SUPPLEMENTARY MATERIAL: ACCURATE IMAGE RESTORATION WITH ATTENTION RETRACTABLE TRANSFORMER

Jiale Zhang[1], **Yulun Zhang**[2]*, **Jinjin Gu**[3,4], **Yongbing Zhang**[5], **Linghe Kong**[1]*, **Xin Yuan**[6]
[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Shanghai AI Laboratory,
[4]The University of Sydney, [5]Harbin Institute of Technology (Shenzhen), [6]Westlake University

APPENDIX

Summary. This appendix provides some vital analyses and additional experimental results. Firstly, we present more experimental results about the ablation study in Sec. 1. Secondly, we give detailed discussion and verification about the differences between our method and the related works in Sec. 2. Thirdly, we provide the comparative results on two other tasks: real image denoising in Sec. 3 and Gaussian grayscale image denoising on Sec. 4. Lastly, we provide more quantitative and visual comparisons about our method in Sec. 5.

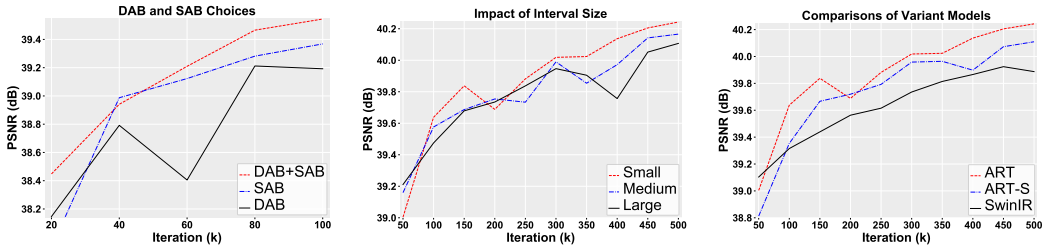

Figure 1: **Left:** PSNR (dB) comparison of our ART using all dense attention block (DAB), using all sparse attention block (SAB), and using alternating DAB and SAB. **Middle:** PSNR (dB) comparison of our ART using large interval size in sparse attention block which is $(8, 8, 8, 8, 8, 8)$ for six residual groups, using medium interval size which is $(8, 8, 6, 6, 4, 4)$, and using small interval size which is $(4, 4, 4, 4, 4, 4)$. **Right:** PSNR (dB) comparison of SwinIR, ART-S, and ART. Note that all the PSNR results are obtained by testing on another benchmark dataset Manga109 under $\times 2$ SR.

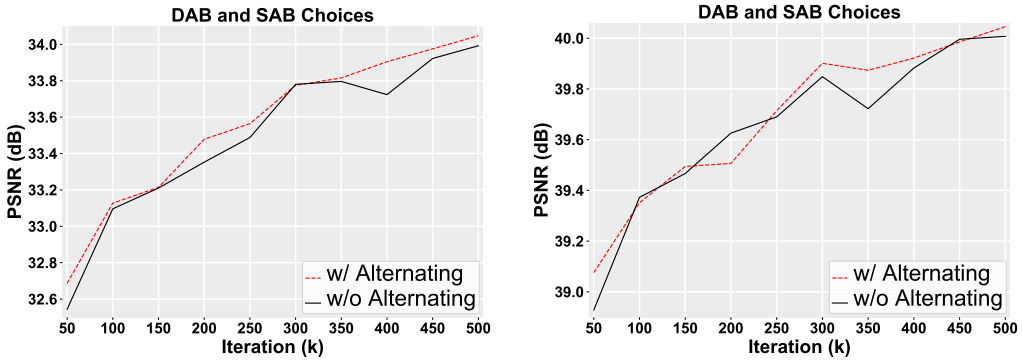

Figure 2: PSNR (dB) comparison of our ART using 3 pairs of alternating dense attention block (DAB) and sparse attention block (SAB), and using 3 successive SABs following 3 successive CABs. The former is with ("w/") alternating and the latter is without ("w/o") alternating. The **left** figure shows the testing results on Urban100 and the **right** one shows the testing results on Manga109.

*Corresponding authors: Yulun Zhang, yulun100@gmail.com; Linghe Kong, linghe.kong@sjtu.edu.cn

1 SUPPLEMENTARY ABLATION STUDY

We provide additional experimental results for ablation study. We train our models for ($\times$2) image super-resolution (SR) based on DIV2K Timofte et al. (2017) and Flicke2K Lim et al. (2017) datasets.

