversity and representativeness of the proposed dataset by incorporating additional real-world degradation conditions.

2. More Details on the Proposed Method

Loss function for generalist model. Inspired by [13, 31], we adopt the L_1 loss to drive the model for directly predicting the residual I_{res}. The training objective is defined as follows:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\left\| I_{res} - I^{\theta}_{res}\left(I_t, t\right)\right\|_1\right], \tag{1}$$

where I_t is the output in timestep t, and I_{res} denotes the residual components between LQ (I_{LQ}) and HQ (I_{HQ}) images.

Task-Incremental pool. In the task-incremental pool, we prioritize grouping degradations with similar attributes as task neighbors, such as Haze-Rain-Raindrop. The complete task-incremental sequence is consistent with that provided in Table 2. This setup offers several benefits compared to a random order: (1) By arranging tasks with similar attributes sequentially, the network can gradually adapt its understanding from one degradation type to another. This reduces the learning complexity and helps the model transfer knowledge effectively between related tasks. (2) Similar degradations share underlying patterns and feature representations. Thus, learning them in proximity allows the model to leverage shared information, improving the efficiency and stability of model training. (3) Since neighboring tasks have related features, the risk of forgetting previously learned tasks is minimized. This ensures smoother transitions and better retention of earlier knowledge.

Class-Incremental flow. We find that diverse coupled degradations naturally share some overlapping feature space distributions due to the interaction of multiple degradation effects, whereas diverse isolated degradations typically exhibit more distinct and dispersed distributions, as shown in Figure 7. Our class-incremental flow, which begins with isolated degradations and uses the learned information to guide the learning of coupled degradations, offers several advantages over learning both types simultaneously: (1) By starting with isolated degradations, the model can learn more distinct and dispersed feature space distributions. This way simplifies the initial learning process and avoids the complexities introduced by overlapping feature spaces in coupled degradations. (2) Leveraging the learned information from isolated degradations provides a structured foundation for tackling coupled degradations. This guidance reduces the difficulty of learning interactions between multiple degradation effects, as the model already has a solid understanding of individual degradation patterns. (3) Learning both isolated and coupled degradations simultaneously could lead to conflicting gradients and slower convergence due to overlapping feature spaces. The incremental approach avoids this by sequentially addressing simpler tasks (isolated degradations) before moving on to the more complex coupled degradations.

Expert pool. Our expert pool is highly extensible, allowing researchers to easily integrate more specialist models tailored to specific restoration tasks. For example, researchers can add experts specialized in tasks such as face image restoration, underwater image restoration, or other domain-specific challenges. This adaptability ensures that the framework can rapidly adjust to new and emerging restoration needs, improving both the efficiency and accuracy of the model for a wide range of real-world applications. By incorporating both generalist and specialist models, our approach not only improves restoration quality but also offers the flexibility to handle a variety of degradation types.

Testing pipeline of FoundIR. The proposed FoundIR consists of a generalist model and multiple specialist models, obtaining high-quality restored images in various real-world scenarios. Specifically, the generalist model allows users to restore images with unknown degradations. Furthermore, to meet the users' specific needs in challenging and complex inputs, multiple specialist models are provided to refine the generalist model's results in parallel. Users have the flexibility to either directly use the result of the generalist model or the refined results obtained by different specialist models.

3. More Experimental Results

Quantitative comparisons of perceptual metrics. Table 4 shows the quantitative comparisons of perceptual metrics, including LPIPS [29], NIQE [17], NIMA [19], MUSIQ [9], CLIPIQA [21], MANIQA [26], and FID [7] on the proposed benchmark. The proposed FoundIR achieves the best performance compared with recent state-of-the-art methods [3, 10, 15, 18, 24, 31].

Table 4. Quantitative comparisons of perceptual metrics on the proposed benchmark.

Method	AirNet [10]	PromptIR [18]	DiffIR [24]	DiffUIR [31]	DA-CLIP [15]	X-Restormer [3]	FoundIR
LPIPS ↓ [29]	0.5250	0.3872	0.3222	0.2903	0.4600	0.4061	**0.2693**
NIQE ↓ [17]	7.6521	6.7485	4.9767	4.9472	7.1613	6.6968	**4.8761**
NIMA ↑ [19]	4.4301	4.4228	4.4668	4.7531	4.4594	4.4131	**4.7673**
MUSIQ ↑ [9]	29.4592	34.6049	41.7078	50.6280	31.7528	33.7725	**50.9194**
CLIPIQA ↑ [21]	0.3407	0.3348	0.3259	0.3777	0.3601	0.3245	**0.3953**
MANIQA ↑ [26]	0.4614	0.4589	0.4919	0.5565	0.4570	0.4617	**0.5580**
FID ↓ [7]	56.2712	31.0074	29.7050	15.2832	40.5125	36.5438	**14.7455**

Model efficiency. Table 5 shows the comparisons of model efficiency (*i.e.*, Parameters, FLOPs, and Testing time) with other SOTA universal restoration methods [14, 15, 24], where the testing time is averaged on 100 images (with $1080 \times 1920 \times 3$ pixels), evaluated on a machine with a NVIDIA GeForce RTX 4090 GPU, using PyTorch 2.0.

Table 5. Comparison of model efficiency with recent methods.

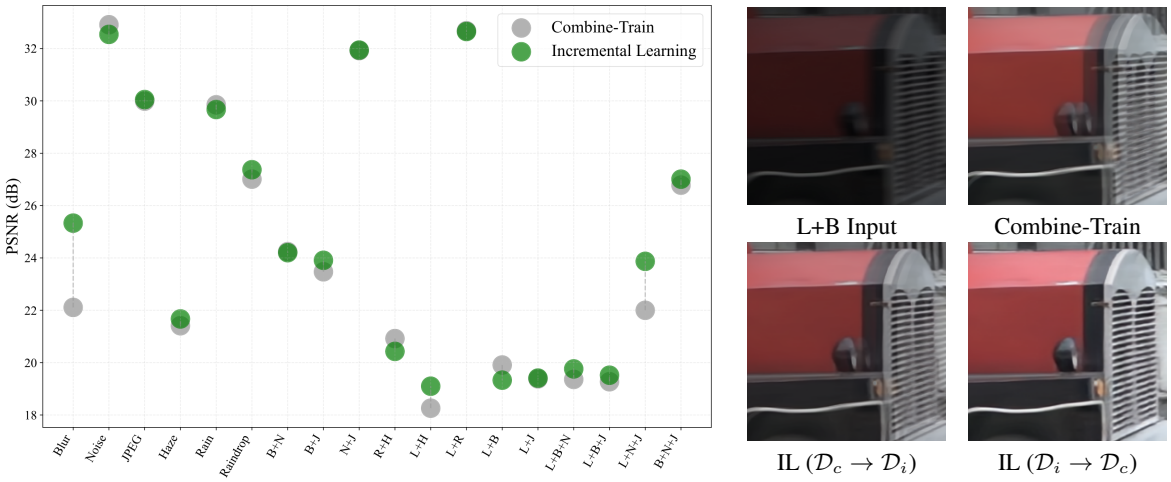
Method	DiffIR [24]	IR-SDE [14]	DA-CLIP [15]	FoundIR
Parameters (M)	35.59	36.22	174.10	36.30
FLOPs (G)	499	3610	3674	310
Testing time (s)	21.37	132.78	145.22	10.88

Extend FoundIR to other restoration task. Since our goal is to develop a general image restoration method, we explore its potential for extending to other restoration tasks. To achieve this, we extend our FoundIR to the real-world image super-resolution (SR) task by incorporating some SR specialist models [2, 23, 27]. Specifically, we first use the generalist model to restore the low-resolution (LR) input, and then fine-tune the SR specialist models to obtain the high-resolution (HR) output. Table 6 shows that by integrating the generalist model of our FoundIR, existing methods [2, 23, 27] can achieve significant performance improvement on real-world SR ($\times 4$) benchmark [1].

Table 6. Quantitative comparisons of no-reference metrics (*i.e.*, MUSIQ [9], CLIPIQA [21]) on the real-world SR ($\times 4$) benchmark [1]. Where $\mathcal{G}$ and $\mathcal{S}$ denote the generalist model and specialist model, respectively.

Method	ResShift [27]	FoundIR ($\mathcal{G}$) + ResShift ($\mathcal{S}$)	OSEDiff [23]	FoundIR ($\mathcal{G}$) + OSEDiff ($\mathcal{S}$)	FaithDiff [2]	FoundIR ($\mathcal{G}$) + FaithDiff ($\mathcal{S}$)
MUSIQ ↑	31.46	35.82	32.62	40.93	31.41	37.54
CLIPIQA ↑	0.4772	0.4841	0.6613	0.6817	0.5422	0.5621

Impact of incremental learning strategy. Beyond the 0.30dB PSNR improvement averaged across all degradation types reported in Table 4(a) of the main paper, Figure 2(a) shows detailed comparisons on each degradation type, where our incremental learning strategy outperforms the commonly used 'Combine-Train' on a larger number of degradation categories. Moreover, we compare performing various training strategies (*i.e.*, Combine-Train, IL ($\mathcal{D}_c \rightarrow \mathcal{D}_i$), and IL ($\mathcal{D}_i \rightarrow \mathcal{D}_c$)) in Figure 2(b). Our training strategy (IL ($\mathcal{D}_i \rightarrow \mathcal{D}_c$)) achieves a clearer result on a challenging example with coupled degradations (*i.e.*, lowlight+blur).



(a) Quantitative comparisons on different degradation type.

(b) Visual comparisons.

Figure 2. Effect of incremental learning strategy.

Qualitative comparisons on the proposed dataset. To demonstrate the effectiveness and generalization of the proposed method, Figures 11-32 present qualitative comparisons with state-of-the-art methods [3, 8, 10, 15, 18, 24, 31] across all degradation types in the proposed million-scale dataset. The proposed FoundIR can handle various degradations and generate much clearer images with finer details and structures.

Qualitative comparisons on public benchmarks. We conduct additional qualitative comparisons with recent methods [8, 10, 18, 20, 24, 31] on public benchmarks [5, 11, 12, 30] to evaluate the generalization ability of the proposed training data. Note that 'DiffUIR-Official' and 'DiffUIR-Our data' represent the official pre-trained model and the model retrained on the proposed dataset, respectively. Figures 33-38 show that 'DiffUIR-Our data' restores better results compared to 'DiffUIR-Official', while the proposed FoundIR is able to effectively handle out-of-distribution data.

Qualitative comparisons in ablation studies. Figure 39 shows that the models trained on existing public datasets (see Figures 39(b)-(c)) struggle to address coupled degradation (*e.g.*, Lowlight+Haze) effectively. While Figure 40 presents that these models fail to remove the complex real-world rain streaks. In contrast, the models trained on the proposed training data can handle different real-world complex degradations, and the quality of the restored image improves progressively as the scale of the training data increases (see Figures 39(d)-(f)).

