Supplementary Material – Restormer: Efficient Transformer for High-Resolution Image Restoration

Syed Waqas Zamir[1] Aditya Arora[1] Salman Khan[2] Munawar Hayat[2,3]
Fahad Shahbaz Khan[2,4] Ming-Hsuan Yang[5,6,7]
[1]Inception Institute of AI [2]Mohamed bin Zayed University of AI [3]Monash University
[4]Linköping Univeristy [5]University of California, Merced [6]Yonsei University [7]Google Research

1. Datasets and Experimental Details

In Table 1, we list the datasets used for training and evaluation. Next we describe them for each individual task.

Image Deraining. Following [16, 28, 39], we train Restormer on 13,712 clean-rainy image pairs collected from numerous datasets [13, 20, 36, 41, 42], as shown in Table 1. With this single trained model, we perform evaluation on Rain100H [36], Rain100L [36], Test100 [42], Test2800 [13], and Test1200 [41]. PSNR/SSIM scores are computed on the Y channel in YCbCr color space as in other works [16, 39].

Single-Image Motion Deblurring. Consistent with existing methods [17, 33, 34, 39, 40], we use the GoPro [25] dataset for training our Restormer. It contains 2,103 blurry-sharp image pairs for training and 1,111 pairs for validation. To test the generalization ability of Restormer, we take our GoPro trained model and *directly evaluate* it on the testsets of HIDE [32] and RealBlur [31] datasets. The testset of the HIDE dataset [32] consists of 2,025 images, and it is particularly gathered for the human-aware motion deblurring. The blurry images in both the GoPro and HIDE datasets are synthetically generated. Whereas, the blurry-sharp image pairs of RealBlur dataset [31] are acquired in real-world conditions. The RealBlur dataset has two subsets: (1) RealBlur-J contains 980 images that are obtained directly as camera JPEG outputs, and (2) RealBlur-R is generated offline by applying white balance, demosaicking, and denoising operations to the RAW images. It also has 980 images.

Defocus Deblurring. For this task, we use a recently presented DPDD dataset [2]. DPDD consists of 500 indoor/outdoor scenes captured with a DSLR camera. Each scene contains three defocus input images and a corresponding all-in-focus ground-truth image. Three input images are labeled as left, right and center views. The left and right defocused sub-aperture views are acquired with a wide camera aperture setting, and the corresponding all-in-focus ground-truth image captured with a narrow aperture. We use this sub-aperture data to train Restormer for the dual-pixel defocus deblurring task. Whereas, the center input image and corresponding ground-truth is used for training Restormer for the single-image defocus deblurring task. The DPDD dataset [2] contains 350 images for training, 74 images for validation, and 76 images for testing (37 indoor and 39 outdoor).

In the dual-pixel defocus deblurring task, the input to Restormer is of 6-channel (concatenated left and right sub-aperture views) and the output is a 3-channel deblurred image. Furthermore, instead of using the residual learning on images $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$ (as shown in Fig. 2 of the main paper), we use skip connection on features, *i.e.*, $\mathbf{F_0} + \mathbf{F_r}$.

Gaussian Image Denoising. Following [22, 43], both for the grayscale and color denoising, we use a combined set of 800 images from DIV2K [4], 2,650 images of Flickr2K, 400 BSD500 images [5] and 4,744 images from WED [23]. Noisy images are generated by adding additive white Gaussian noise with noise level σ to clean images. Testing is performed on Set12 [45], BSD68 [24], Urban100 [15], Kodak24 [12] and McMaster [46] benchmark datasets.

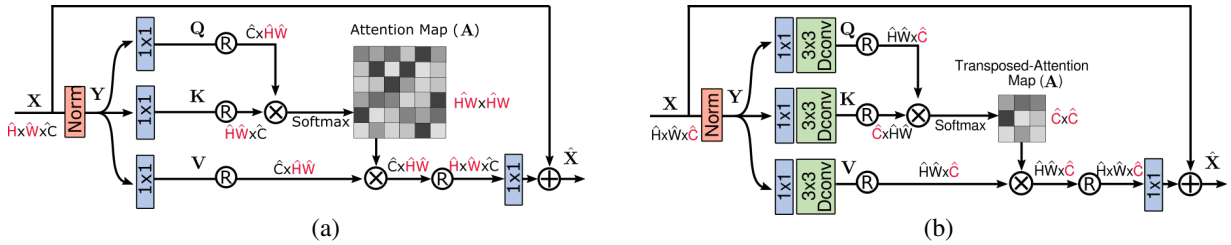
Real Image Denoising. To train our Restormer, we use 320 high-resolution images of the SIDD dataset [1]. With this SIDD trained model, we perform evaluation on 1,280 patches from the SIDD validation set [1] and 1,000 patches from the DND benchmark dataset [27]. These test patches on both datasets are extracted from the full resolution images by the original authors. Since the ground-truth of DND images is not publicly available, the PSNR/SSIM scores are obtained by uploading results to the online server https://noise.visinf.tu-darmstadt.de/.

2. Baseline vs Proposed Transformer Block

We provide illustrations in Fig. 1 and Fig. 2 to demonstrate the transition from the baseline to the proposed transformer block components *i.e.*, multi-Dconv head transposed attention (MDTA) and gated-Dconv feed-froward network (GDFN) to better illustrate our design contributions.

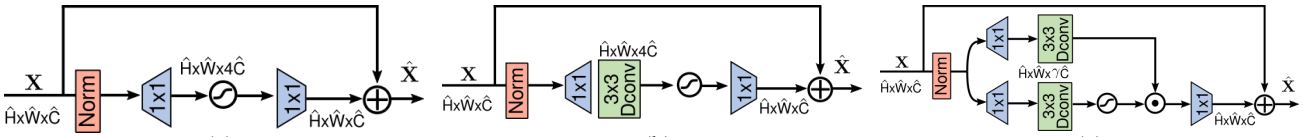
Table 1. Dataset description for various image restoration tasks.

Tasks	Datasets	Train Samples	Test Samples	Testset Rename
Deraining	Rain14000 [13]	11200	2800	Test2800
	Rain1800 [36]	1800	0	-
	Rain800 [42]	700	100	Test100
	Rain100H [36]	0	100	Rain100H
	Rain100L [36]	0	100	Rain100L
	Rain1200 [41]	0	1200	Test1200
	Rain12 [20]	12	0	-
Motion Deblurring	GoPro [25]	2103	1111	-
	HIDE [32]	0	2025	-
	RealBlur [31]	0	1960	-
Denoising	SIDD [1]	320	1280 patches from 40 images	-
	DND [27]	0	1000 patches from 50 images	-
	DIV2K [4]	800	0	-
	Flickr2K	2650	0	-
	BSD500 [5]	400	0	-
	WED [23]	4744	0	-
	Set12 [45]	0	12	-
	BSD68 [24]	0	68	-
	Urban100 [15]	0	100	-
	Kodak24 [12]	0	24	-
	McMaster [46]	0	18	-
Defocus Deblurring	DPDD [2]	350	76	-



(a) (b)

Figure 1. Comparisons between (a) the conventional multi-head self attention [11] and (b) the proposed multi-Dconv head transposed attention (MDTA). Our MDTA module implicitly models global context by applying self-attention across channels rather than the spatial dimension, thus having linear complexity rather than quadratic.