Design Choices for DAB and SAB. We set two different experiment conditions about the application of dense attention block (DAB) and sparse attention block (SAB). The first one is using 3 successive SABs following 3 successive CABs in each residual group module. The second one is using 3 pairs of alternating DAB and SAB. We keep else experiment setting the same and train these two models for 500k iterations. Fig. 2 shows the PSNR results on Urban100 and Manga109. We can see that the model with alternating DAB and SAB structure achieves better performance. It is mainly because that the alternating application of these two blocks enables residual group module to obtain local and global receptive field simultaneously. While, successive DABs and SABs structure has limited ability to capture wider receptive fields and thus shows poor performance. Besides, we provide more results to compare models with different blocks. As shown in Fig. 1(Left), the evaluation results on Manga109 also validate that the simultaneous usage of DAB and SAB is necessary.

Impact of Interval Size. We provide more comparisons about the models with different interval size settings. In detail, we evaluate the corresponding models that have been introduced in Ablation study of the main paper on Manga109. The results are shown in Fig. 1(Middle). As we can see, smaller intervals in our model bring more performance gains.

Comparisons of Variant Models. We provide a new version of our model for fair comparisons and name it ART-S. Different from ART, the MLP ratio in ART-S is set to 2 and the interval size is set to 8. We aim to keep the same computation cost with SwinIR in training phase. It is known that the input training image size is 64$\times$64 and the window size of SwinIR is 8. Therefore, our sparse attention module will not introduce additional computational cost by extracting the same 8$\times$8 tokens with SwinIR. In practice, our ART-S has the same Mult-Adds with SwinIR (e.g., 51.3G) when training for ($\times$2) image SR. We show PSNR comparison on Manga109 datasets as the training iterations increase in Fig. 1(Right). It further demonstrates that our ART-S outperforms SwinIR. Compared with SwinIR, our ART-S does not introduce additional computational cost in training phase but achieves better performance. It is validated to be new promising Transformer-based network.

2 DISCUSSION OF RELATED WORKS

As the core component of Transformer, self-attention module plays an important role in modeling long-range dependencies. Since our proposed method achieves promising performance, the sparse attention has been validated to be effective in dealing with low-level vision tasks. To demonstrate the differences, we compare our method to the usage of sparse attention in related works. Specially, we consider CrossFormer Wang et al. (2022a) as a representative work. We give detailed analysis and comparisons in the following parts.

2.1 COMPARISONS WITH CROSSFORMER

Inspired by recent local window self-attention scheme proposed by Swin Transformer Liu et al. (2021), CrossFormer proposed Cross-scale Embedding Layer (CEL) and Long Short Distance Attention (LSDA) for high-level vision tasks. In our paper, we proposed sparse attention modules for low-level vision tasks, e.g., image SR, denoising and image compression artifact reduction. We compare these two methods as follows.

Different Tasks. Different tasks have different requirements. The purpose of our method is to recover as more high-frequency information of original images as possible. Meanwhile, we directly forward the low-frequency information of LR images to the final HR outputs. Take image denoising as an example. We want to learn the real noise and remove it while retaining original image details. However, the purpose of CrossFormer is to make dense prediction. By contrast, the sparse attention modules in our network focus on pixel-level information while the long-distance attention modules in CrossFormer focus on the semantic-level information.

Different Model Architectures. CrossFormer has a pyramid structure, like most ViT backbones. The downsampling become a basic operation to reduce the computational cost. As the layer in CrossFormer becomes deeper, the interval of sparse tokens becomes smaller and the channel dimension becomes larger. This pyramid structure enables the whole model to focus on the semantic-level

Method	$scale$	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CrossFormer	$\times 2$	38.20	0.9613	34.04	0.9228	32.32	0.9017	32.91	0.9358	39.28	0.9785
ART-S	$\times 2$	**38.39**	**0.9622**	**34.33**	**0.9253**	**32.49**	**0.9038**	**33.70**	**0.9415**	**39.88**	**0.9800**

Table 1: Quantitative comparison (PSNR (dB)/SSIM) of CrossFormer and ART-S under $\times 2$ SR task with training 300k iterations, while the total training iterations are 500k

Method	MIRNet		Uformer-S		Uformer-B		Restormer		ART (ours)		ART+ (ours)	
Metrics	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SIDD	39.72	0.959	39.77	0.959	39.89	0.960	40.02	0.960	39.96	0.960	39.99	0.960
DND	39.88	0.956	39.96	0.956	40.04	0.956	40.03	0.956	40.05	0.956	40.08	0.956
Params	31.79M		20.63M		50.88M		26.13M		25.70M		25.70M	

Table 2: Quantitative comparison (PSNR (dB)/SSIM) with state-of-the-art methods for real image denoising. Best and second best results are colored with red and blue.

information. By contrast, our method does not use downsampling. We keep the size of feature map unchanged. It enables our model to focus on pixel-level information. Besides, we use long-distance residual connection to reserve low-frequency information.

