XFORMER: HYBRID X-SHAPED TRANSFORMER FOR IMAGE DENOISING– SUPPLEMENTARY MATERIAL

Jiale Zhang[1], **Yulun Zhang**[1*†], **Jinjin Gu**[2,3], **Jiahua Dong**[4], **Linghe Kong**[1*], **Xiaokang Yang**[1]
[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory, [3]University of Sydney,
[4]Shenyang Institute of Automation, Chinese Academy of Sciences

APPENDIX

In this supplementary material, we firstly provide more experimental results for ablation studies in Sec. 1. We present comprehensive contents including additional quantitative comparisons and convergence analyses. Secondly, we provide more comparisons with recent state-of-the-art (SOTA) method to validate the superiority of our method in Sec. 2. Thirdly, we employ our Xformer to train a single model to solve Gaussian grayscale or color image denoising with various noise levels in Sec. 3. Experimental results further demonstrate that our proposed method achieves better performance. Fourthly, we provide additional experimental results of motion deblurring task and provide more visual results in Sec. 4 and 5 . Lastly, we discuss the limitations and future work. Code and models are available in the the website https://github.com/gladzhang/xformer.

1 ADDITIONAL ABLATION RESULTS

In this section, we present more comprehensive results to demonstrate the effectiveness of the Bidirectional Connection Unit (BCU). We also provide more comparisons to validate the necessity of dual branches. Then, we discuss the results of the model with smaller window size. Lastly, we discuss the comparisons between the new model architecture with alternating blocks and our Xformer.

1.1 IMPACT OF BCU

As introduced in the Ablation Study section of the main paper, we conduct comparative experiments about whether to use the BCU. Note that training iterations are 100k on Gaussian color image denoising with noise level σ=50. We provide more comparative results here.

Quantitative Comparisons. Table 1a shows the complete evaluation results on commonly-used four benchmark datasets, including CBSD68 (Martin et al., 2001), Kodak24 (Franzen, 1999), McMaster (Zhang et al., 2011), and Urban100 (Huang et al., 2015). As we can see, our Xformer with BCU achieves better performance across all benchmarks. Equipped with BCU, our proposed concurrent network can fuse two styles of deep features and simultaneously capture patch-level and channel-level information. Therefore, it plays an important role in improving the performance of our proposed Transformer-based network.

Convergence Analyses. We provide the validation curve comparisons for the corresponding ablation experiments. We show validation curves on all four benchmarks. As shown in Fig. 1, our Xformer with BCU can achieve obvious performance gains over that without BCU.

1.2 IMPACT OF DIFFERENT BRANCHES

We conduct comparative experiments to investigate the importance of different branches in our network. The details have been introduced in the Ablation Study section of the main paper. The training iterations are 100k on Gaussian color image denoising with noise level σ=50. We provide more quantitative comparisons here.

*Corresponding authors: Yulun Zhang, yulun100@gmail.com; Linghe Kong, linghe.kong@sjtu.edu.cn.
†The work was mainly done when Yulun Zhang was at ETH Zürich.

Method	w/o BCU		w/ BCU	
	PSNR	SSIM	PSNR	SSIM
CBSD68	28.55	0.8094	28.57	0.8108
Kodak24	29.76	0.8190	29.79	0.8203
McMaster	30.17	0.8479	30.22	0.8487
Urban100	29.82	0.8842	29.94	0.8865

(a) Ablation study about the BCU.

Method	CBSD68		Kodak24		McMaster		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
single STB-branch	28.52	0.8103	29.74	0.8201	30.14	0.8478	29.84	0.8853
single CTB-branch	28.53	0.8099	29.74	0.8189	30.14	0.8477	29.68	0.8829
two-branch w/o BCU	28.55	0.8094	29.76	0.8190	30.17	0.8479	29.82	0.8842
two-branch w/ BCU	28.57	0.8108	29.79	0.8203	30.22	0.8487	29.94	0.8865

(b) Ablation study about using different branches.