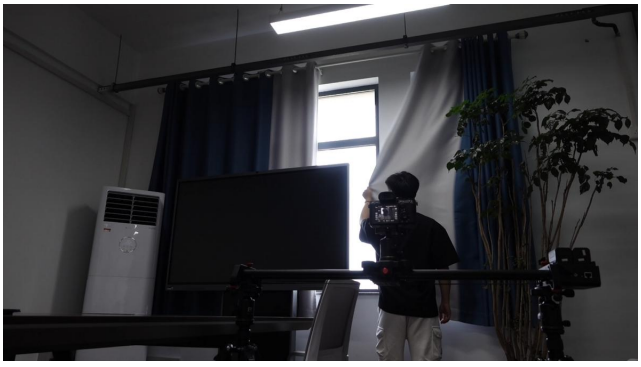
Furthermore, to demonstrate the effectiveness of our ensemble framework, we provide qualitative comparisons in Figure 41. It demonstrates that our method produces clearer results, particularly excelling at handling coupled degradations.

(a) GT capture (Round I)

(b) Adjust camera settings (Round II)

(c) Record object movements (Round II)

(d) Record pedestrian movements (Round II)



(e) Close the curtains (Round III)

(f) Turning off the lights (Round III)

(g) Use electric sprinklers to generate rain (Round III)

(h) Artificial light source (Round III)

Figure 3. Examples of data collection in different rounds. We adjust internal camera settings (Round II) and external imaging conditions (Round III) to capture various degradation.

(a) Start-marker (b) End-marker

(c) Markers in data collection scenarios (d) Markers in data collection scenarios

Figure 4. Examples of placing reference objects in the data alignment pipeline. To mitigate misalignment between each GT-LQ frames, we place recognizable reference objects as start-marker and end-marker before and after the uniform moving phase.

Figure 5. Word cloud statistics derived from real-world data collection scenarios. It can be observed that our million-scale dataset encompasses a wide variety of rich indoor and outdoor scenes.

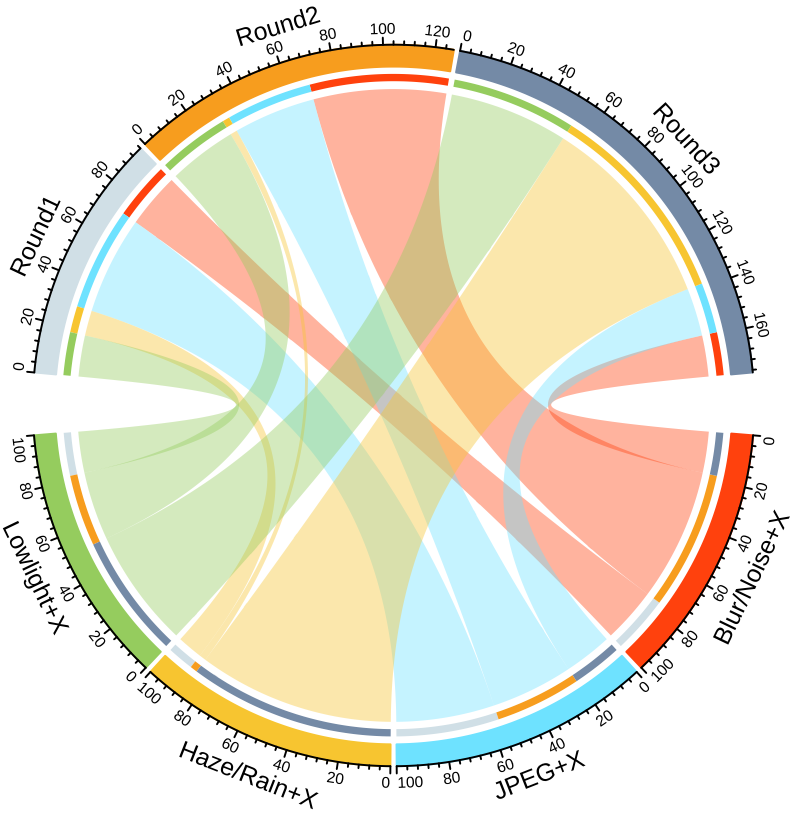
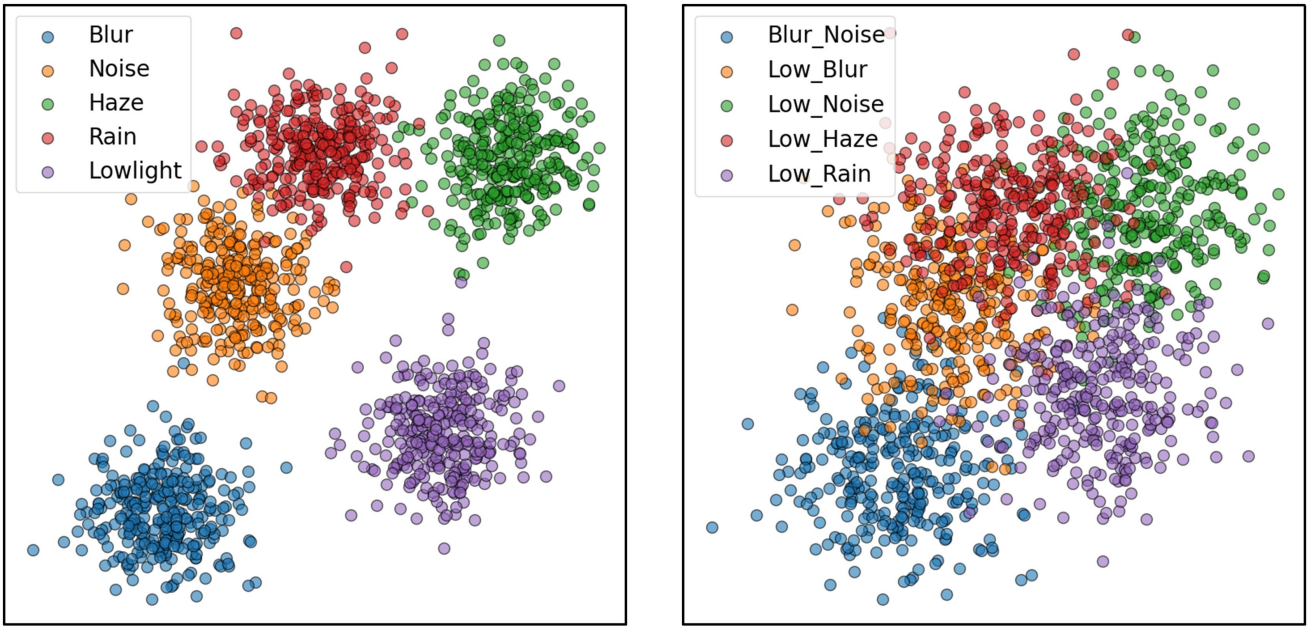


Figure 6. Statistics of degradation captured by multi-round data collection using our system. The upper semicircle represents the shooting round, while the lower semicircle indicates the types of degradations captured.



(a) Isolated degradation

(b) Coupled degradation

Figure 7. Distribution of different degradations visualized by t-SNE [16].

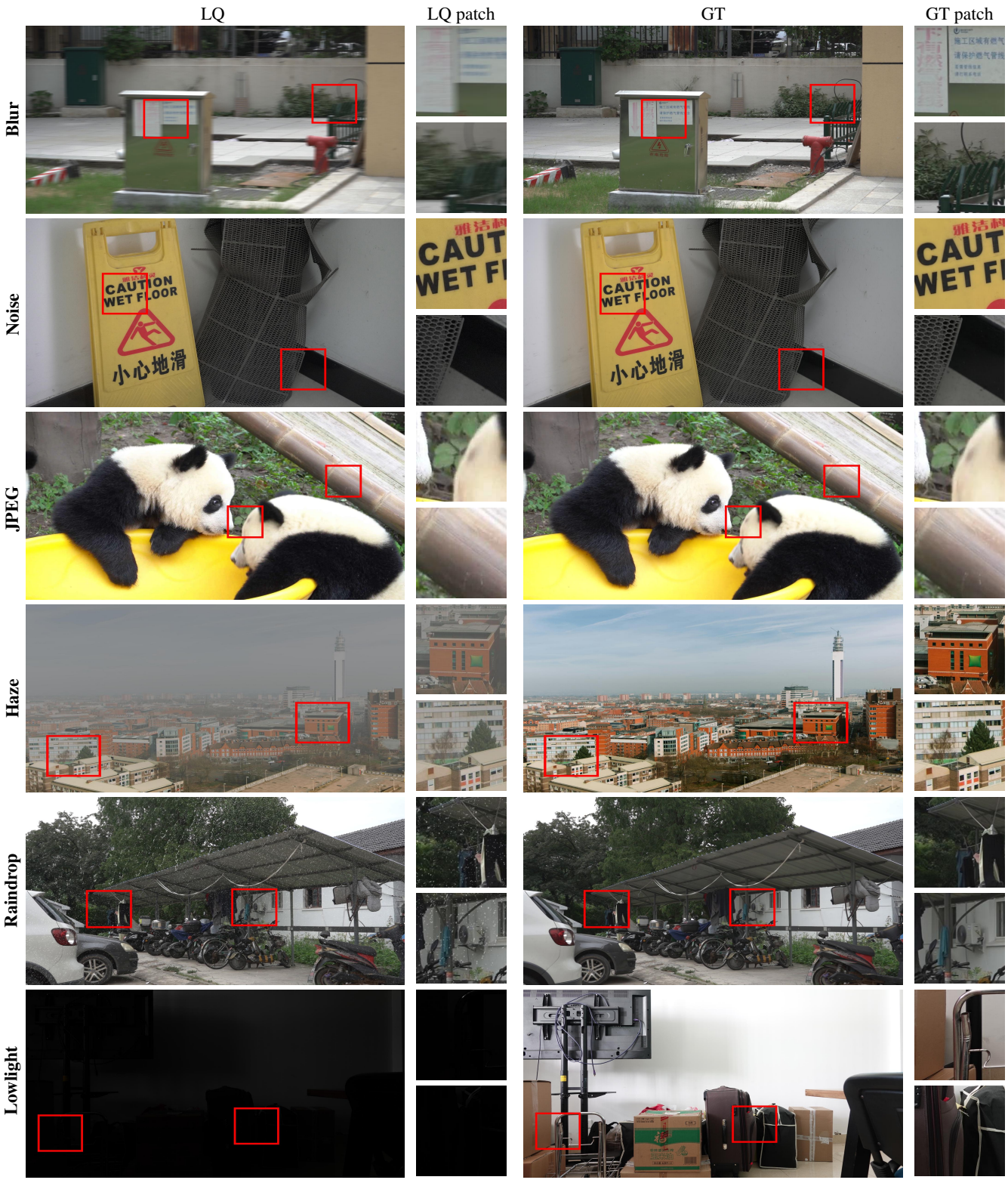


Figure 8. Example LQ-GT paired images in the proposed million-scale dataset. Compared to existing training data, the proposed dataset offer twofold advantages: (i) **larger-scale real-world samples**, and (ii) **higher-diversity data types**. Zoom in for a better view.