(a) (b) (c)

Figure 2. Comparisons among (a) conventional feed-forward network [11], (b) Dconv feed-forward network [21] and (c) the proposed gated-Dconv feed-forward network (GDFN). Since the proposed GDFN performs more operations as compared to (a) and (b), we reduce the expansion ratio γ so as to have similar parameters and identical compute burden.

3. Computational Comparisons

Table 2 shows that, compared to existing Transformer-based methods, our Restormer is more efficient and effective.

4. Additional Visual Results

We present images reproduced by the proposed Restormer and those of other competing approaches for different image restoration tasks as qualitative examples.

Table 2. Computational comparison of Transformer-based image restoration models.

	Params (M)	FLOPs (B)	Time (s) 256×256 patch	PSNR (Denoising; Table 5) Urban100, $\sigma = 50$
IPT [8]	115.3	379	3.35	29.71
SwinIR [22]	11.50	444	1.80	29.82
Restormer	26.11	141	0.11	30.02

- **Image deraining:** Figures 3,4,5
- **Single-image motion deblurring:** Figures 6, 7, 8, 9.
- **Dual-pixel defocus deblurring:** Figures 10, 11.
- **Gaussian grayscale image denoising:** Figures 12, 13.
- **Gaussian color image denoising:** Figures 14, 15.
- **Real image denoising:** Figures 16, 17, 18.

5. Limitations and Future Work

While Restormer emerges as a competitive backbone architecture across several benchmarks, it can be further improved with the specific complimentary modules. For example, feature aggregation approaches can be incorporated which seek to resolve the spatial feature misalignment that occurs when the *high-resolution* encoder features are aggregated with the *low-resolution* decoder features (via skip connections). This could be achieved by employing deformable convolutions [10] or by using cross-scale attention mechanism [7] instead of concatenation used in our architecture. While the potential feature misalignment is not specific to Restormer, but a common issue of encoder-decoder designs, our approach can further improve with better feature alignment techniques.

In this work, we use L_1 loss, AdamW optimizer, GeLU non-linearity and Layer normalization following default choices in the existing literature. Other choices (specific to image restoration) might yield improved results, that can be explored in the future work. As an example, one simple experiment using the H-Swish [14,21] non-linearity can lead to better than our reported results, however, we opt for GeLU since our goal is to provide a generic and strong backbone instead of a heavily tuned architecture. Restormer employs deep-narrow architecture design which applies sequential operations and have higher latency which can be improved. Further, a multi-stream Transformer (better parallelizable) can be explored without compromising accuracy.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 1, 2, 18, 19

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020. 1, 2, 12, 13

[3] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *ICCV*, 2021. 12, 13

[4] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 1, 2

[5] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI.*, 2011. 1, 2

[6] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *ECCV*, 2020. 20

[7] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 3

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 3, 16, 17

[9] 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. 8, 9, 10, 11

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[12] Rich Franzen. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`, 1999. Online accessed 24 Oct 2021. 1, 2

[13] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 1, 2

[14] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 3

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 1, 2, 14, 15, 16, 17

[16] Kui Jiang, Zhongyuan Wang, Peng Yi, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020. 1, 5, 6, 7

[17] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 1

[18] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021. 12, 13

[19] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018. 5, 6, 7

[20] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *CVPR*, 2016. 1, 2

[21] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: bringing locality to vision transformers. *arXiv:2104.05707*, 2021. 2, 3

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 1, 3, 14, 15, 16, 17

[23] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *TIP*, 2016. 1, 2

[24] 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*, 2001. 1, 2

[25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2, 8, 9

[26] 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, 9

[27] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 1, 2, 20

[28] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, 2021. 1

[29] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *CVPR*, 2021. 14, 15, 18, 19

[30] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019. 5, 6, 7

[31] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 1, 2, 10, 11

[32] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, 2019. 1, 2

[33] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 2020. 1, 8, 9

[34] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1

[35] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv:2106.03106*, 2021. 12, 13, 18, 19, 20

[36] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 1, 2, 5, 6, 7

[37] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *CVPR*, 2019. 5, 6, 7

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 18, 19, 20

[39] 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, 9, 10, 11, 18, 19, 20

[40] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 1, 8, 9

[41] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, 2018. 1, 2, 5, 7

[42] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *TCSVT*, 2019. 1, 2, 5, 6

[43] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *TPAMI*, 2021. 1, 14, 15, 16, 17

[44] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, 2020. 8, 9, 10, 11

[45] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 2017. 1, 2

[46] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *JEI*, 2011. 1, 2

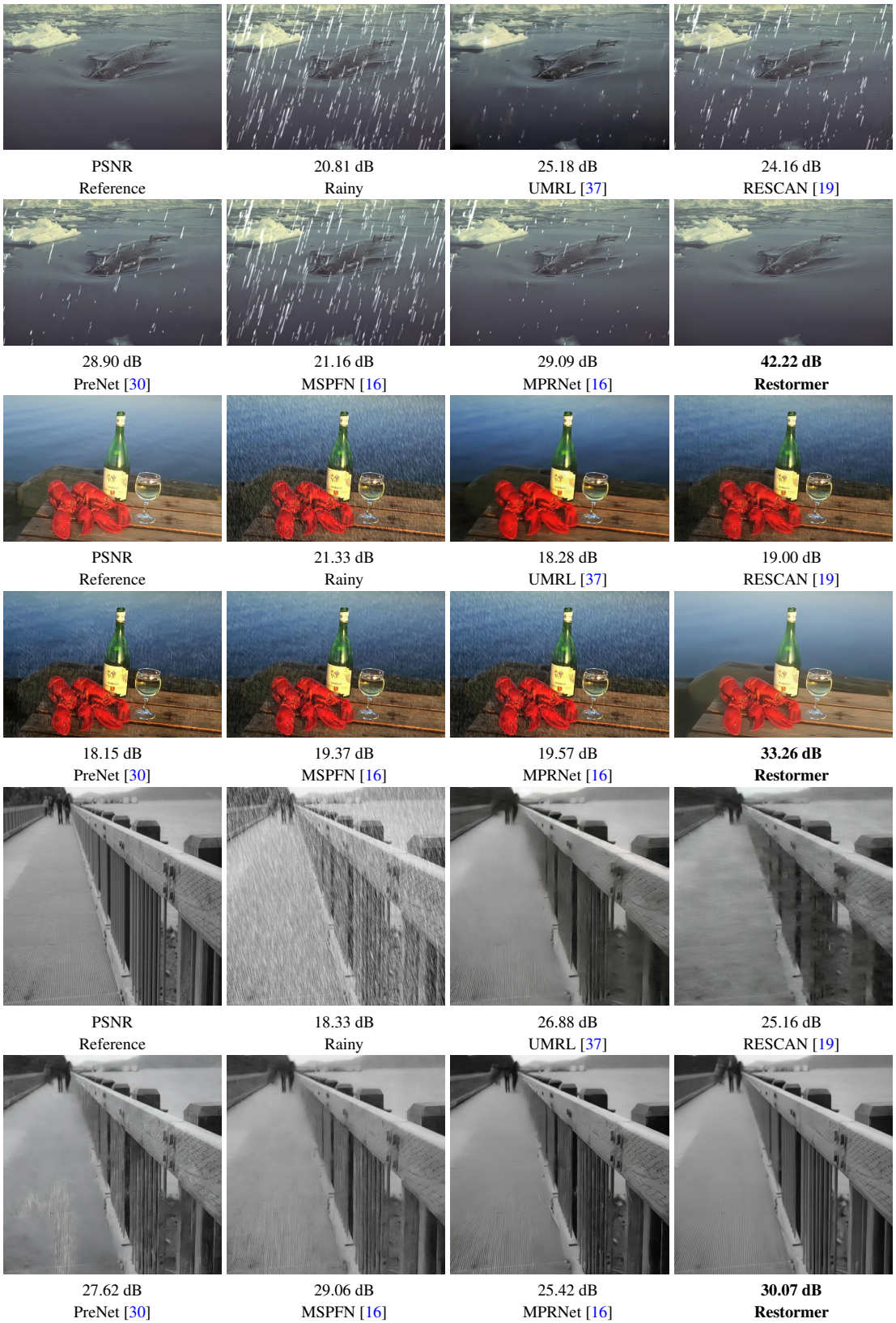


Figure 3. **Image deraining**. Top image is from Rain100L [36], middle is from Test100 [42] and the bottom is from Test1200 [41].