Table 1: Ablation experiments. For ablation, we train models on Gaussian color image denoising task with σ=50 for 100k iterations and test on four benchmark datasets. We report the PSNR (dB) and SSIM scores.

Quantitative Comparisons. Table 1b shows the complete evaluation results on four benchmark datasets, which are CBSD68 (Martin et al., 2001), Kodak24 (Franzen, 1999), McMaster (Zhang et al., 2011), and Urban100 (Huang et al., 2015). The PSNR and SSIM scores are reported. As we can see, the model using two branches with BCU achieves the best performance across all benchmarks. It achieves the obvious performance improvement over the model using the single STB-based branch or CTB-based branch. Besides, we find that the model using dual branches without BCU has the limited performance. Although the dual branches enable the network to capture two levels of information, the direct connection of these two branches by concatenating has limited ability to utilize the information. Equipped with BCU, the performance is greatly enhanced. Therefore, the effective fusion of patch-level information from STB and channel-level information from CTB is very important for achieving state-of-the-art results. In conclusion, the joint designs of dual branches and the BCU bring promising performance improvement for our method.

1.3 Impact of Window Size

We use the window-based self-attention block in the spatial-wise branch to capture the patch-level information. Therefore, the size of window affects the performance of the proposed model. For further comparisons, we also provide the results of Xformer with smaller window size in STB. The new model variant named Xformer-SW only changes the window size to 8. We train the model on the task of the 15-level gaussian color image denoising. We provide the comparisons with recent state-of-the-art methods, which are SwinIR (Liang et al., 2021), Uformer (Wang et al., 2022) and Restormer (Zamir et al., 2022). Since the Uformer has not provided the results of 15-level color denoising in their paper, we train the basic version of Uformer by ourselves. We use the officially provided model settings and keep the training conditions the same with Restormer and Xformer. The detailed results are shown in Tab. 2.

Quantitative Comparisons. Table 2 shows the evaluation results on four benchmark datasets for 15-level gaussian color denoisng. The PSNR scores are reported. As we can see, our method Xformer-SW still achieves the best performance. Besides, the model parameters and FLOPs of our model are acceptable as shown in the table. In conclusion, it is validated that our proposed model with smaller window size still has promising performance on image denoising tasks.

1.4 Comparisons with Alternating Strategy

Since the joint usage of STB and CTB is important, we also try another model architecture with alternating STB and CTB in a complete U-shaped network. We compare this model architecture with our proposed Xformer. To make a fair comparison, we keep the specific settings in STB or CTB the same. Besides, the layers number in corresponding stages of the network are the same. We name the new model architecture as AlternateNet. We train this model under the same settings with Xformer on the task of 15-level color image denoising. For ablation, we provide the evaluation results of these two different models with 200k training iterations on four commonly-used benchmark datasets. The detailed results are shown in Tab. 4.

Quantitative Comparisons. As shown in Tab. 4, our Xformer achieves better results on all benchmark datasets. It shows that the alternating usage of STB and CTB has suboptimal performance. It is because that the direct alternating connection of these two blocks fails to effectively fuse the patch-level and channel-level information. In contrast, our proposed concurrent network Xformer is able to achieve promising performance since it has powerful ability to utilize the fused information.

Method	Params (M)	FLOPs (G)	CBSD68	Kodak24	McMaster	Urban100
SwinIR (Liang et al., 2021)	11.50	201.2	34.42	35.34	35.61	35.13
Uformer* (Wang et al., 2022)	50.88	21.7	34.41	35.32	35.55	35.06
Restormer (Zamir et al., 2022)	26.11	38.7	34.40	35.35*	35.61	35.13
Xformer-SW	25.15	38.8	**34.42**	**35.37**	**35.65**	**35.22**

Table 2: We train our model while setting the window size to 8. Parameters and FLOPs are also reported. We calculate FLOPs while setting the input size to $3\times128\times128$. The best results are **bolded**. * denotes results obtained by testing with officially provided pre-trained models. * means that we train the Uformer by ourselves under the same settings.