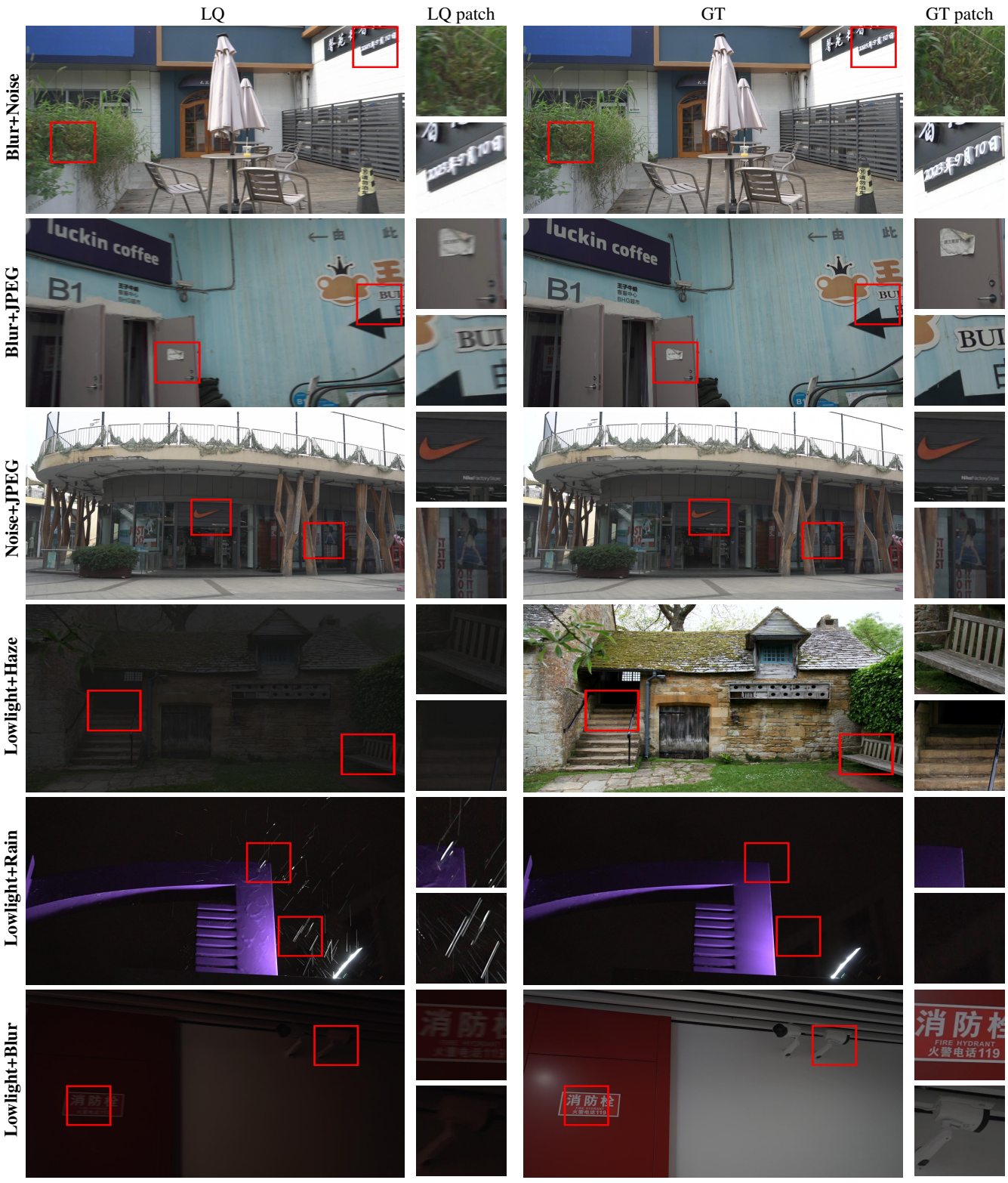


Figure 9. Example LQ-GT paired images in the proposed million-scale dataset. Compared to existing training data, the proposed dataset offer twofold advantages: (i) **larger-scale real-world samples**, and (ii) **higher-diversity data types**. Zoom in for a better view.

Figure 10. Example LQ-GT paired images in the proposed million-scale dataset. Compared to existing training data, the proposed dataset offer twofold advantages: (i) **larger-scale real-world samples**, and (ii) **higher-diversity data types**. Zoom in for a better view.

Figure 11. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant blur effects, our approach generates a clearer image. Zoom in for a better view.

Figure 12. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant blur effects, our approach generates a clearer image. Zoom in for a better view.

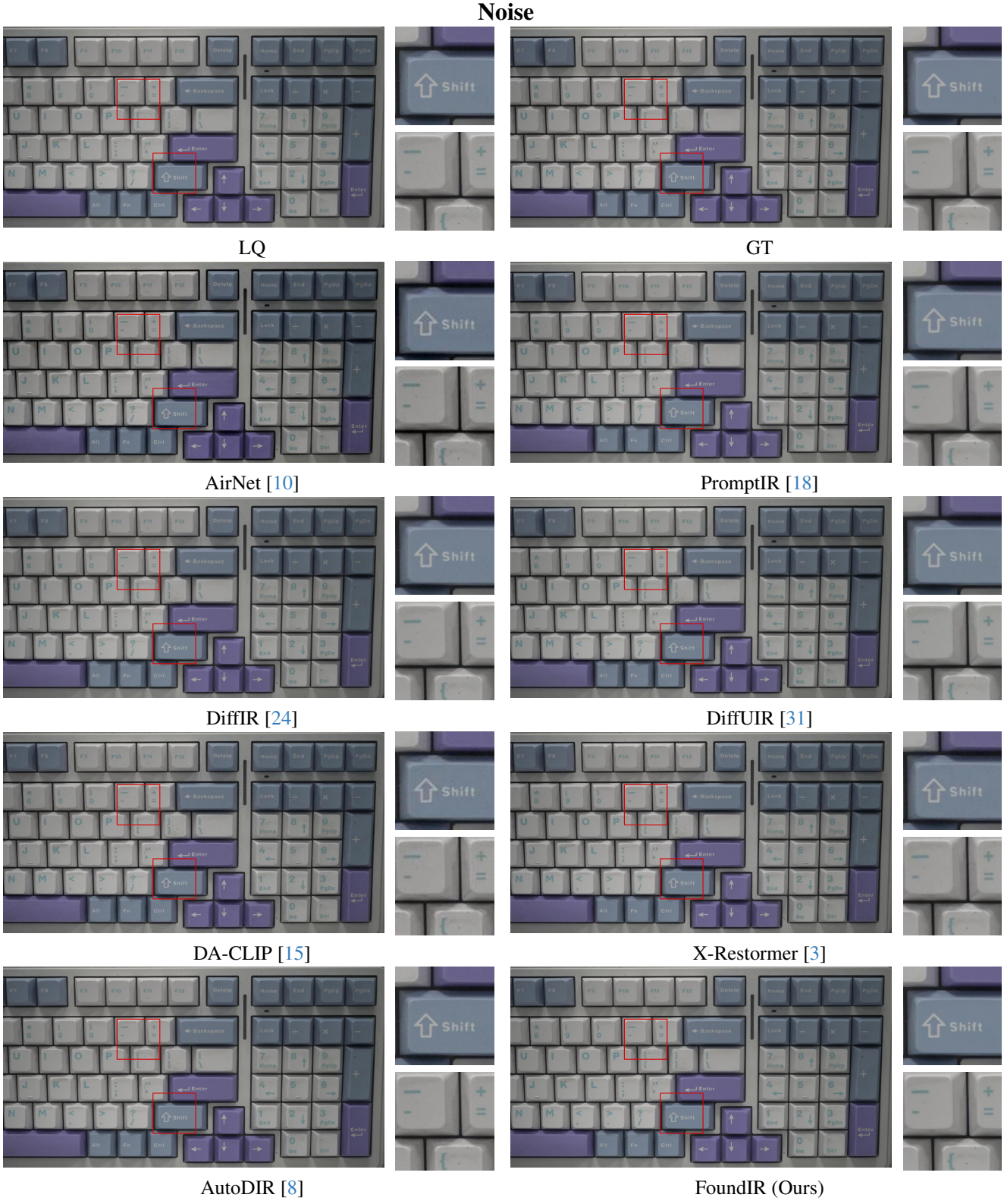


Figure 13. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant noise effects, our approach generates a clearer image. Zoom in for a better view.

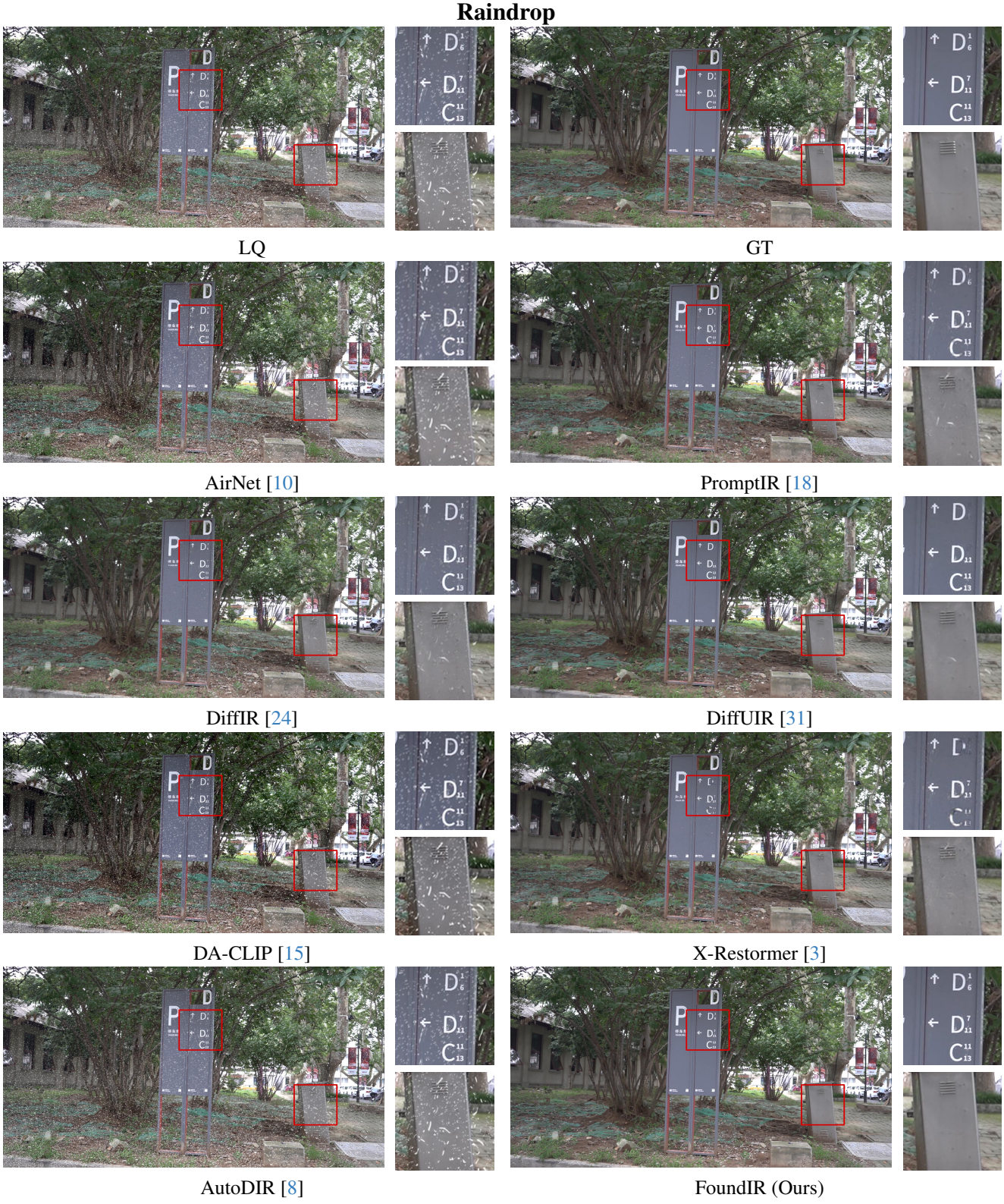


Figure 14. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant raindrops, our approach generates a clearer image. Zoom in for a better view.