Figure 4. **Image deraining**. Top image is from Rain100L [36], middle is from Test100 [42] and the bottom is from Rain100H [36].

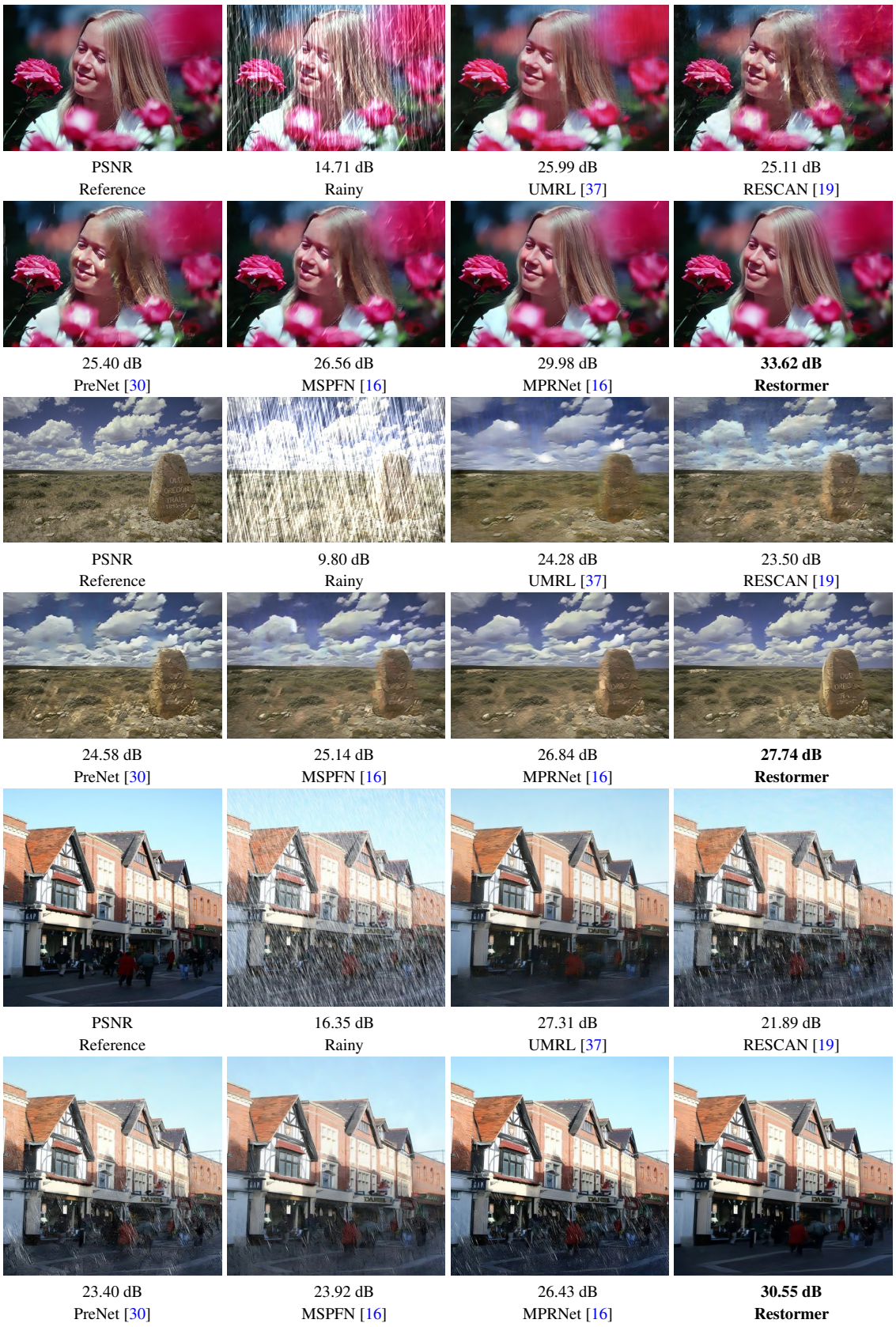


Figure 5. **Image deraining**. Top two images are from Rain100H [36], and the bottom is from Test1200 [41].

Figure 6. **Single-image motion deblurring** comparisons on the GoPro dataset [25].

Figure 7. **Single-image motion deblurring** comparisons on the GoPro dataset [25].