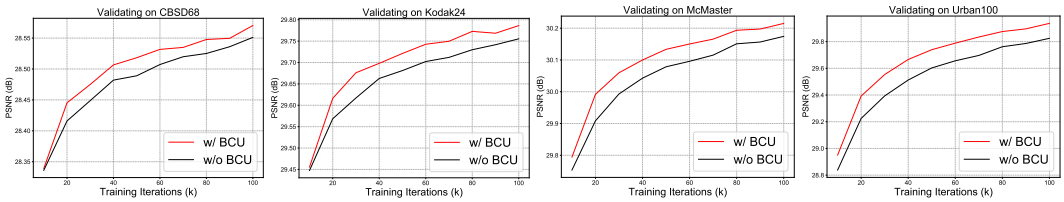

Figure 1: Convergence analyses on four benchmark datasets. All the validating curves are obtained by sampling every 10k training iterations. Comparisons are based on the ablation study about whether to use BCU. Training task is Gaussian color image denoising with noise level σ=50. We train these two models under the same training settings for 100k iterations.

2 Additional Comparisons with SOTA

In this section, we provide detailed comparisons with recent state-of-the-art method Restormer (Zamir et al., 2022). We train the model of Restormer based on the officially provided training settings and code. In order to make fair comparisons, we also train our proposed model under the same training settings. In detail, the training is based on Gaussian color image denoising with noise level σ=15. Total training iterations are 300k and the progressive training strategy is used. The detailed comparisons are introduced as follows.

Convergence Analyses. We provide the validation curve comparisons for the trained two models. We show validation curves on all four benchmarks, which are CBSD68 (Martin et al., 2001), Kodak24 (Franzen, 1999), and Urban100 (Huang et al., 2015). As shown in Fig. 2, our Xformer can achieve significant performance improvement over Restormer across all testing datasets. Experimental results validate that our proposed Xformer is becoming a new promising Transformer-based image denoising network. Thanks to the fusion of channel-level and patch-level information, our proposed method can exploit stronger global information modeling ability in two branches and achieve state-of-the-art performance.

3 One Model for Various Noise Levels

We conduct new experiments on Gaussian grayscale and color image denoising about using one model to solve different levels of noise. In detail, we employ the proposed Xformer to train two new models while solving the synthetic noise removal tasks with various noise levels. Specifically, we train a single model to handle different noise levels, including 15, 25, and 50. As introduced in the main paper, we do not change any parameters settings of Xformer. During training process, the input noise levels are randomly determined between 0 and 50. Therefore, the trained models enjoy robustness to handle various noise levels. We present the comparative results as follows. Note that the evaluation results are based on the commonly-used benchmarks, including Set12 (Zhang et al., 2017a), BSD68 (Martin et al., 2001), Kodak24 (Franzen, 1999), McMaster (Zhang et al., 2011), and Urban100 (Huang et al., 2015).

Quantitative Comparisons. We compare our Xformer to recent leading methods, including DRUNet (Zhang et al., 2021) and Restormer (Zamir et al., 2022). Note that DRUNet asks the inputs to include the noise level map while our model only needs the degraded images as inputs. As Restormer also trained a single model to solve various noise levels and achieved SOTA results, we mainly compare our method to it. As shown in Tab. 3, our proposed Xformer obtains the best