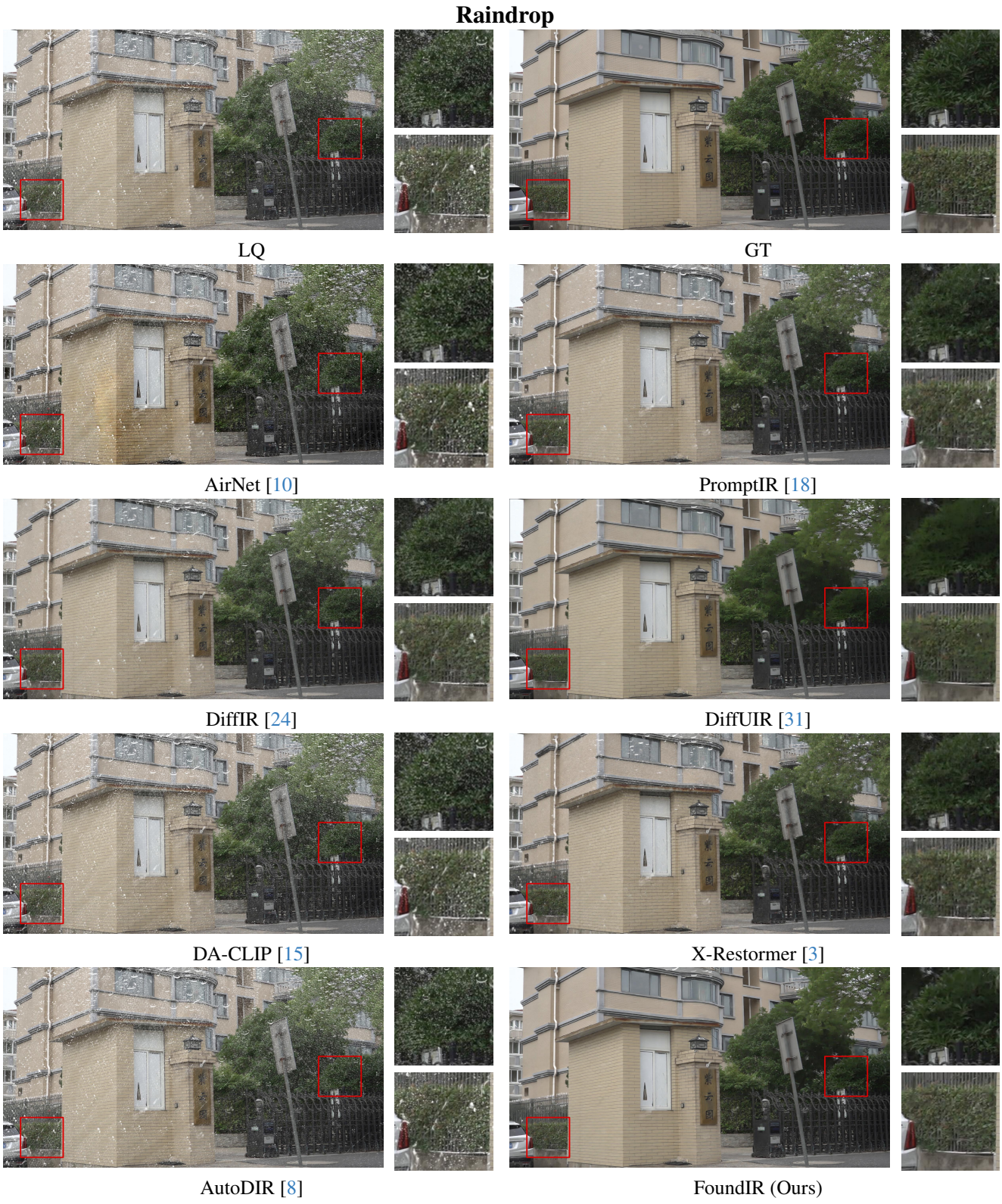


Figure 15. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant raindrops, our approach generates a clearer image. Zoom in for a better view.

Figure 16. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], which still contain significant raindrops, our approach generates a clearer image. Zoom in for a better view.