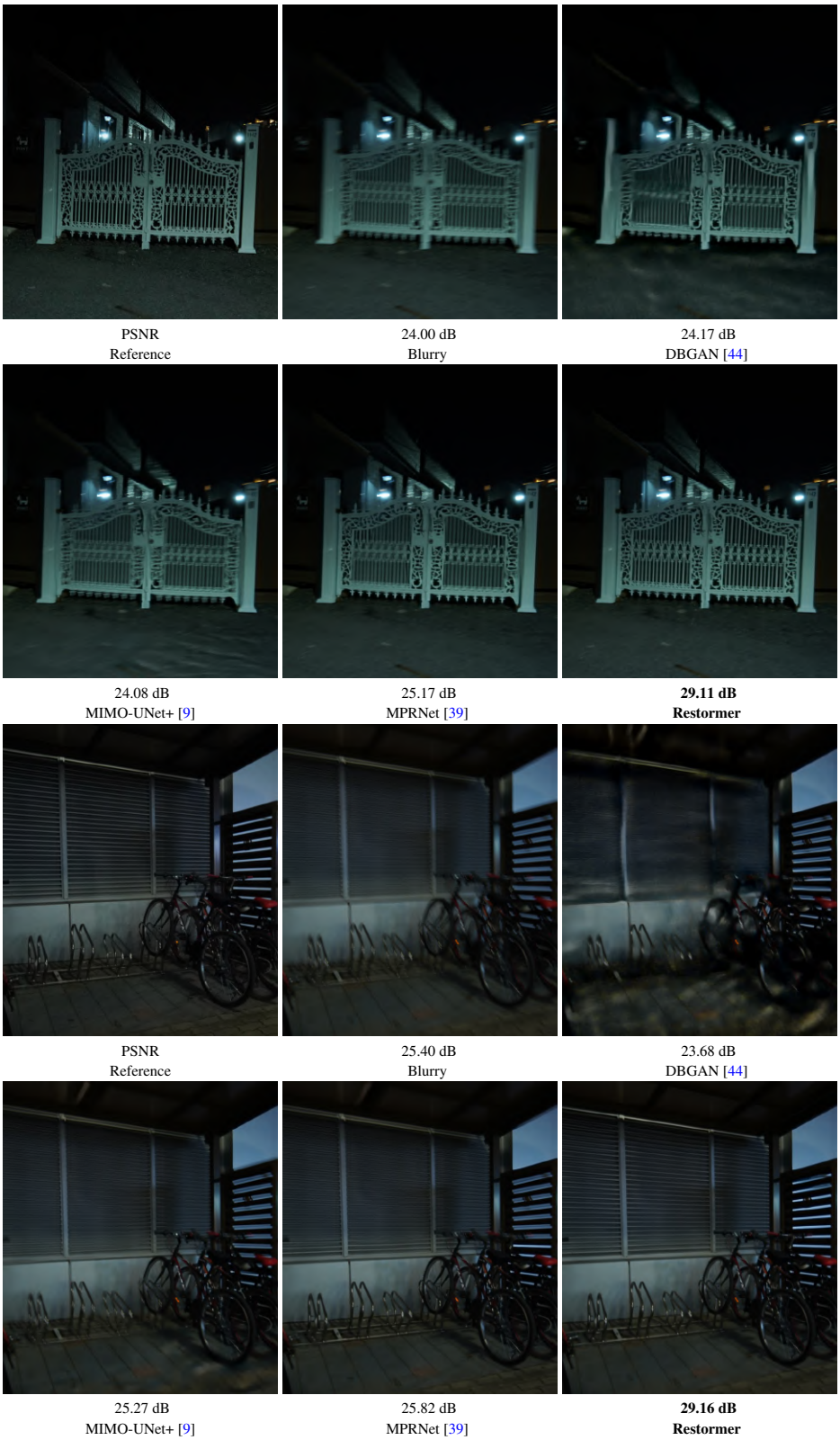


Figure 8. **Single-image motion deblurring** comparisons on the RealBlur dataset [31].

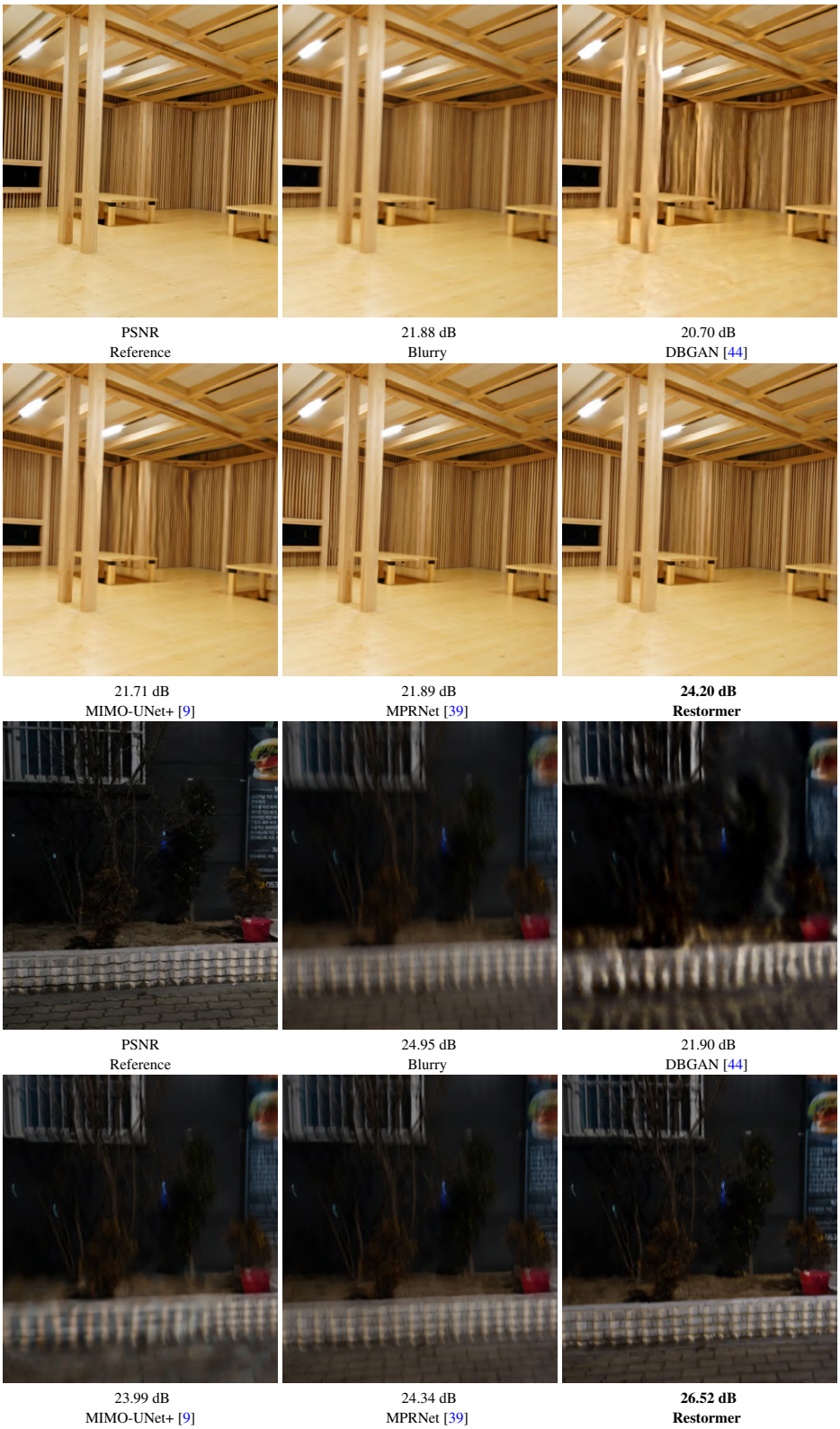
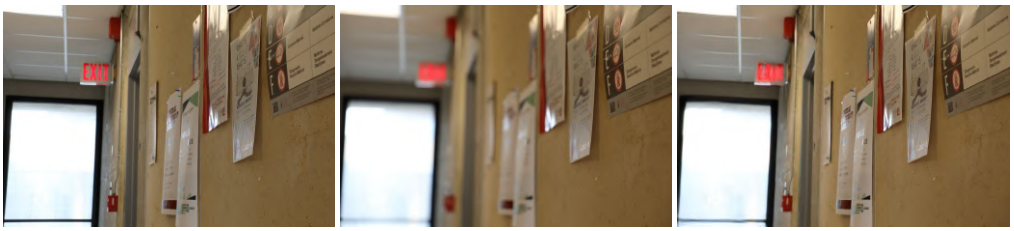
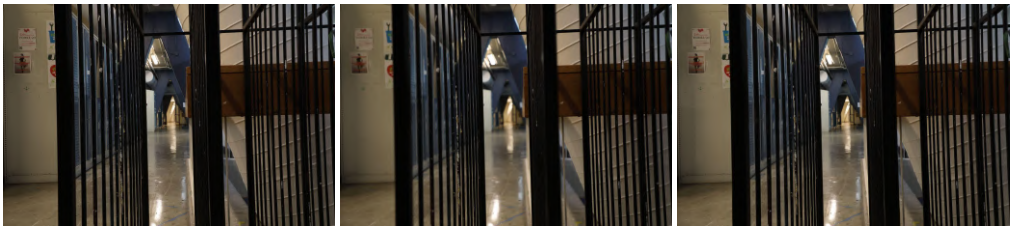


Figure 9. **Single-image motion deblurring** comparisons on the RealBlur dataset [31].