Different Roles of Sparse Attention. CrossFormer uses long-distance attention to build interactions of multi-scale features. Our method uses sparse attention to build interactions of equal-scale features as we does not use cross-scale embedding layer. Compared with the extracted tokens in CrossFormer, the tokens from our sparse attention can represent the real pixel in original feature map. The interactions of tokens in our module achieve the real spatial sparsity. By contrast, the sparsity of long-distance attention in CrossFormer is temporary as the feature map gradually shrinks.

2.2 Experimental Verification

Furthermore, we provide comparative experiments to demonstrate that the model design of CrossFormer with CEL and LSDA is not suitable for low-level vision tasks. We want to validate that our model design is better. Detailed introductions are provided and experimental results are shown in Tab. 1 and Fig. 3(Left).

Experimental Settings. In detail, we use SwinIR Liang et al. (2021) as backbone and use the core components of CrossFormer to replace the modules in SwinIR. We change the pyramid structure of CrossFormer to make a fair comparison. Meanwhile, we set the related parameters of CrossFormer and ART the same. In detail, we set the window size and the interval size to 8. The number of Transformer blocks in each residual group are 6. We also use 6 successive residual groups. The MLP ratio is 2. We train these two models under super-resolution $\times 2$ task. We use DIV2K Timofte et al. (2017) and Flickr2K Lim et al. (2017) as training data, Set5 Bevilacqua et al. (2012), Set14 Zeyde et al. (2010), B100 Martin et al. (2001), Urban100 Huang et al. (2015), and Manga109 Matsui et al. (2017) as testing data, which are consistent with SwinIR.

Convergence Analyses. We train these two models for close to 300k iterations. In Fig. 3(Left), we show the validation curves of CrossFormer and our ART-S with validation dataset as Set5. We can see that the convergence of our ART-S is faster than CrossFormer. This result shows that better performance is achieved by our ART-S. Therefore, the CrossFormer is validated to be not suitable for low-level vision task.

Quantitative Comparisons. Table 1 shows the PSNR/SSIM of CrossFormer and our ART-S. We can see that CrossFormer achieves poor performance in solving image SR. The long-distance attention modules in CrossFormer focus on cross-scale features and learn more semantic-level information. However, the multi-scale features are not necessary in restoring HR images. By contrast, the sparse attention modules in our network focus on pixel-level information so that our method can achieve better performance.

3 Experiments on Real Image Denoising

We conduct experiments on another image restoration task to show the superiority of our method. In practice, we employ our method to solve Real Image Denoising problems. We give detailed introduction as follows.

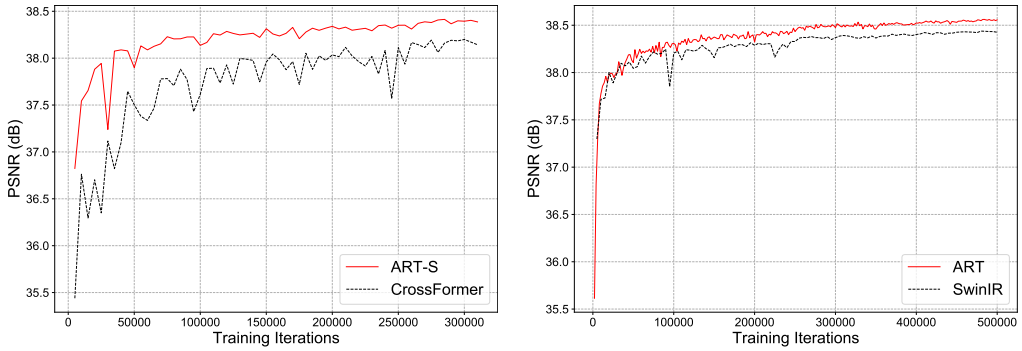

Figure 3: Convergence Analyses on $\times 2$ SR. **Left:** comparisons of CrossFormer and ART-S with training 300k iterations. **Right:** comparisons of SwinIR and ART with training 500k iterations.

Results for denoising on sRGB Data

Search: ART-300K

Name	Uses VST	Denoised on	PSNR on sRGB	SSIM on sRGB	Details
ART-300k-ensemble	0	srgb	40.0775	0.9558	
ART-300K	0	srgb	40.0462	0.9557	

Figure 4: Screenshot of the testing result of our ART in the DND testing website.