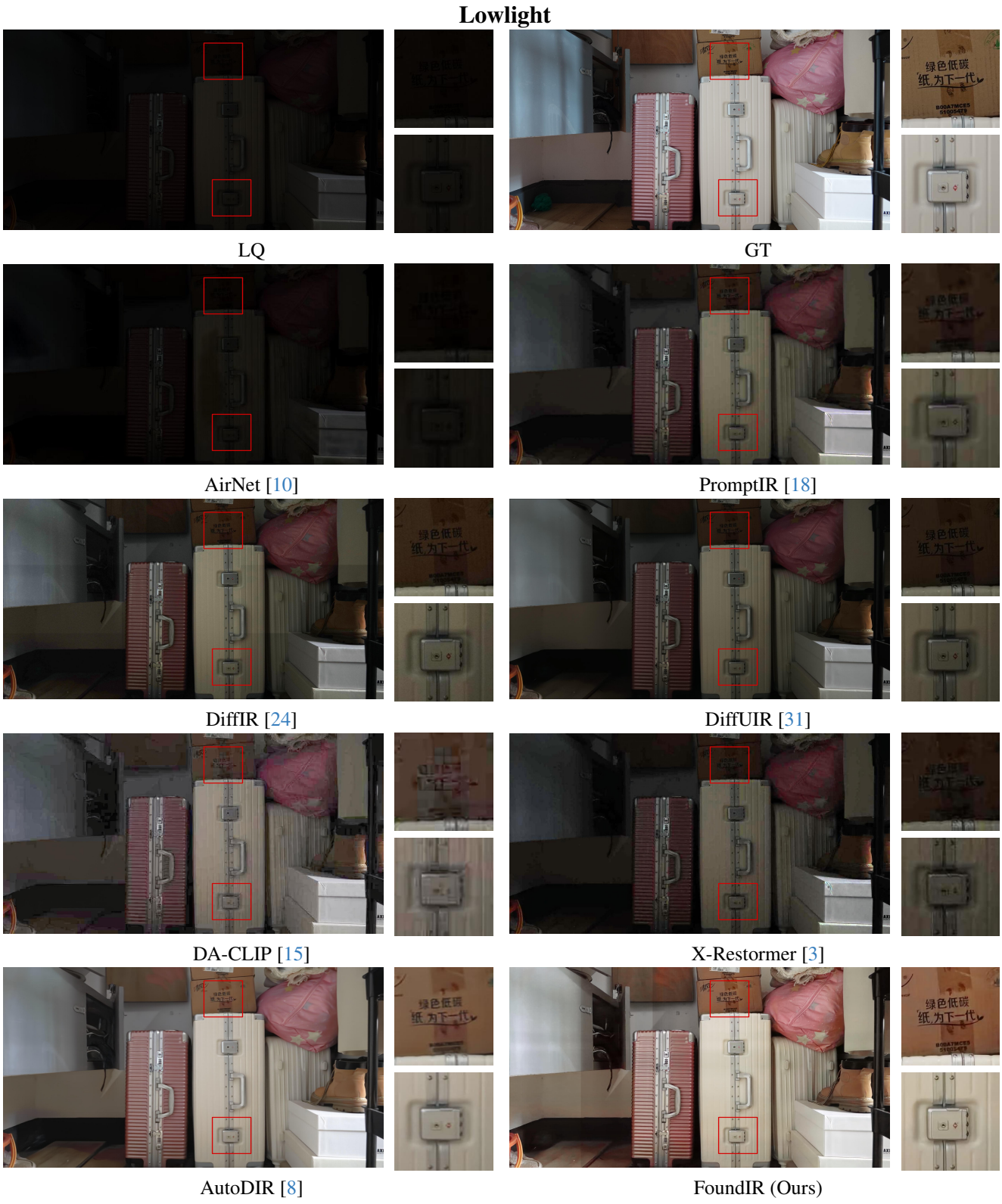


Figure 17. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

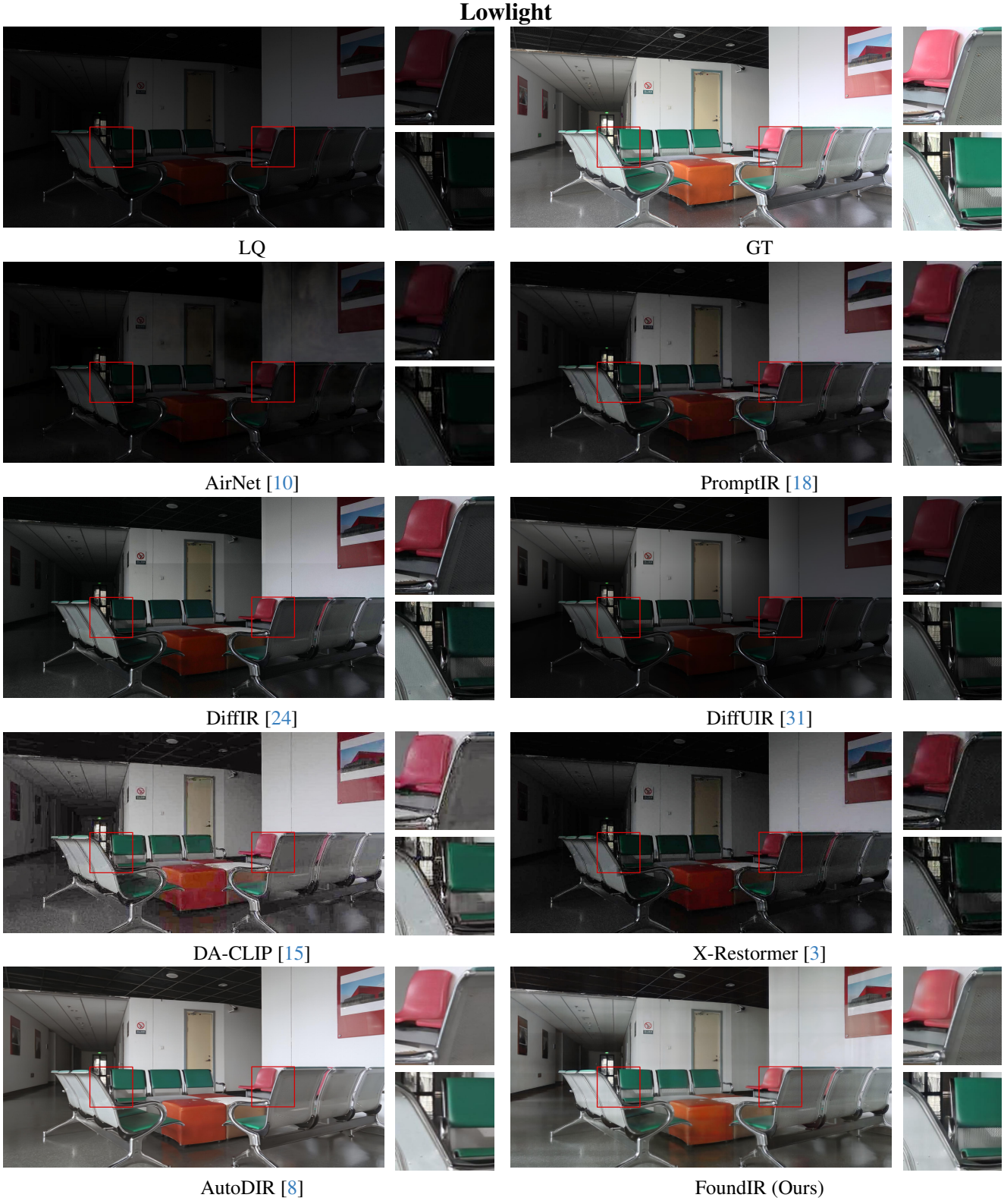


Figure 18. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

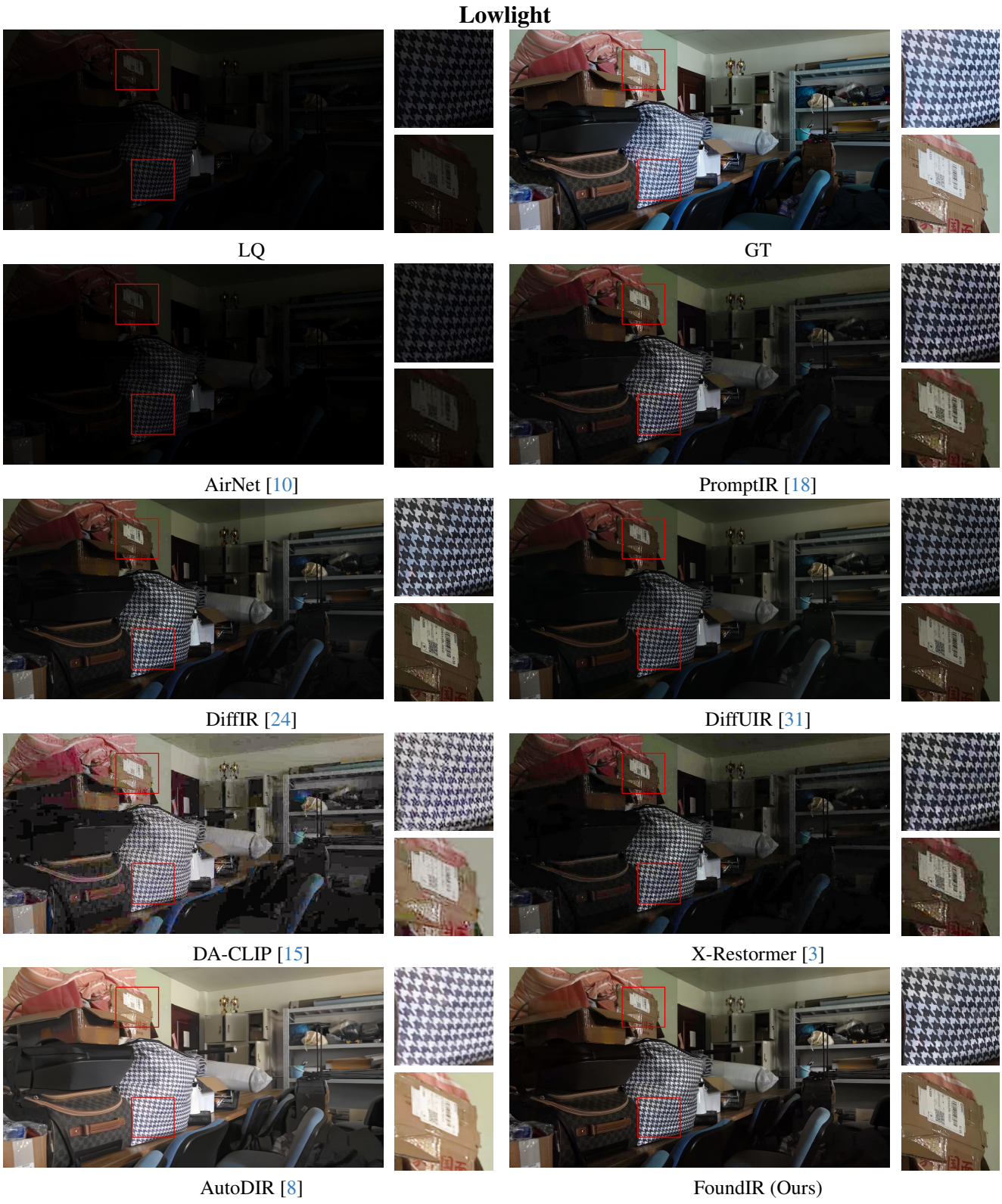


Figure 19. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

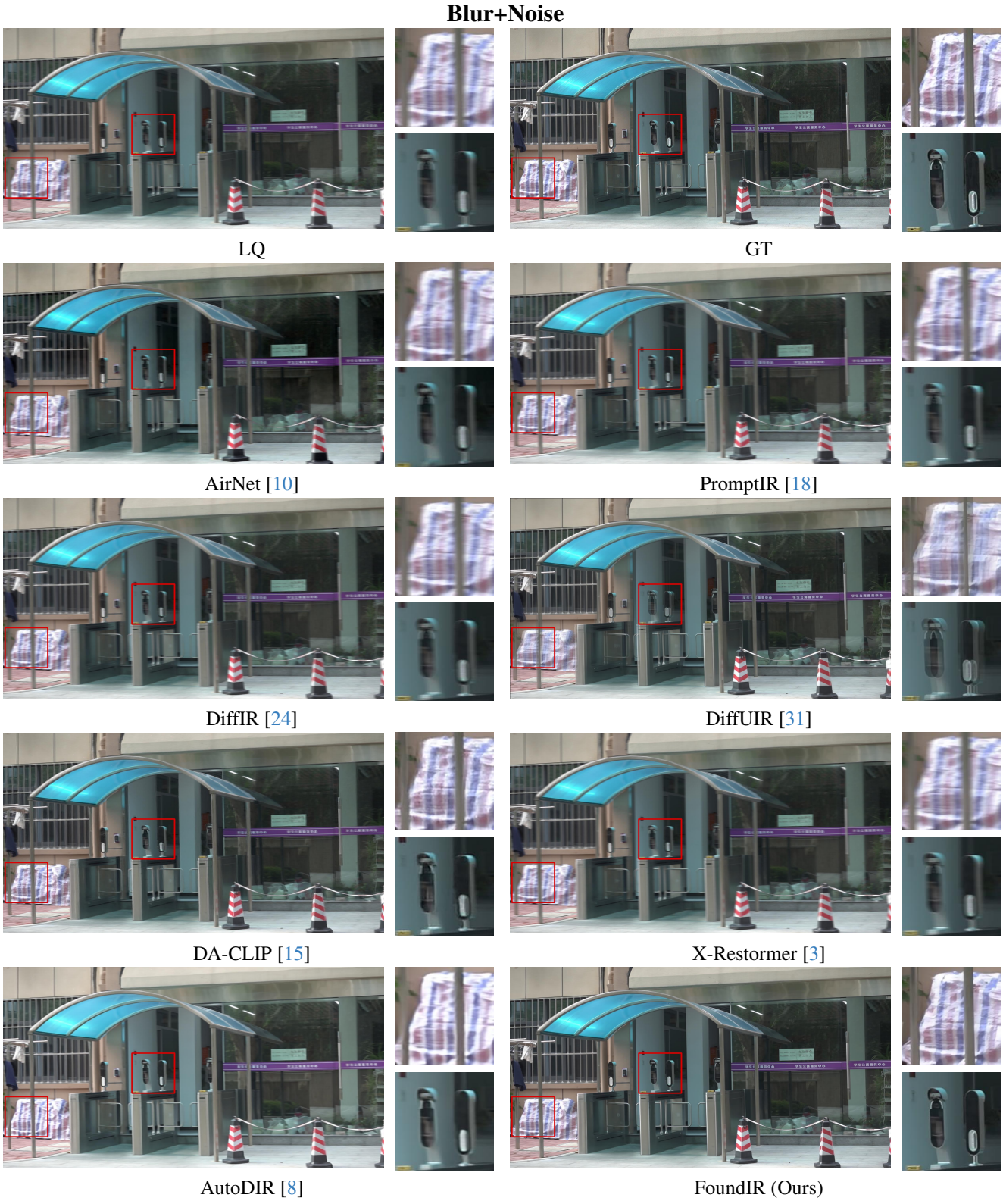


Figure 20. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 21. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 22. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