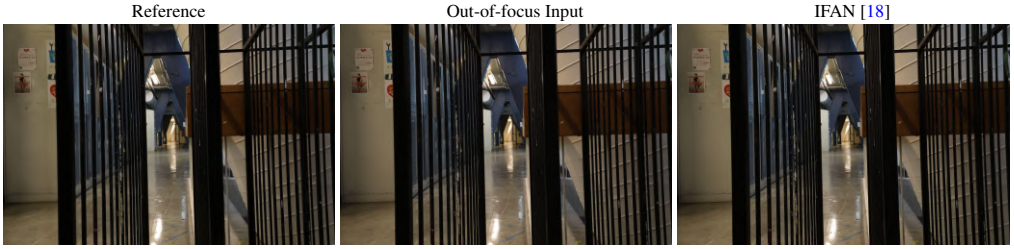


PSNR | 22.17 dB | 26.26 dB
Reference | Out-of-focus Input | IFAN [18]

25.89 dB | 26.51 dB | **29.88 dB**
RDPD [3] | Uformer [35] | **Restormer**



PSNR | 28.46 dB | 29.14 dB
Reference | Out-of-focus Input | IFAN [18]



29.72 dB | 30.26 dB | **32.10 dB**
RDPD [3] | Uformer [35] | **Restormer**

PSNR | 26.91 dB | 29.54 dB
Reference | Out-of-focus Input | IFAN [18]

29.41 dB | 29.15 dB | **32.44 dB**
RDPD [3] | Uformer [35] | **Restormer**

Figure 10. **Dual-pixel defocus deblurring** on DPDD [2].

PSNR | 18.30 dB | 23.39 dB
Reference | Out-of-focus Input | IFAN [18]

21.79 dB | 22.04 dB | **26.29 dB**
RDPD [3] | Uformer [35] | **Restormer**

PSNR | 30.73 dB | 30.90 dB
Reference | Out-of-focus Input | IFAN [18]

29.59 dB | 30.93 dB | **33.74 dB**
RDPD [3] | Uformer [35] | **Restormer**

PSNR | 21.75 dB | 25.88 dB
Reference | Out-of-focus Input | IFAN [18]

24.22 dB | 25.16 dB | **28.47 dB**
RDPD [3] | Uformer [35] | **Restormer**

Figure 11. **Dual-pixel defocus deblurring** on DPDD [2].

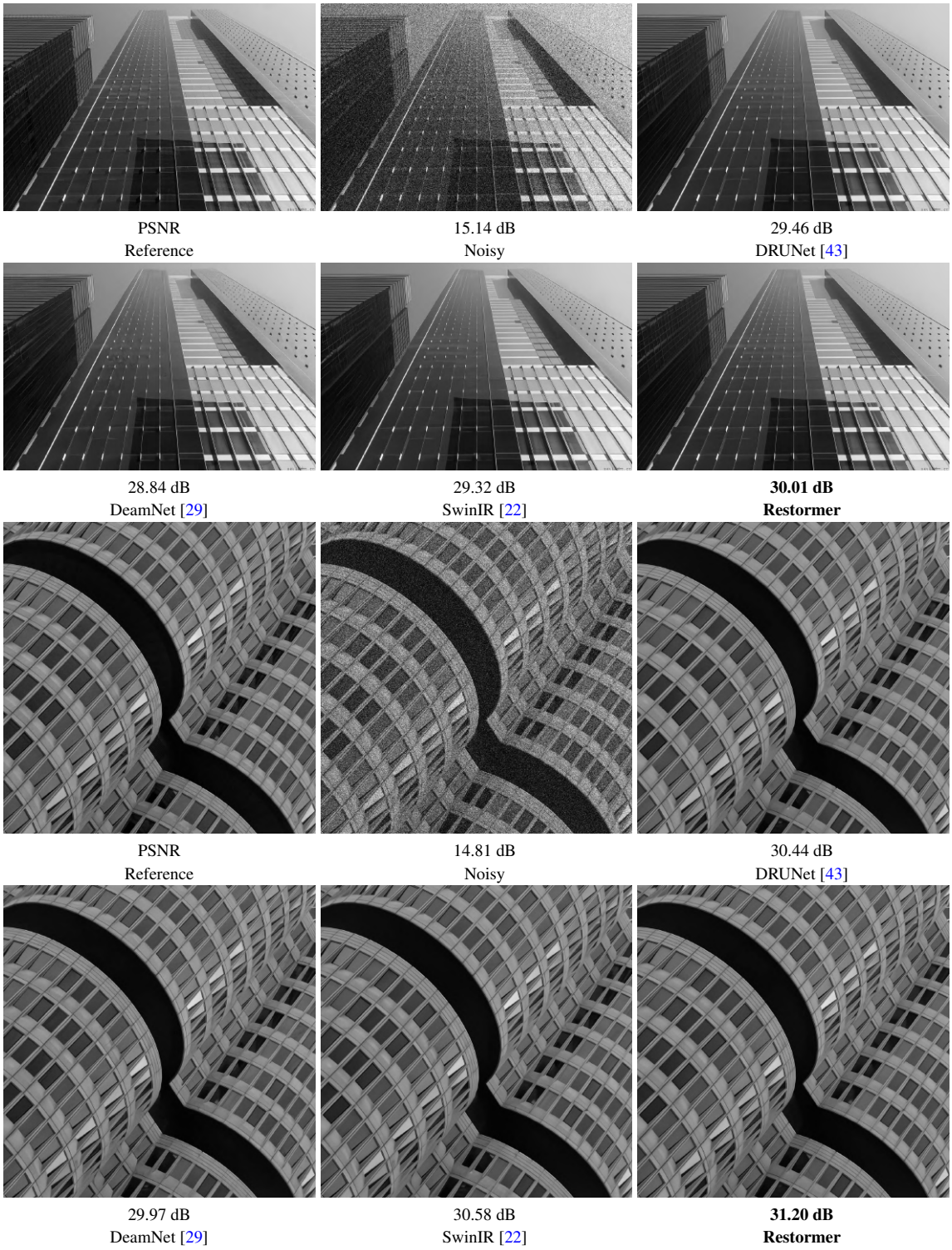


Figure 12. **Gaussian grayscale image denoising** comparisons on the Urban100 dataset [15].