Experimental Settings. To make a fair comparison, we utilize the model architecture of Restormer Zamir et al. (2022) to train our ART. We use the proposed dense and sparse attention blocks to replace the Transformer blocks in Restormer. We keep other model components the same with Restormer. We also provide the implementation details. We use a 4-level symmetric encoder-decoder with the number of Transformer blocks as $[4, 6, 6, 8]$, from level-1 to level-4. The blocks number in refinement stage is 4. The number of attention heads is $[1, 2, 4, 8]$ and the MLP expansion ratio is 4 (2.66 in Restormer). We set the interval size in four successive encoders as (32, 16, 8, 4). For the dense attention block, we set the window size to 8.

We train our model under the same training conditions with Restormer. We use AdamW optimizer with β_1=0.9 and β_2=0.99. The initial learning rate is 3×10^{-4}. According to the cosine annealing Loshchilov & Hutter (2017), it gradually decreases to 1×10^{-6}. The model is trained within 300K iterations. Note that the progressive learning proposed by Restormer is also employed. We train ART on 320 high-resolution images from SIDD Abdelhamed et al. (2018) datasets. The testing datasets include SIDD test set and DND dataset Plotz & Roth (2017).

Quantitative Comparisons. We compare our ART with state-of-the-art methods including MIR-Net Zamir et al. (2020), Uformer Wang et al. (2022b), and Restormer. In Tab. 2, we can see that our proposed method has comparable performance with existing state-of-the-art models Uformer-B and Restormer. Higher performance is achieved by ART+ using self-ensemble. Besides, the parameter of our ART is smaller than all compared models except Uformer-S. It indicates that our method can also have promising performance in Real Image Denoising.

Further Analysis. In fact, our ART model has not been extended to larger model. If our ART has the same model size with Uformer (50.88M), the performance gains will be higher. To confirm the validity of our results, We also show the online test results of ART on the DND Plotz & Roth (2017) in Fig. 4. In conclusion, our proposed method shows strong ability to restore high-quality images in real image denoising.

4 Experiments on Gaussian Grayscale Image Denoising

We also employ our method to solve Gaussian grayscale image denoising task. The training datasets are same with the task of Gaussion color image denoising. We use Set12 Zhang et al. (2017a), BSD68 Martin et al. (2001), and Urban100 Huang et al. (2015) to perform evaluations. We compare our method to recent leading methods quantitatively and visually. We give more details as follows.

Dataset	σ	BM3D	DnCNN	IRCNN	FFDNet	RNAN	RDN	DRUNet	SwinIR	Restormer	ART	ART+
Set12	15	32.37	32.86	32.76	32.75	N/A	N/A	33.25	33.36	33.42	33.42	33.44
	25	29.97	30.44	30.37	30.43	N/A	N/A	30.94	31.01	31.08	31.10	31.12
	50	26.72	27.18	27.12	27.32	27.70	27.60	27.90	27.91	28.00	28.02	28.05
BSD68	15	31.08	31.73	31.63	31.63	N/A	N/A	31.91	31.97	31.96	31.97	32.00
	25	28.57	29.23	29.15	29.19	N/A	N/A	29.48	29.50	29.52	29.53	29.56
	50	25.60	26.23	26.19	26.29	26.48	26.41	26.59	26.58	26.62	26.60	26.65
Urban100	15	32.35	32.64	32.46	32.40	N/A	N/A	33.44	33.70	33.79	33.89	33.98
	25	29.70	29.95	29.80	29.90	N/A	N/A	31.11	31.30	31.46	31.68	31.78
	50	25.95	26.26	26.22	26.50	27.65	27.40	27.96	27.98	28.29	28.56	28.67

Table 3: PSNR (dB) comparisons for Gaussian grayscale image denoising on three benchmark datasets. The best and second best results are in red and blue.

Method	Parameters (M)	Mult-Adds (G)	PSNR on Urban100 (dB)
SwinIR Liang et al. (2021)	11.50	201	27.98
Restormer Zamir et al. (2022)	26.11	39	28.29
ART (ours)	20.82	44	28.56

Table 4: Model size comparisons. PSNR scores are reported by testing on Gaussian gray image denoising (σ=50). Input size is $1\times128\times128$ for Mult-Adds calculation.

Experimental Settings. We also utilize the model architecture of Restormer Zamir et al. (2022) to train our ART. We use U-net structure to design our model and train it on the codebase of Restormer. The training settings have been introduced in Sec. 3. We declare some differences here. We adjust some parameters in ART. We set the window size to 16 and the initial interval size to 8. As the stage grows, the interval size is reduced by half. To make fair comparisons, we keep the same layers number in encoder-decoder module and MLP expansion ratio with Restormer, which are [4,6,6,8] and 2.66, respectively.