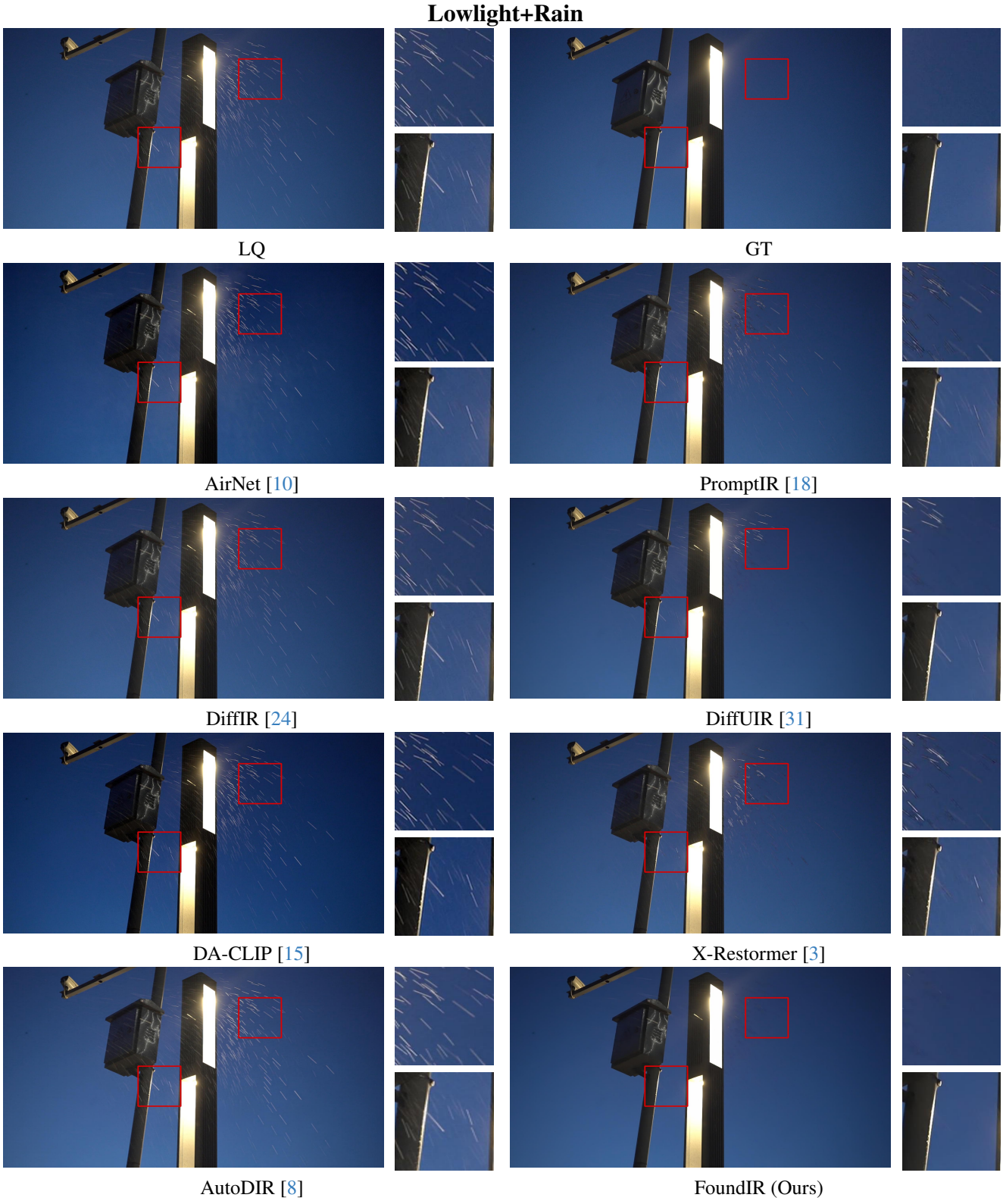


Figure 23. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

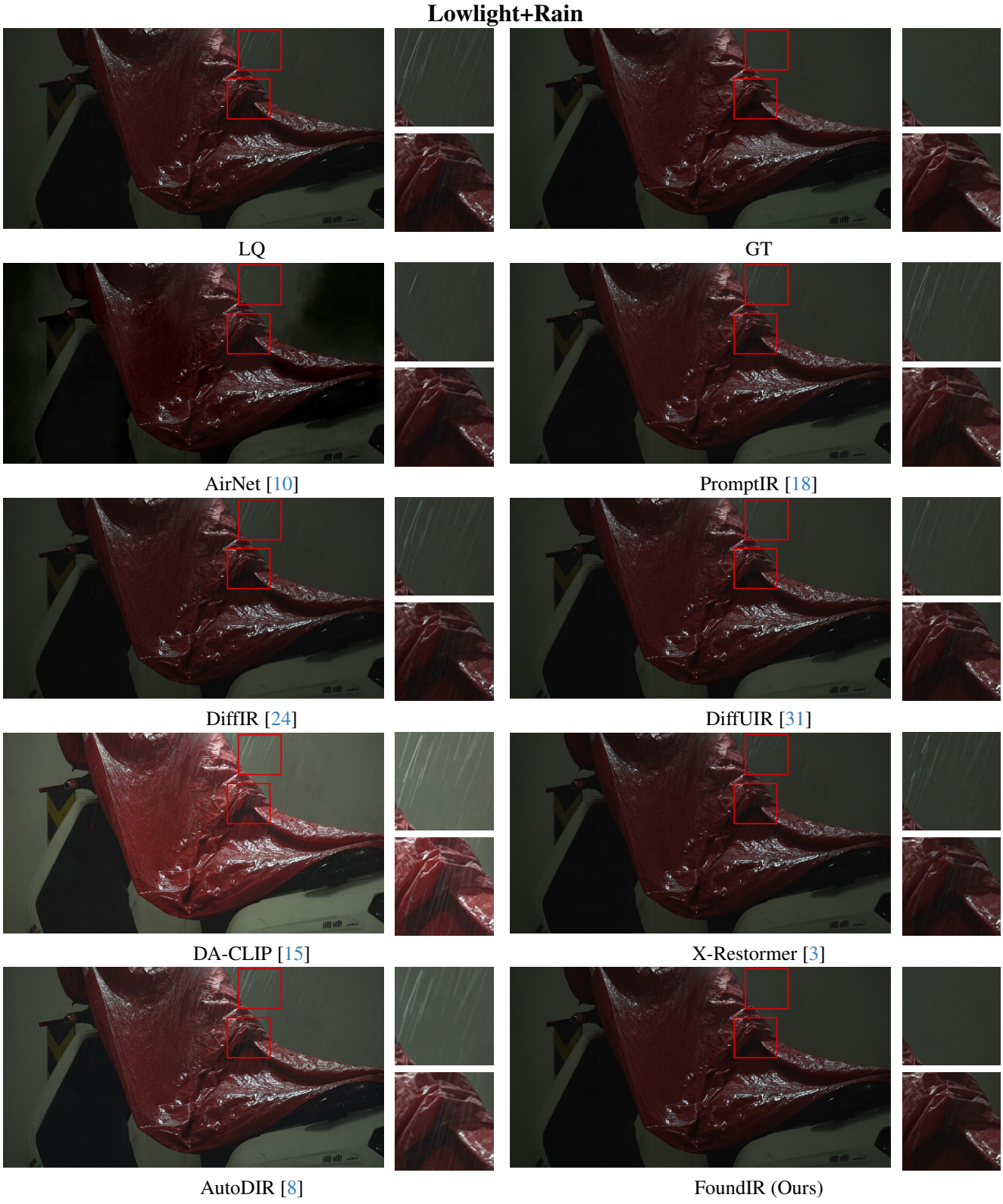


Figure 24. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 25. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 26. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 27. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

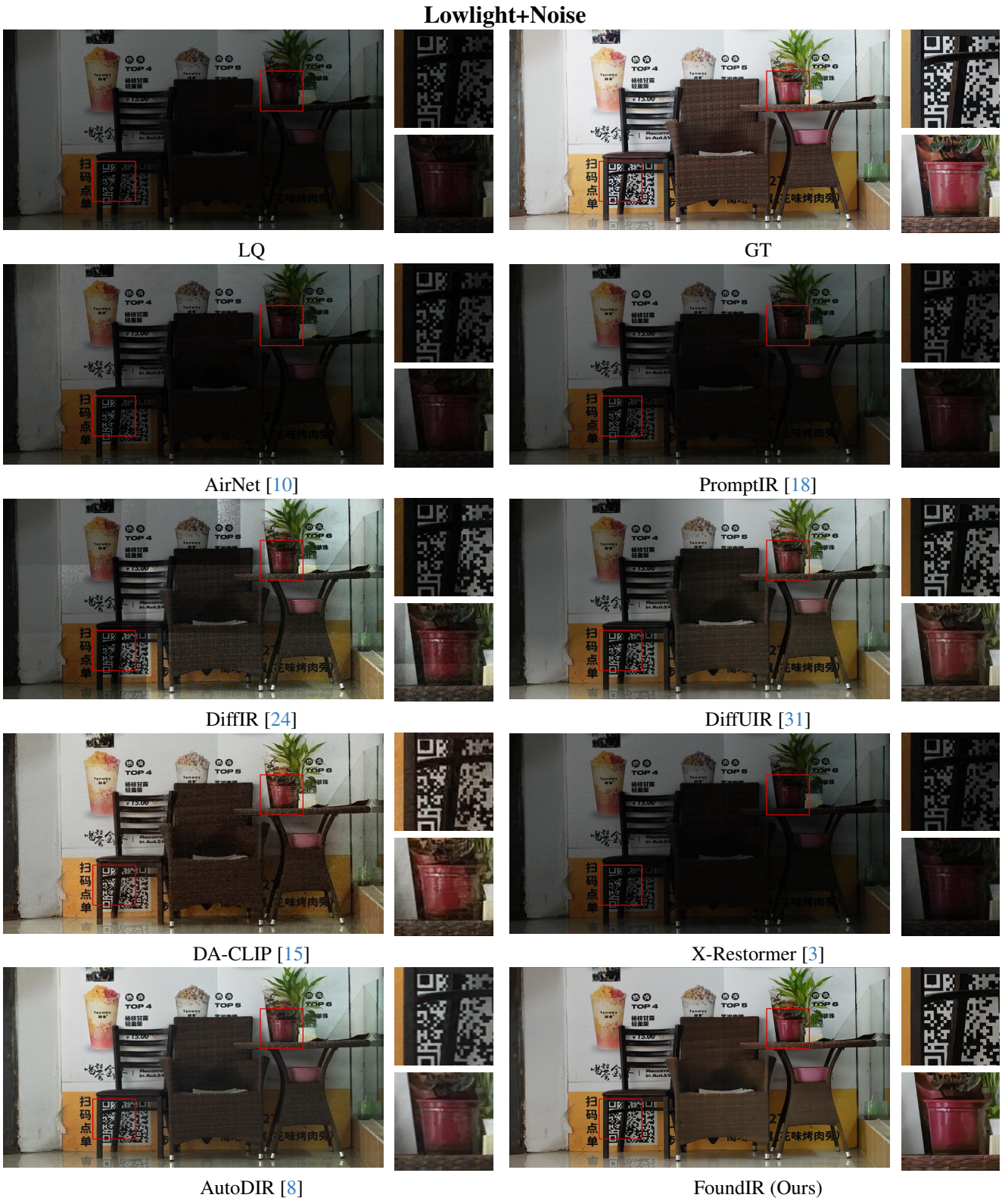


Figure 28. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 29. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 30. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