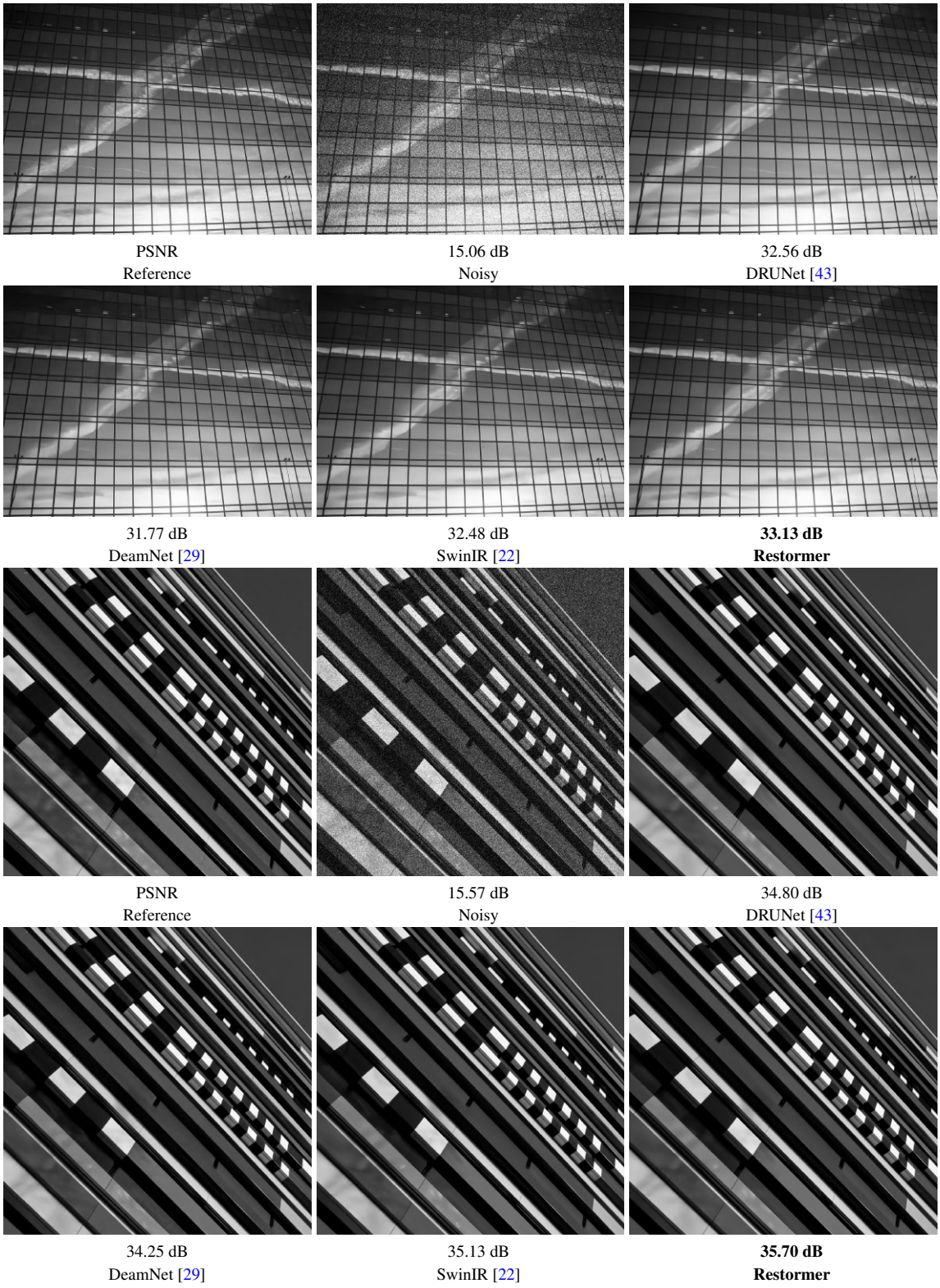


Figure 13. **Gaussian grayscale image denoising** comparisons on the Urban100 dataset [15].

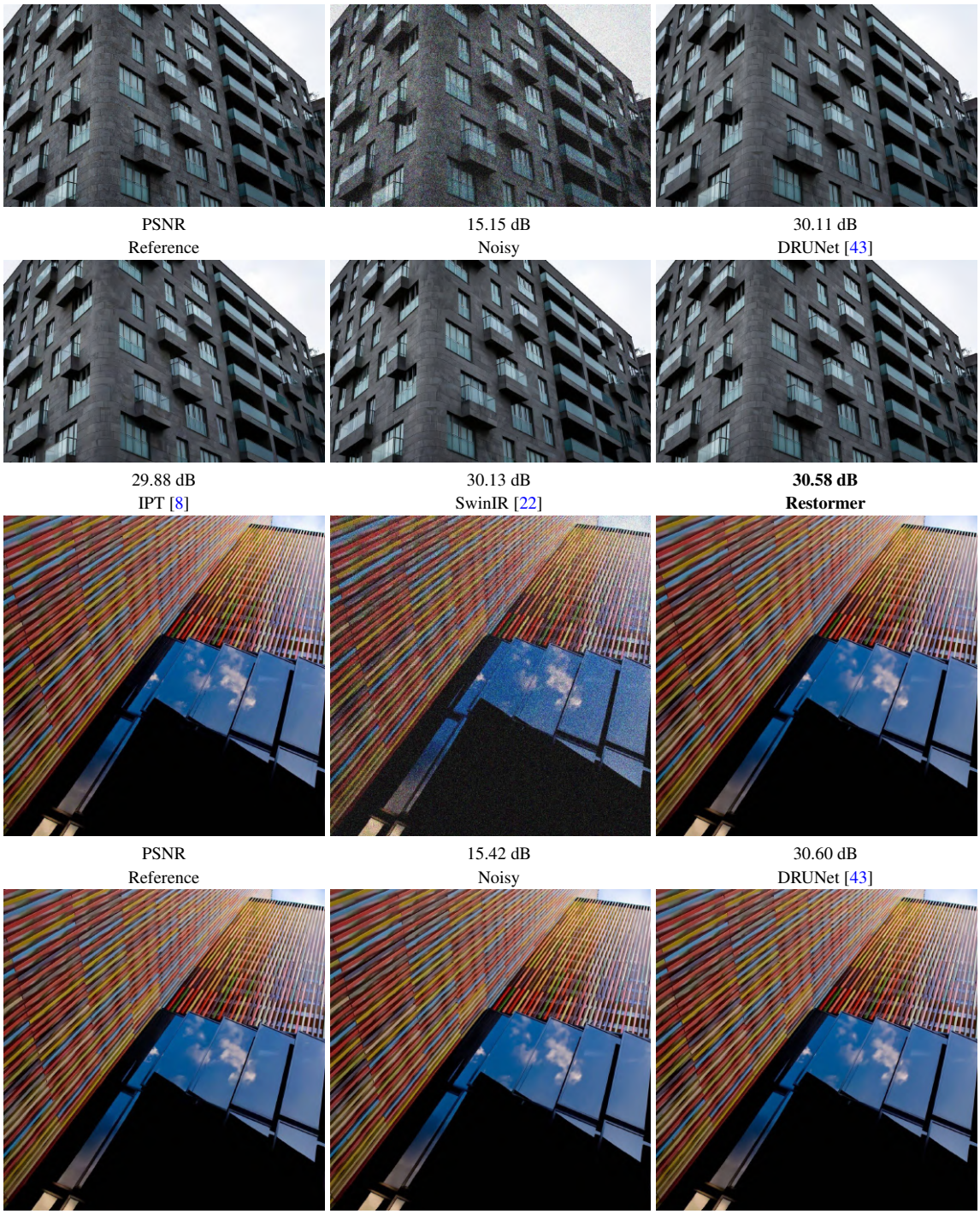
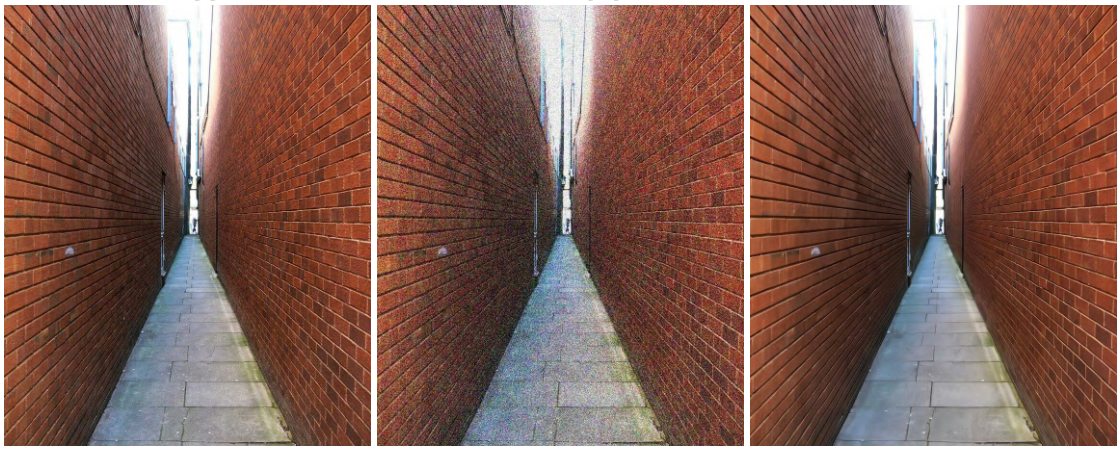