Quantitative Comparisons. We compare our ART with state-of-the-art methods including BM3D Dabov et al. (2007), IRCNN Zhang et al. (2017b), FFDNet Zhang et al. (2018a), DnCNN Zhang et al. (2017a), RNAN Zhang et al. (2019), RDN Zhang et al. (2020), DRUNet Zhang et al. (2021a), SwinIR Liang et al. (2021), and Restormer Zamir et al. (2022). From Table 3, we can see that our proposed ART achieves better performance than existing state-of-the-art methods, including CNN-based and Transformer-based networks. Specifically, our ART yields 0.58 dB and 0.27 dB performance gain over SwinIR and Restormer, respectively, on Urban100 with noise level σ=50. Higher performance is achieved by ART+ using self-ensemble. It is worth mentioning that our ART has comparable Mult-Adds with Restormer but 1.25$\times$ fewer parameters. It indicates that our method can also have impressive performance in Gaussian grayscale mage Denoising.

Model Size Comparisons. Table 4 provides comparisons of parameters number and Mult-Adds. We mainly compare our ART to recent leading Transformer-based methods including SwinIR and Restormer. We calculate the Mult-Adds assuming that the input size is $1\times128\times128$. We find that our ART enjoys very low Mult-Adds when compared to SwinIR. Compared to Restormer, our method has comparable Mult-Adds but less model parameters. It is seen that our ART can achieve the best performance among them. It indicates that our method owns promising computational and memory efficiency while obtaining promising performance.

Visual Comparisons. We also provide numerous visual comparisons with recent state-of-the-art methods on Gaussian grayscale image denoising with noise level σ=50. We show these visual results in Fig. 8-10. As we can see, our proposed ART can obtain visually pleasing results when compared to other methods. Especially, compared to SwinIR and Restormer, our ART can restore more high-frequency components. It indicates that our ART with a U-Net structure can also achieve promising denoising results. In conclusion, our method share a similar model architecture with Restormer and owns comparable computational cost and less model parameters. However, our ART can outperform Restormer both quantitatively and visually.

5 Additional Experimental Results

We provide more quantitative and visual comparisons about our proposed model as follows.

Quantitative Comparisons. Table 5 shows quantitative comparisons for $\times2$, $\times3$, and $\times4$ image super-resolution (SR). All the results are provided by publicly available data. We compare our ART

with 16 state-of-art methods: EDSR Lim et al. (2017), SRMDNF Zhang et al. (2018b), D-DBPN Haris et al. (2018), RDN Zhang et al. (2020), RCAN Zhang et al. (2018c), SAN Dai et al. (2019), SRFBN Li et al. (2019), HAN Niu et al. (2020), IGNN Zhou et al. (2020), CSNLN Mei et al. (2020), RFANet Liu et al. (2020), NLSA Mei et al. (2021), CRAN Zhang et al. (2021b), DFSA Magid et al. (2021), IPT Chen et al. (2021), and SwinIR Liang et al. (2021). Symbol "+" means that results are produced with self-ensemble Lim et al. (2017) in test phase. Note that DFSA Magid et al. (2021) only provides self-ensemble scores. As we can see, our ART achieves the best performance on all the benchmark datasets across all scale factors. ART+ gains even better results using self-ensemble. Besides, ART-S has comparable mode size with SwinIR and also performs outstandingly. It is mainly because our proposed models have stronger representation ability than SwinIR.

Convergence Analyses. We provide the validation curve comparisons of our ART and SwinIR in Fig. 3(Right). We keep the training settings the same and the total training iterations are 500k. We report the validation results on Set5 based on the $\times 2$ SR task. As we can see, our ART achieves better performance than SwinIR. It further reveals that our proposed method has stronger global representation ability by using sparse attention. Thus, it outperforms other Transformer-based methods mainly using dense attention.

Visual Comparisons. We provide more visual comparisons. Figures 5 and 6 show some challenging examples for ($\times 4$) image super-resolution (SR). As we can see, most of previous state-of-the-art SR methods suffer from heavy blurring artifacts and have difficulty in recovering high-frequency details. Taking "image_008" and "image_076" as an example, all the compared methods have heavy blurring artifacts. While, our ART is able to alleviate blurring artifacts to some degree while recovering more details. Besides, in "image_047" and "image_059", our ART can reconstruct more detailed lines. However, all the compared methods are hard to deal with heavy blurring and thus fail to recover crisp lines. These comparisons indicate that previous leading SR methods have limited ability to restore high-quality images when handling some challenging cases. On the other hand, our proposed ART can obtain visually pleasing results and recover more high-frequency details.