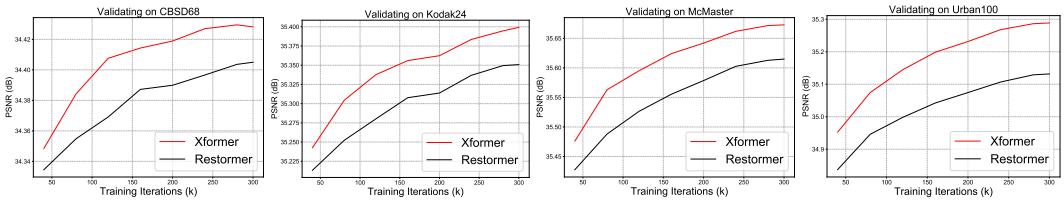

Figure 2: Convergence analyses on four benchmark datasets. All the validating curves are obtained by sampling every 40k training iterations. Comparisons are based on Restormer and our proposed Xformer. Training task is Gaussian color image denoising with noise level σ=15. All the training settings are the same. We train these two models for 300k iterations.

Method	Set12			BSD68			Urban100		
	σ =15	σ =25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DRUNet	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
Restormer	33.35	31.04	28.01	31.95	29.51	26.62	33.67	31.39	28.33
Xformer (ours)	**33.45**	**31.15**	**28.09**	**31.97**	**29.54**	**26.64**	**33.97**	**31.75**	**28.72**

(a) Gray image denoising comparisons.

Method	CBSD68			Kodak24			McMaster			Urban100		
	σ =15	σ =25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DRUNet	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
Restormer	34.39	31.78	28.59	35.32*	32.91*	29.86*	35.55	33.31	30.29	35.06	32.91	30.02
Xformer (ours)	**34.43**	**31.82**	**28.63**	**35.40**	**32.99**	**29.93**	**35.67**	**33.43**	**30.37**	**35.29**	**33.19**	**30.35**

(b) PSNR (dB) comparisons on color image denoising.

Table 3: Quantitative comparisons for **learning a single model to solve various noise levels**. PSNR scores are reported in (a) and (b). The best results are **bolded**. * denotes results obtained by testing with officially provided pre-trained models.

performance across three noise levels on all provided benchmarks. It is informed that we use the officially provided models of Restormer to evaluate results on Kodak24. For both grayscale and color image denoising, our method can achieve promising performance. Compared to Restormer, our method has obvious performance gains, e.g., +0.39 dB on Gaussian grayscale image denoising with noise level σ=50. It is worth mentioning that our Xformer has the comparable model parameters and computational cost with Restormer. In conclusion, experimental results validate that our proposed method enjoys the superiority and robustness to solve the synthetic noise removal tasks with various noise levels. Therefore, our Xformer is also good at handling a wide range of noise levels via a single model.

4 Additional Image Restoration Task

To further demonstrate the effectiveness of our proposed method, we train our model to solve another image restoration task. We train our Xformer on the task of motion deblurring. We keep the same training settings with the state-of–art method Restormer (Zamir et al., 2022). The specific settings of Xformer are not changed. We evaluate the model on the commonly-used datasets, which are Gopro (Nah et al., 2017), HIDE (Shen et al., 2019), and the real-world datasets (RealBlur-R (Rim et al., 2020) and RealBlur-J (Rim et al., 2020)). The detailed results are shown in Tab. 5.

Quantitative Comparisons. We compare our method to recent representative vision Transformer method Restormer. As shown in Tab. 5, our proposed method Xformer can achieve comparable performance with Restormer. Although our proposed method does not focus on the task of motion deblurring, it still has promising performance. It is because that our Xformer has strong ability to model global information in the Transformer network.

5 Additional Visual Results

In this section, we provide more visual comparisons on Gaussian color and grayscale image denoising tasks. We compare our approach to recent state-of-the-art methods. We show the visual results in Figures 3 and 4. Besides, we also analyze some failure cases of Xformer in Figure 5. The detailed comparisons are as follows.