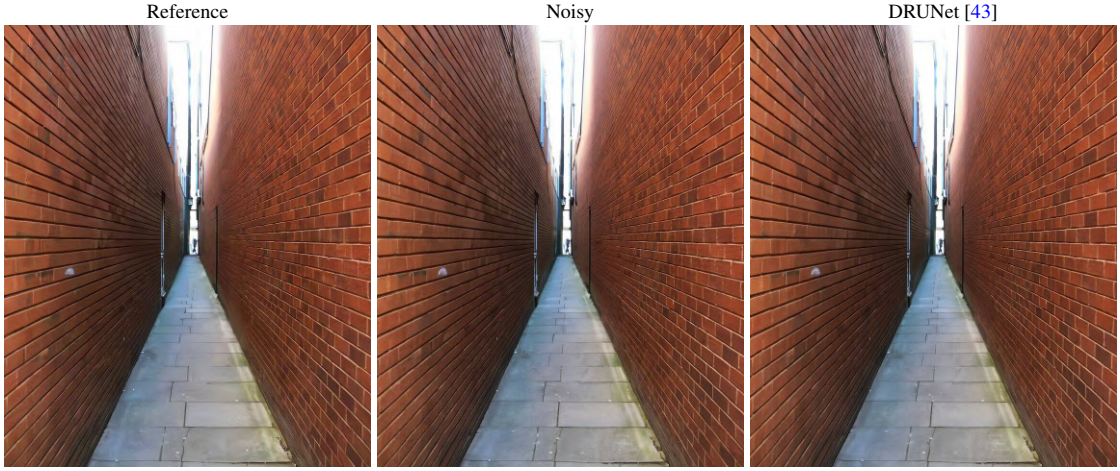
Figure 14. **Gaussian color image denoising** comparisons on the Urban100 dataset [15].

PSNR | 14.92 dB | 29.41 dB
Reference | Noisy | DRUNet [43]

29.51 dB | 29.64 dB | **29.89 dB**
IPT [8] | SwinIR [22] | **Restormer**



PSNR | 14.95 dB | 28.21 dB
Reference | Noisy | DRUNet [43]



28.17 dB | 28.32 dB | **28.46 dB**
IPT [8] | SwinIR [22] | **Restormer**

Figure 15. **Gaussian color image denoising** comparisons on the Urban100 dataset [15].

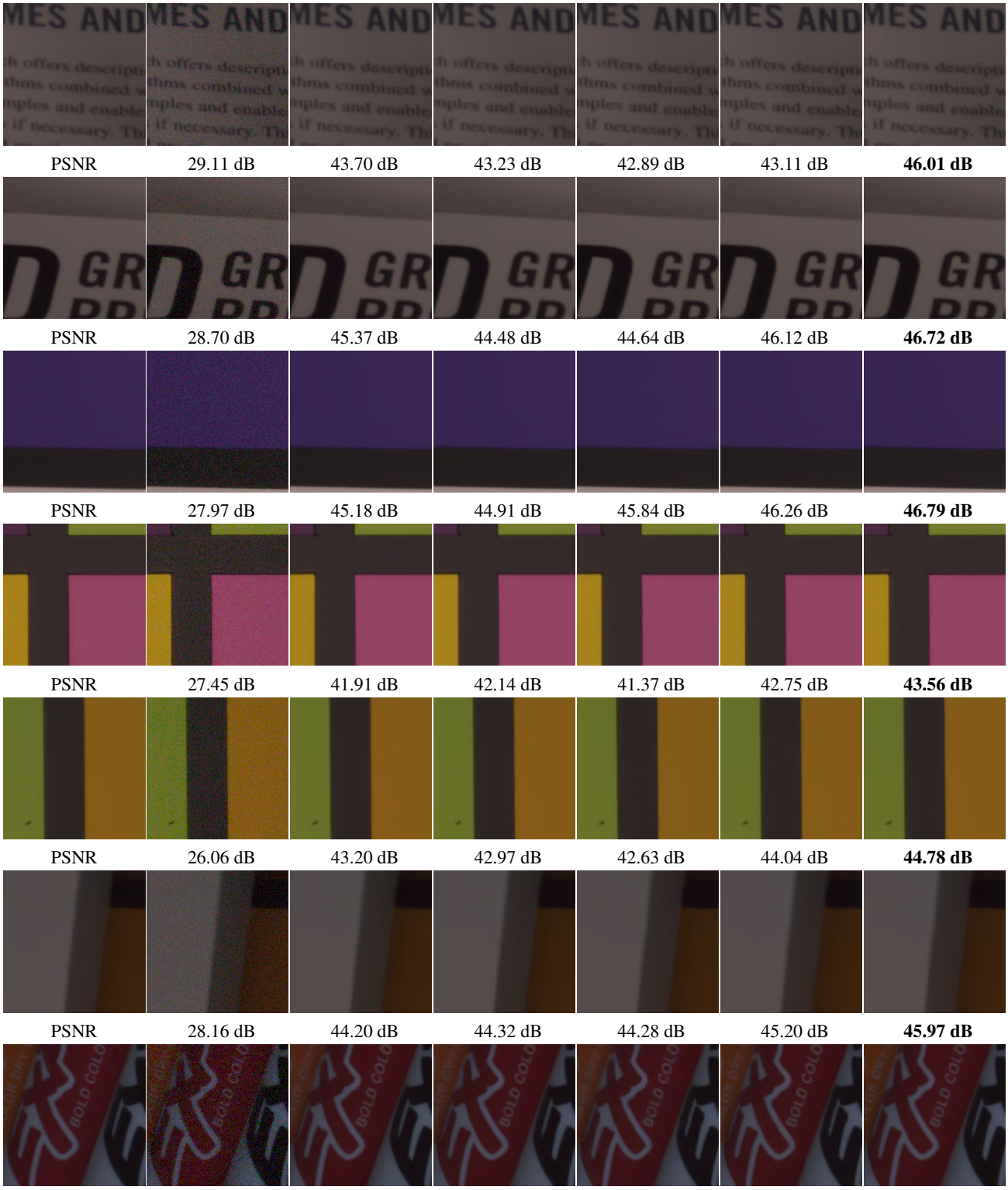


Figure 16. **Real image denoising** on the SIDD dataset [1].