Figure 7 shows visual comparisons of some challenging examples for color image denoising with noise level 50. As we can see, some previous denoising methods have difficulty in removing heavy noise corruption. While, our proposed ART can reserve detailed textures and high-frequency components after denoising. For example, in "image_008" and "image_060", all the compared methods suffer from heavy noise corruption. However, our ART can remove these noise corruption to some degree and restore clean and crisp images.

Figures 8-10 show visual comparisons of some challenging examples for grayscale image denoising with noise level 50. Results are obtained by testing on Urban100. We make comparisons with recent leading CNN-based and Transformer-based methods. As we can see, our ART can achieve better visual results.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 4

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012. 3

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. 6, 8

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *ICIP*, 2007. 5

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 6, 8

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 6, 8

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 3, 4

Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. 6, 8

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 3, 5, 6, 8

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2, 3, 6, 8

Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 6, 8

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 4

Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, 2021. 6, 8

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. 3, 4

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 2017. 3

Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 6, 8

Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 6, 8

Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 6, 8

Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 4

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 2, 3

Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *ICLR*, 2022a. 2

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022b. 4

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. 4

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. 4, 5

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. 7th Int. Conf. Curves Surf.*, 2010. 3

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. 4, 5

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017b. 5

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *TIP*, 2018a. 5

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018b. 6, 8

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *TPAMI*, 2021a. 5

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018c. 6, 8

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 5

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 2020. 5, 6, 8

Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *ICCV*, 2021b. 6, 8

Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, 2020. 6, 8

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
EDSR Lim et al. (2017)	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
D-DBPN Haris et al. (2018)	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
SRMDNF Zhang et al. (2018b)	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
RDN Zhang et al. (2020)	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
RCAN Zhang et al. (2018c)	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN Dai et al. (2019)	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
SRFBN Li et al. (2019)	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
HAN Niu et al. (2020)	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IGNN Zhou et al. (2020)	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
CSNLN Mei et al. (2020)	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
RFANet Liu et al. (2020)	×2	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
NLSA Mei et al. (2021)	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
CRAN Zhang et al. (2021b)	×2	38.31	0.9617	34.22	0.9232	32.44	0.9029	33.43	0.9394	39.75	0.9793
DFSA+ Magid et al. (2021)	×2	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
IPT Chen et al. (2021)	×2	38.37	N/A	34.43	N/A	32.48	N/A	33.76	N/A	N/A	N/A
SwinIR Liang et al. (2021)	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
ART-S (ours)	×2	38.48	0.9625	34.50	0.9258	32.53	0.9043	34.02	0.9437	40.11	0.9804
ART (ours)	×2	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
ART+ (ours)	×2	38.59	0.9630	34.68	0.9269	32.60	0.9050	34.41	0.9457	40.33	0.9810
Bicubic	×3	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349	26.95	0.8556
EDSR Lim et al. (2017)	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
SRMDNF Zhang et al. (2018b)	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN Zhang et al. (2020)	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
RCAN Zhang et al. (2018c)	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN Dai et al. (2019)	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
SRFBN Li et al. (2019)	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
HAN Niu et al. (2020)	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
IGNN Zhou et al. (2020)	×3	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
CSNLN Mei et al. (2020)	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
RFANet Liu et al. (2020)	×3	34.79	0.9300	30.67	0.8487	29.34	0.8115	29.15	0.8720	34.59	0.9506
NLSA Mei et al. (2021)	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
CRAN Zhang et al. (2021b)	×3	34.80	0.9304	30.73	0.8498	29.38	0.8124	29.33	0.8745	34.84	0.9515
DFSA+ Magid et al. (2021)	×3	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
IPT Chen et al. (2021)	×3	34.81	N/A	30.85	N/A	29.38	N/A	29.49	N/A	N/A	N/A
SwinIR Liang et al. (2021)	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
ART-S (ours)	×3	34.98	0.9318	30.94	0.8530	29.45	0.8146	29.86	0.8830	35.22	0.9539
ART (ours)	×3	35.07	0.9325	31.02	0.8541	29.51	0.8159	30.10	0.8871	35.39	0.9548
ART+ (ours)	×3	35.11	0.9327	31.05	0.8545	29.53	0.8162	30.22	0.8883	35.51	0.9552
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
EDSR Lim et al. (2017)	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
D-DBPN Haris et al. (2018)	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
SRMDNF Zhang et al. (2018b)	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
RDN Zhang et al. (2020)	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
RCAN Zhang et al. (2018c)	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN Dai et al. (2019)	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
SRFBN Li et al. (2019)	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
HAN Niu et al. (2020)	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IGNN Zhou et al. (2020)	×4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
CSNLN Mei et al. (2020)	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
RFANet Liu et al. (2020)	×4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.9187
NLSA Mei et al. (2021)	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
CRAN Zhang et al. (2021b)	×4	32.72	0.9012	29.01	0.7918	27.86	0.7460	27.13	0.8167	31.75	0.9219
DFSA+ Magid et al. (2021)	×4	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	0.9266
IPT Chen et al. (2021)	×4	32.64	N/A	29.01	N/A	27.82	N/A	27.26	N/A	N/A	N/A
SwinIR Liang et al. (2021)	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
ART-S (ours)	×4	32.86	0.9029	29.09	0.7942	27.91	0.7489	27.54	0.8261	32.13	0.9263
ART (ours)	×4	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
ART+ (ours)	×4	33.07	0.9055	29.20	0.7964	27.99	0.7513	27.89	0.8339	32.45	0.9291