Visual Comparisons. We show the visual results of color image denoising in Fig. 3. The compared methods include BM3D (Dabov et al., 2007), IRCNN (Zhang et al., 2017b), DnCNN (Zhang et al., 2017a), RNAN (Zhang et al., 2019), RDN (Zhang et al., 2020), SwinIR (Liang et al., 2021), and

Method	Params (M)	FLOPs (G)	CBSD68	Kodak24	McMaster	Urban100
AlternateNet	22.38	42.0	34.40	35.31	35.55	35.08
Xformer	25.23	42.2	**34.42**	**35.36**	**35.64**	**35.23**

Table 4: PSNR (dB) comparisons on 15-level Gaussian color image denoising. We compare the model AlternateNet with alternating STB and CTB to our Xformer. We calculate FLOPs while setting the input size to $3\times128\times128$.

Method	Gopro		HIDE		RealBlur-R		RealBlur-J	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Restormer*	32.92	0.9611	**31.22**	0.9423	**36.19**	0.9572	28.96	0.8786
Xformer	**33.06**	**0.9624**	31.19	**0.9424**	**36.19**	**0.9574**	**29.02**	**0.8829**

Table 5: PSNR (dB)/SSIM comparisons on motion deblurring task. We compare our method Xformer to recent representative method Restormer. * means that we test Restormer with the officially provided pre-trained model.

Restormer (Zamir et al., 2022). As we can see, our proposed Xformer can remove heavy noise and restore the high-frequency components. While, most of previous denoising methods suffer from blurring artifacts and missing details. Taking "img_085" and "img_087" as example, all the compared methods have difficulty in recovering parts of lines and textures while our Xformer can restore more details. Besides, we show the visual comparisons on gray image denoising in Fig. 4. We can find that our proposed method can achieve the best performance to recover clean and crisp images. Taking "img_031" and "img_057" as example, previous state-of-the-art methods fail to recover the obvious lines while our Xformer can do this. In conclusion, our proposed hybrid X-shaped Transformer-based network Xformer is good at solving image denoising tasks. Better visual results could be obtained by using our method. Besides, we provide some failure cases of Xformer in Figure 5. As we can see, Xformer fails to recover good details. As the provided low-quality images suffer from severe blurring, it is difficult for Xformer to restore perfect details.

6 Limitations and Future Work

In this work, we propose an X-shaped vision Transformer named Xformer for image denoising. We exploit the joint usage of spatial-wise self-attention and channel-wise self-attention mechanisms. Our proposed dual branches enable the network to capture patch-level and channel-level information respectively. Furthermore, the proposed BCU module brings the enhanced information fusion so that our method can achieve promising performance. In practice, we have not explored the further improvement about the modifications of specific self-attention modules. While the corresponding improvement is not specific to Xformer, it could be a common issue for existing Transformer-based networks, e.g., enlarging the global receptive field of window-based self-attention or introducing more contextual information to cross-covariance self-attention. Therefore, we will try these improving strategies in future work. Besides, the BCU module is important in our proposed method, which is simple but effective. We will also try more information fusion mechanisms in the future.

References

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *TIP*, 2007.

Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 1999.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021.

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.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020.

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

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022.

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

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

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017b.

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.

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

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 2020.