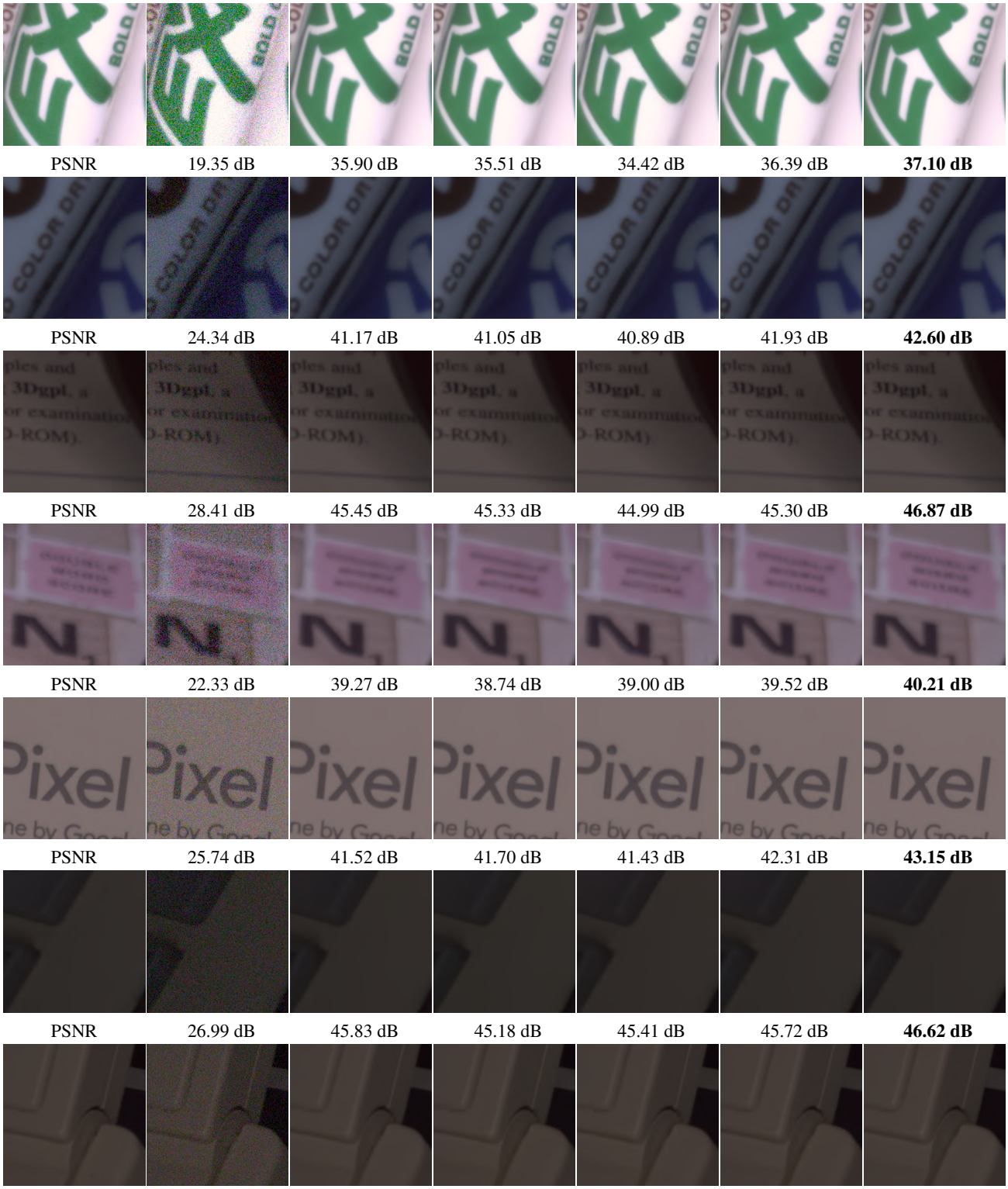


Figure 17. **Real image denoising** on the SIDD dataset [1].

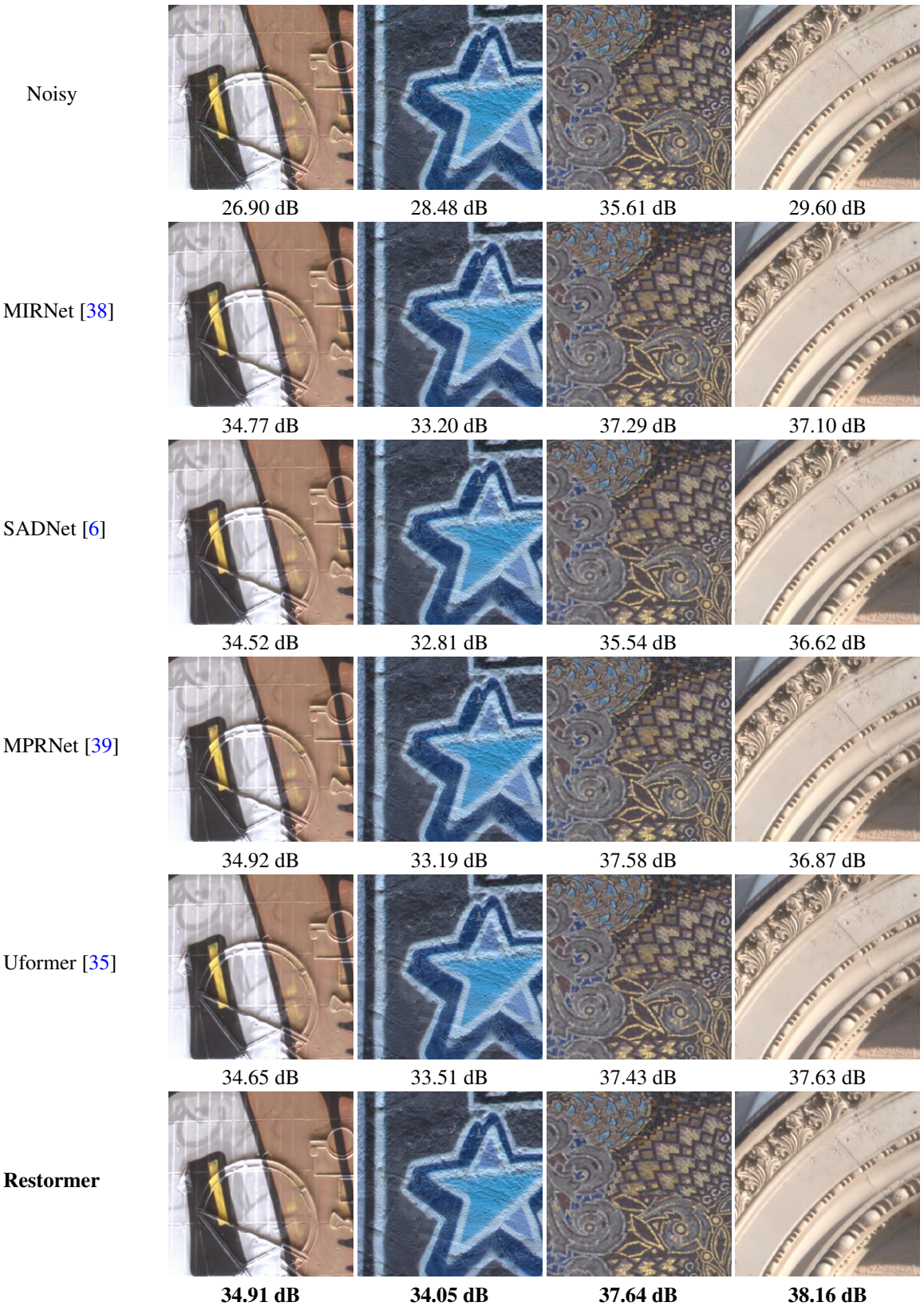


Figure 18. **Real image denoising** on the DND benchmark dataset [27].