Table 5: PSNR (dB)/SSIM comparisons for image super-resolution on five benchmark datasets. We color best and second best results in red and blue.

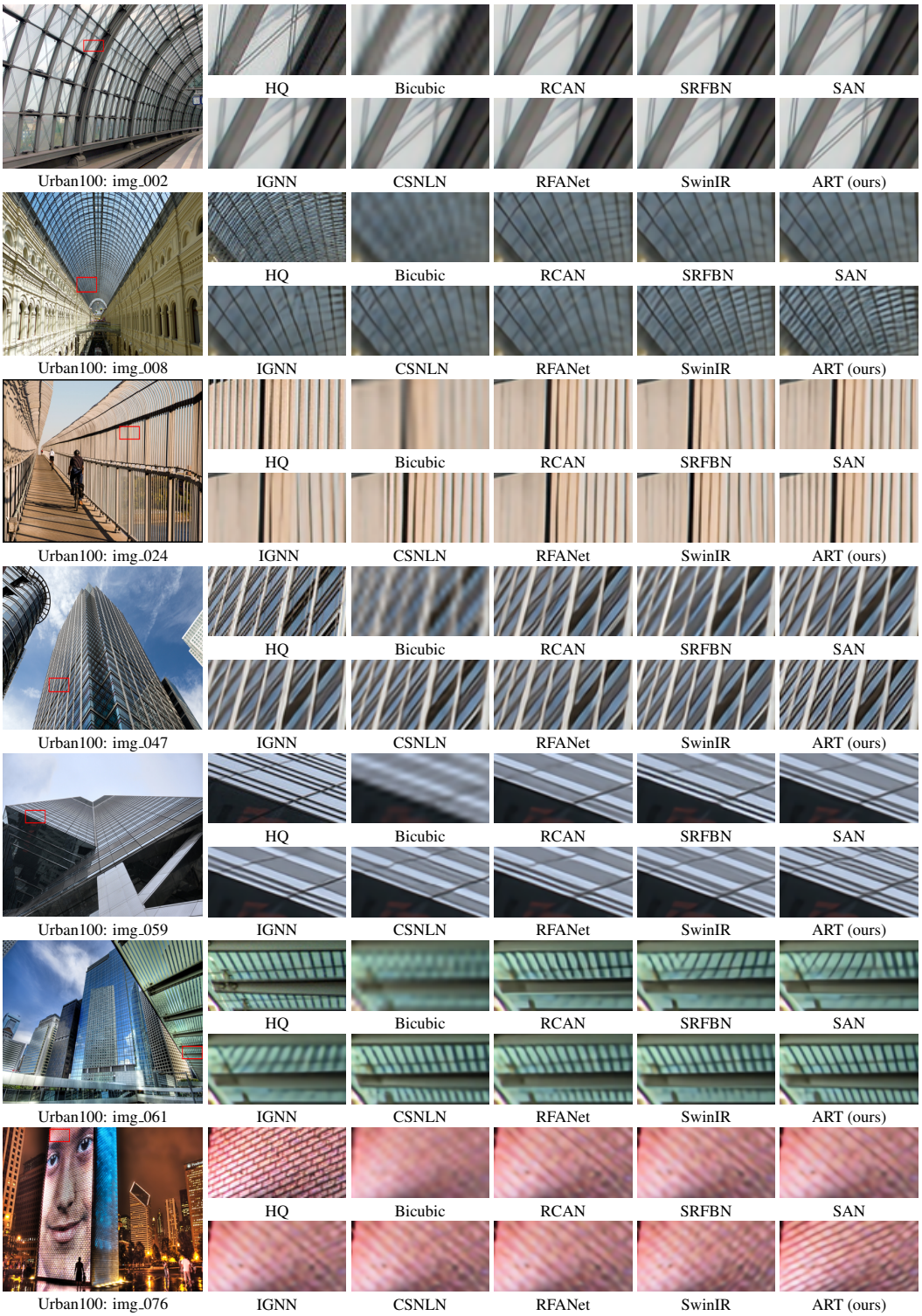

Figure 5: Visual comparison ($\times 4$) with image SR networks on Urban100 dataset.