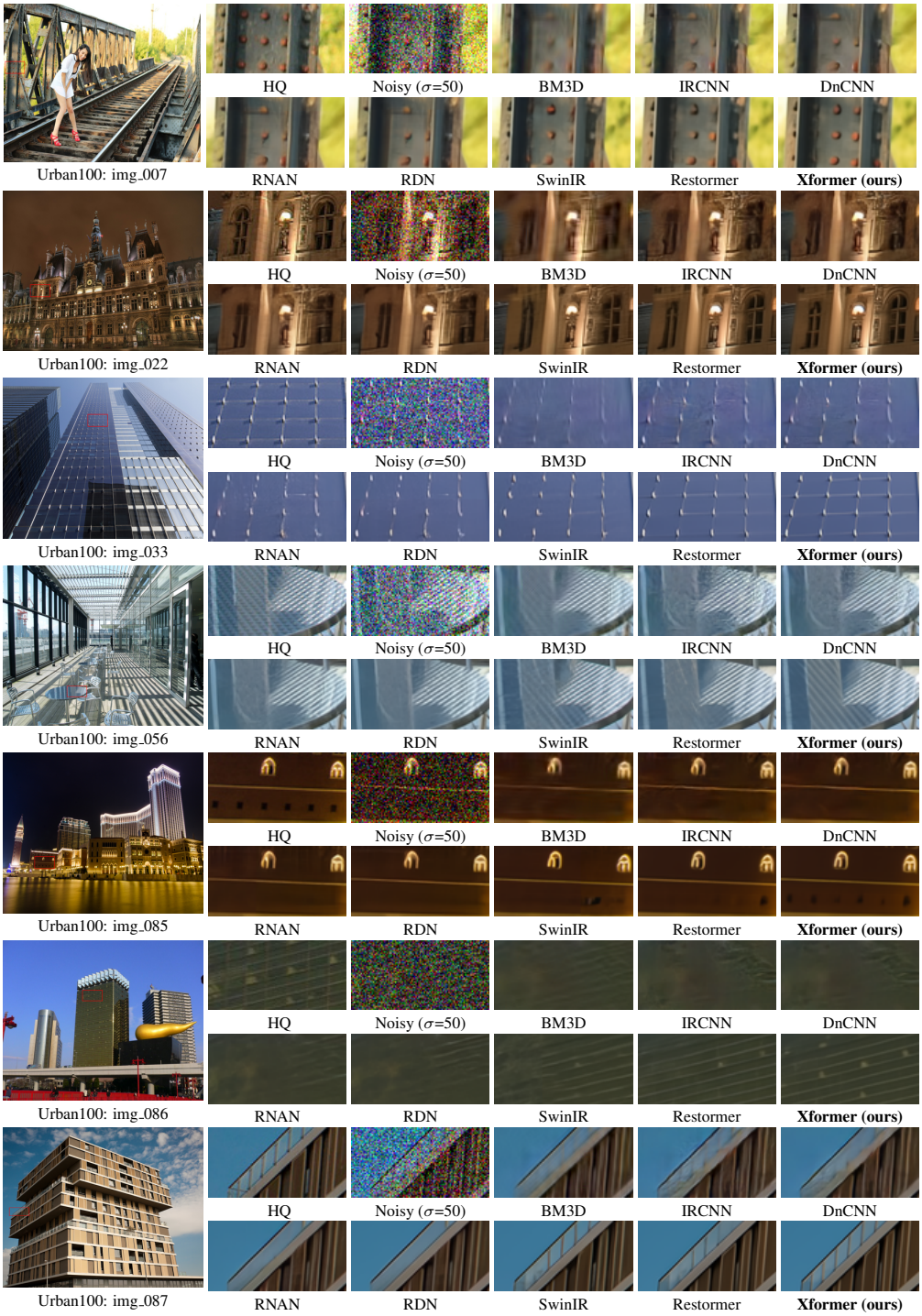

Figure 3: Visual comparisons with challenging examples on color image denoising (σ=50).