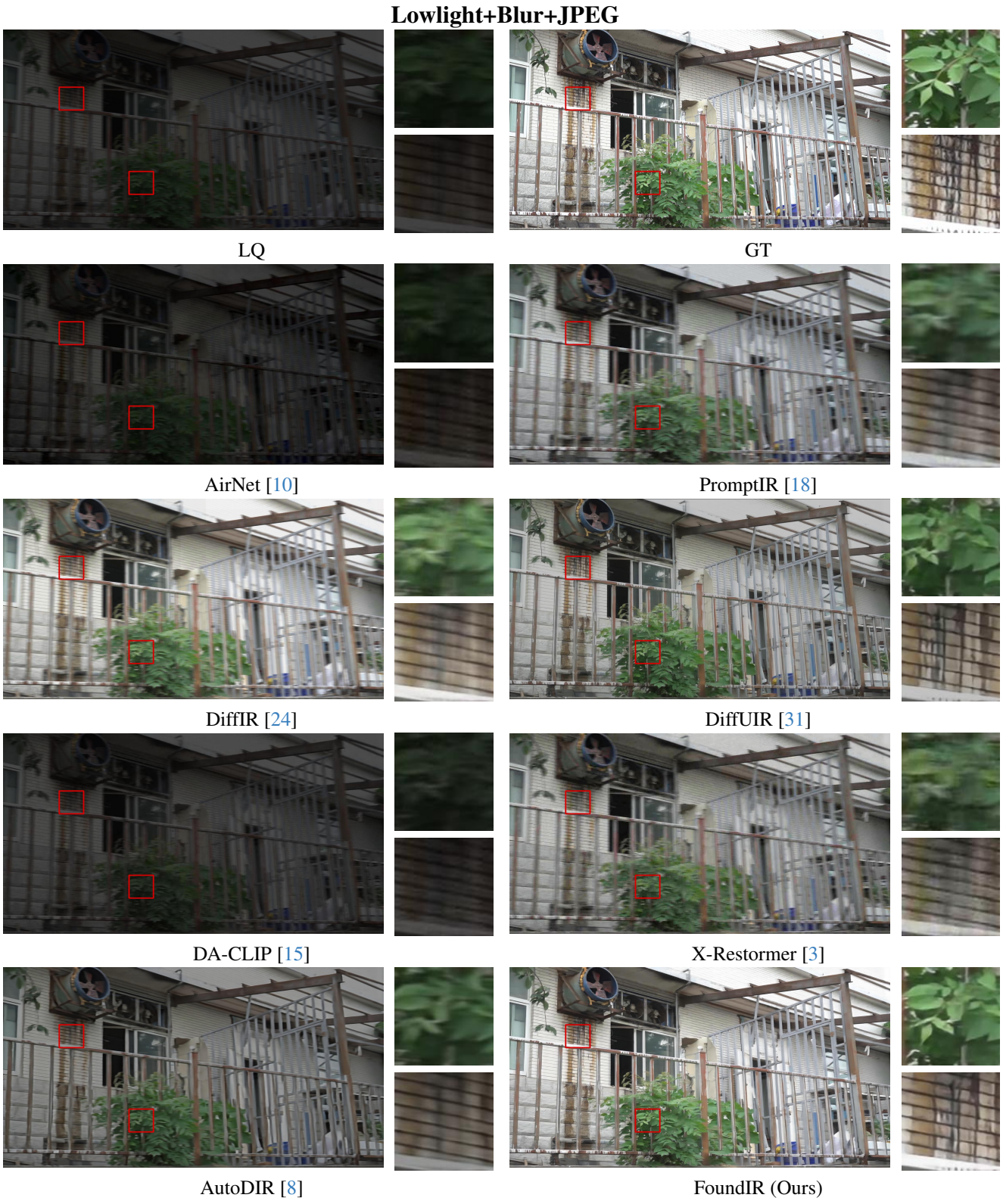


Figure 31. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 32. Visual comparison results. Compared to the results restored by existing methods [3, 8, 10, 15, 18, 24, 31], our approach generates a clearer image. Zoom in for a better view.

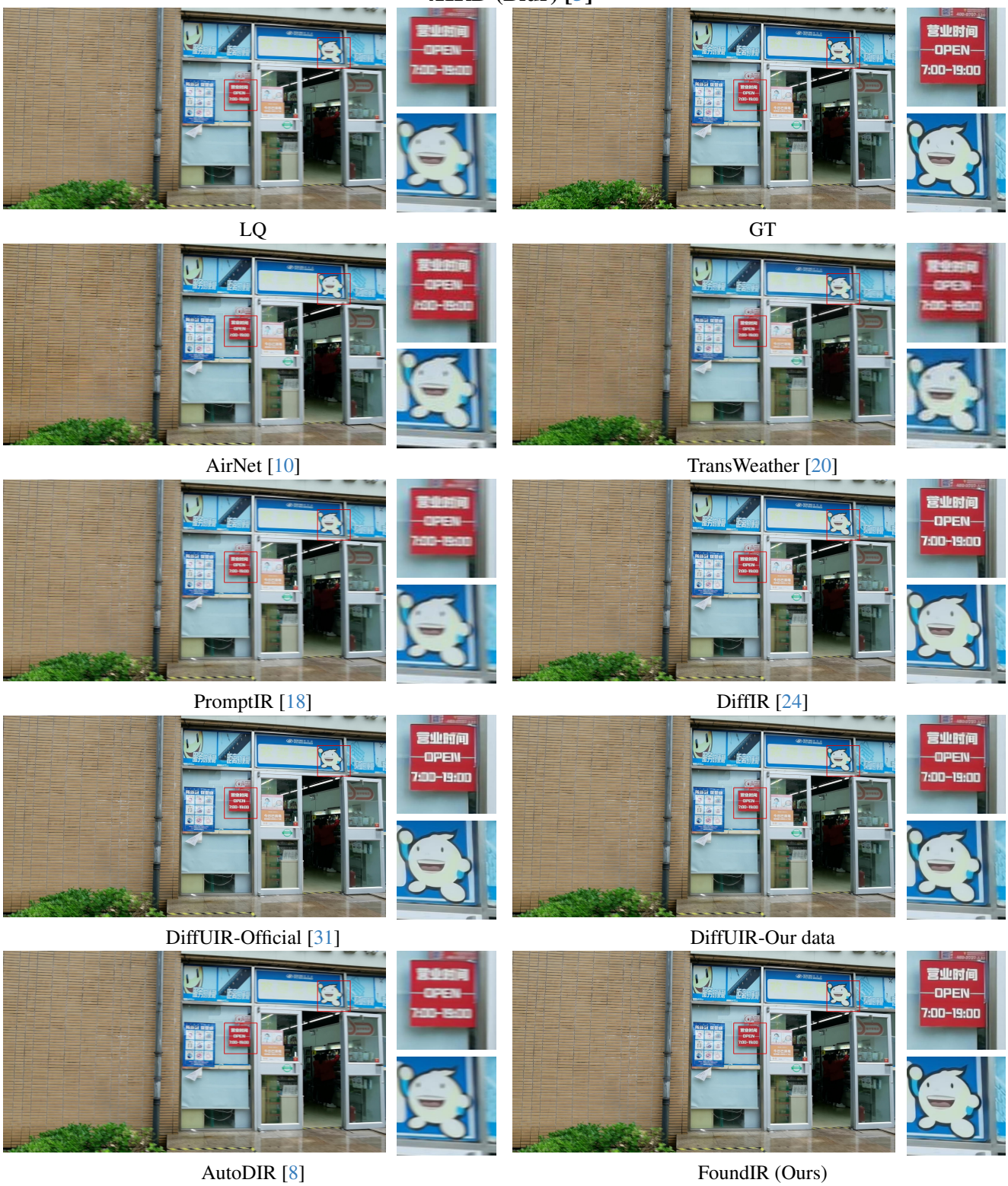


Figure 33. Visual comparison results on the public benchmark - 4KRD [5]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

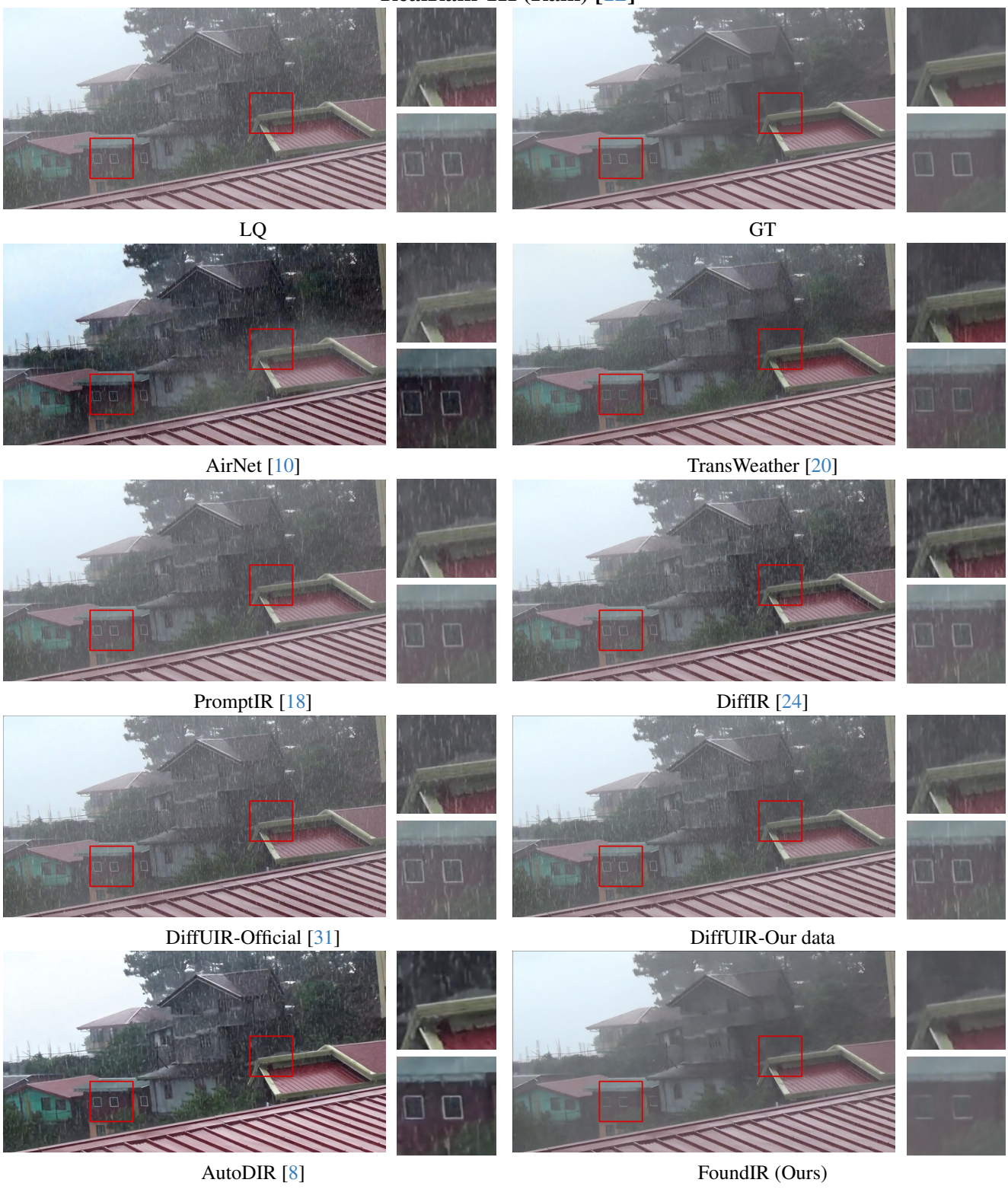


Figure 34. Visual comparison results on the public benchmark - RealRain-1K [12]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 35. Visual comparison results on the public benchmark - RealRain-1K [12]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 36. Visual comparison results on the public benchmark - HazeRD [30]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

Figure 37. Visual comparison results on the public benchmark - UHD-LL [11]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