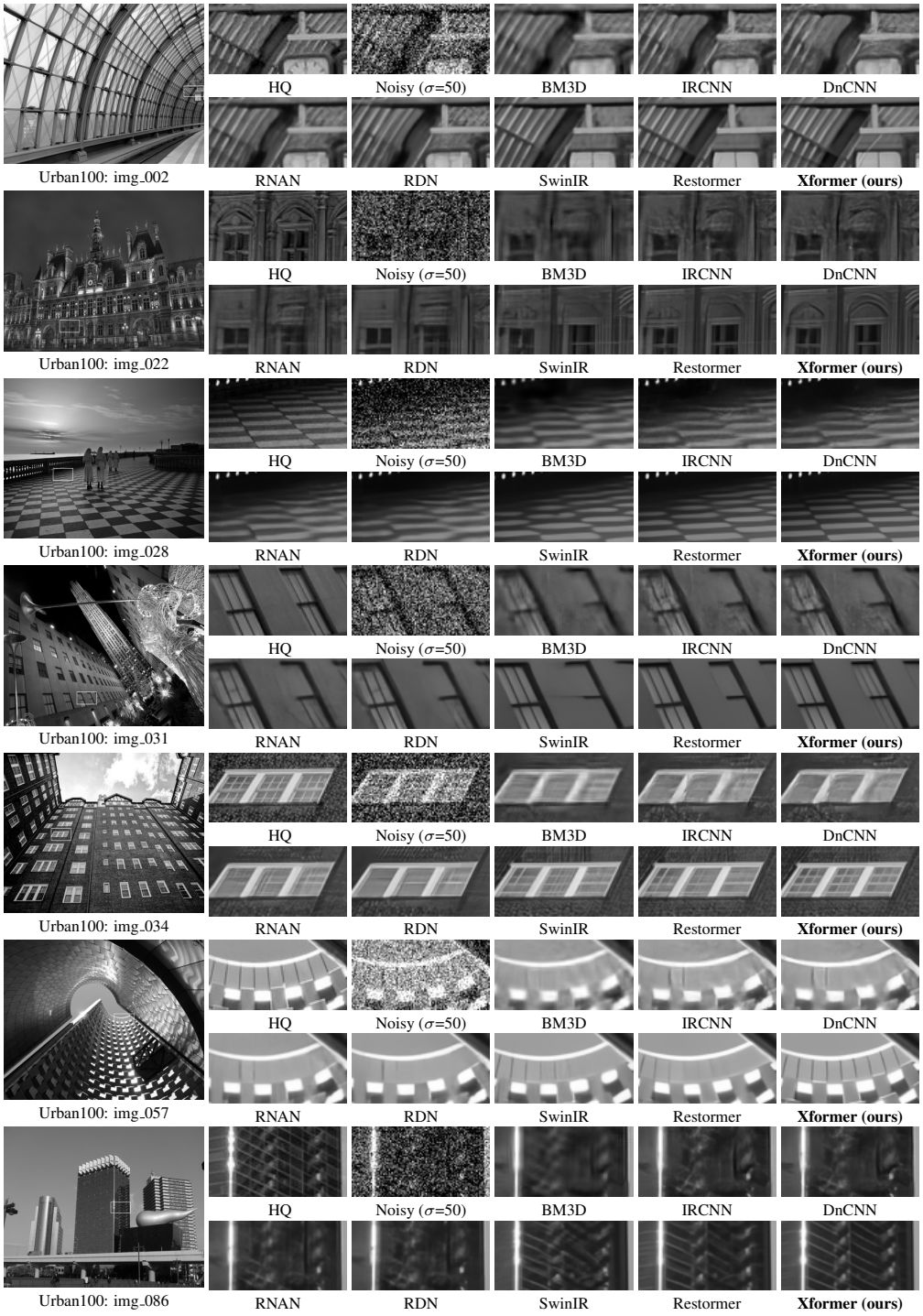

Figure 4: Visual comparisons with challenging examples on gray image denoising (σ=50).

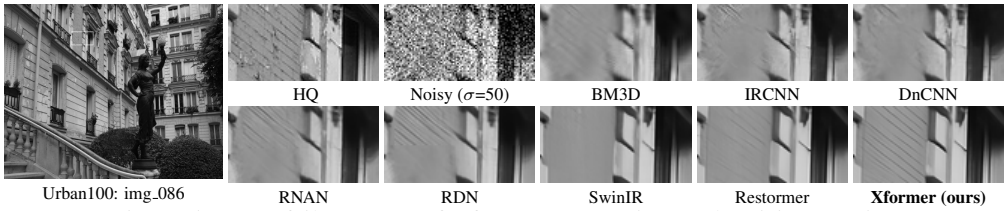

Figure 5: Some failure cases of Xformer on gary image denoising (σ=50).