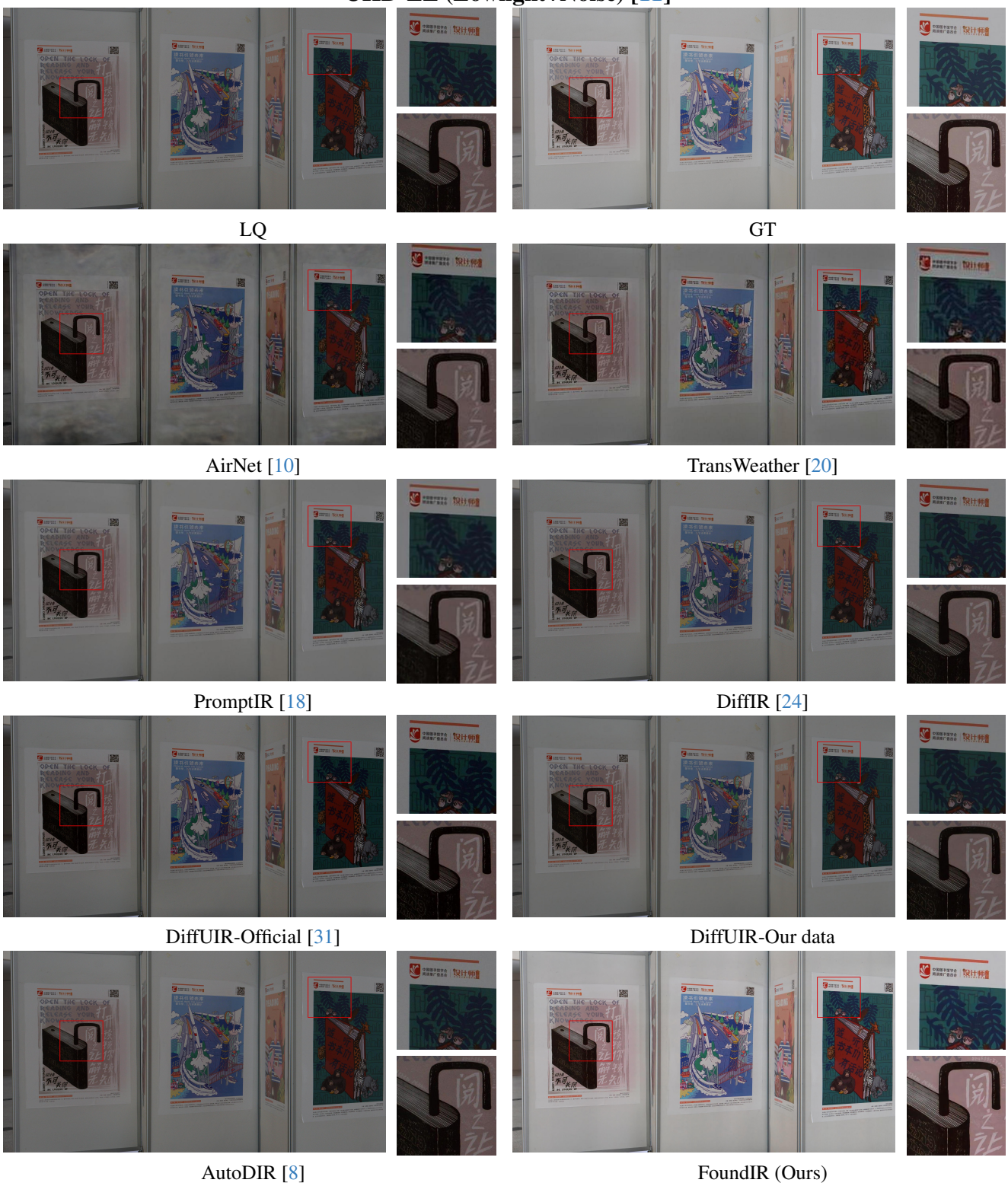


Figure 38. Visual comparison results on the public benchmark - UHD-LL [11]. Compared to the results restored by existing methods [8, 10, 18, 20, 24, 31], our approach generates a clearer image. Zoom in for a better view.

(a) Lowlight+Haze

(b) Mix-trained on existing synthetic datasets (200K)

(c) Mix-trained on existing synthetic+real datasets (200K)

(d) Mix-trained on 20% of our dataset (200K)

(e) Mix-trained on 50% of our dataset (500K)

(f) Mix-trained on our million-scale dataset

Figure 39. Generalization analysis under Lowlight+Haze condition. The results shown in (b)-(c) demonstrate that models trained on these datasets only enhance the image without effectively removing the haze. This indicates that existing training datasets struggle to address coupled degradations effectively. In contrast, the results shown in (d)-(f) illustrate that as the scale of our training dataset increases, the quality of image restoration improves progressively.

(a) Rain

(b) Mix-trained on existing synthetic datasets (200K)

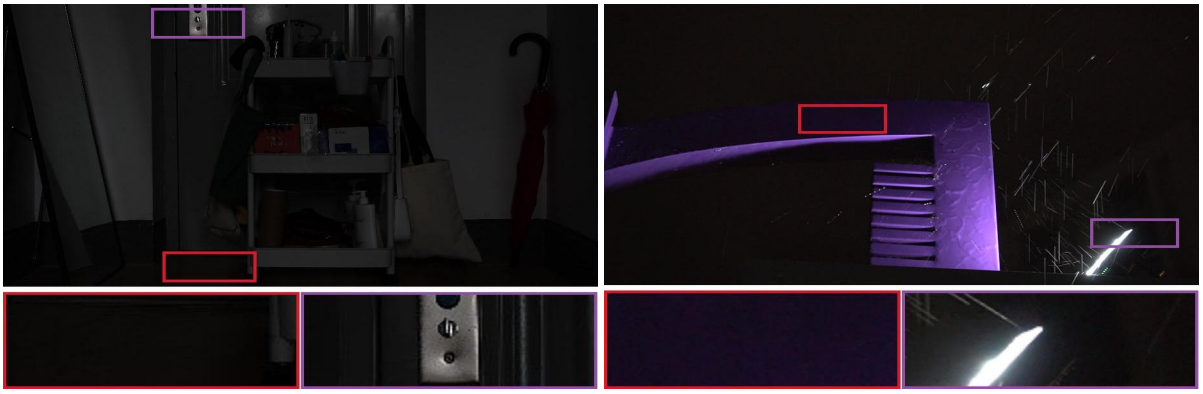
(c) Mix-trained on existing synthetic+real datasets (200K)

(d) Mix-trained on 20% of our dataset (200K)

(e) Mix-trained on 50% of our dataset (500K)

(f) Mix-trained on our million-scale dataset

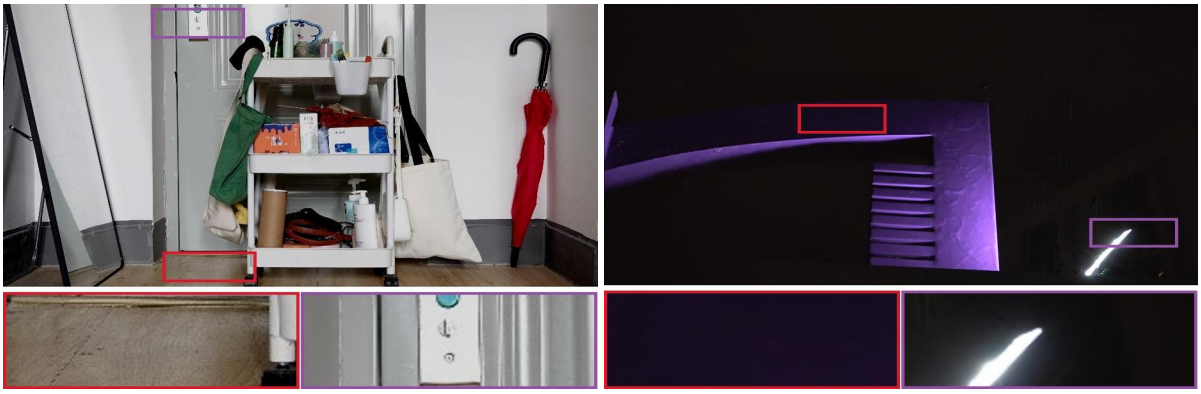
Figure 40. Generalization analysis under Rain condition. The results shown in (b)-(c) demonstrate that models trained on these datasets fail to remove rain streaks. This indicates that existing training datasets struggle to address real-world degradations effectively. In contrast, the results shown in (d)-(f) illustrate that as the scale of our training dataset increases, the quality of image restoration improves progressively.



(a) Lowlight (e) Rain

(b) Only generalist model (f) Only generalist model

(c) Only specialist model (g) Only specialist model



(d) Our ensemble model (h) Our ensemble model

Figure 41. Visual comparisons of different variants of our proposed FoundIR. It can be observed that our ensemble framework effectively removes complex degradations in real-world scenarios, producing much clearer results.

References

[1] Aakerberg, A., Nasrollahi, K., Moeslund, T.B.: Rellisur: A real low-light image super-resolution dataset. In: NeurIPS (2021) 4

[2] Chen, J., Pan, J., Dong, J.: Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In: CVPR (2025) 4

[3] Chen, X., Li, Z., Pu, Y., Liu, Y., Zhou, J., Qiao, Y., Dong, C.: A comparative study of image restoration networks for general backbone network design. In: ECCV (2024) 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[4] Conde, M.V., Geigle, G., Timofte, R.: Instructir: High-quality image restoration following human instructions. In: ECCV (2024) 2

[5] Deng, S., Ren, W., Yan, Y., Wang, T., Song, F., Cao, X.: Multi-scale separable network for ultra-high-definition video deblurring. In: ICCV (2021) 5, 34

[6] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE TPAMI (2010) 3

[7] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 4

[8] Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. In: ECCV (2024) 2, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

[9] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021) 4

[10] Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR (2022) 2, 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

[11] Li, C., Guo, C.L., Zhou, M., Liang, Z., Zhou, S., Feng, R., Loy, C.C.: Embedding fourier for ultra-high-definition low-light image enhancement. In: ICLR (2023) 5, 38, 39

[12] Li, W., Zhang, Q., Zhang, J., Huang, Z., Tian, X., Tao, D.: Toward real-world single image deraining: A new benchmark and beyond. arXiv preprint arXiv:2206.05514 (2022) 5, 35, 36

[13] Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: CVPR (2024) 3

[14] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: ICML (2023) 4

[15] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. In: ICLR (2024) 2, 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[16] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 8

[17] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE SPL (2012) 4

[18] Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: Promptir: Prompting for all-in-one image restoration. NeurIPS (2024) 2, 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

[19] Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE TIP (2018) 4

[20] Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: CVPR (2022) 2, 5, 34, 35, 36, 37, 38, 39

[21] Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023) 4

[22] Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: CVPR (2021) 3

[23] Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS (2024) 4

[24] Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. In: CVPR (2023) 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

[25] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024) 3

[26] Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: CVPR (2022) 4

[27] Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. In: NeurIPS (2023) 4

[28] Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: CVPR (2023) 2

[29] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 4

[30] Zhang, Y., Ding, L., Sharma, G.: Hazerd: an outdoor scene dataset and benchmark for single image dehazing. In: ICIP (2017) 5, 37

[31] Zheng, D., Wu, X.M., Yang, S., Zhang, J., Hu, J.F., Zheng, W.S.: Selective hourglass mapping for universal image restoration based on diffusion model. In: CVPR (2024) 2, 3, 4, 5, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

[32] Zheng, Z., Ren, W., Cao, X., Hu, X., Wang, T., Song, F., Jia, X.: Ultra-high-definition image dehazing via multi-guided bilateral learning. In: CVPR (2021) 3