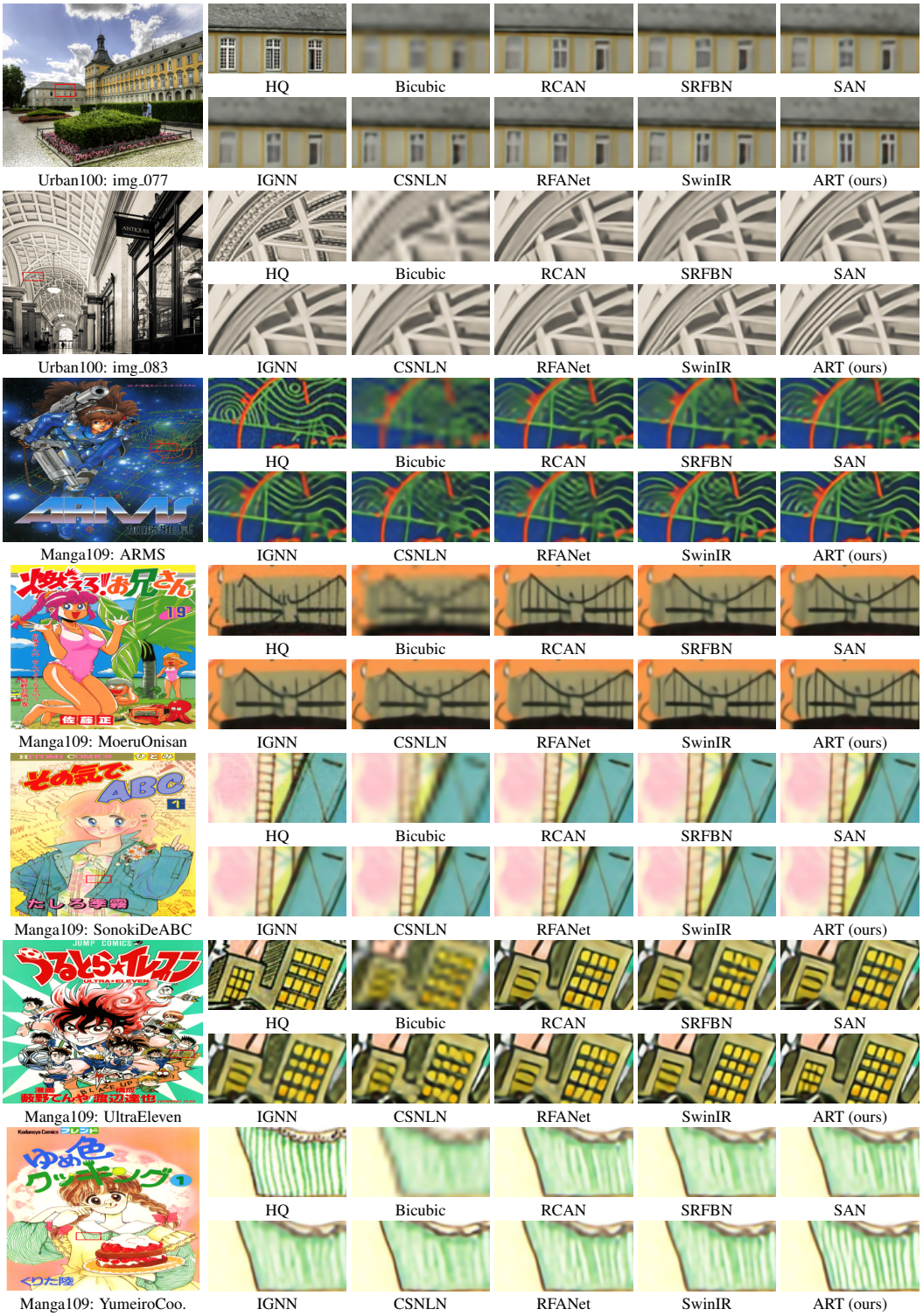

Figure 6: Visual comparison (×4) with image SR networks on Urban100 and Manga109 dataset.

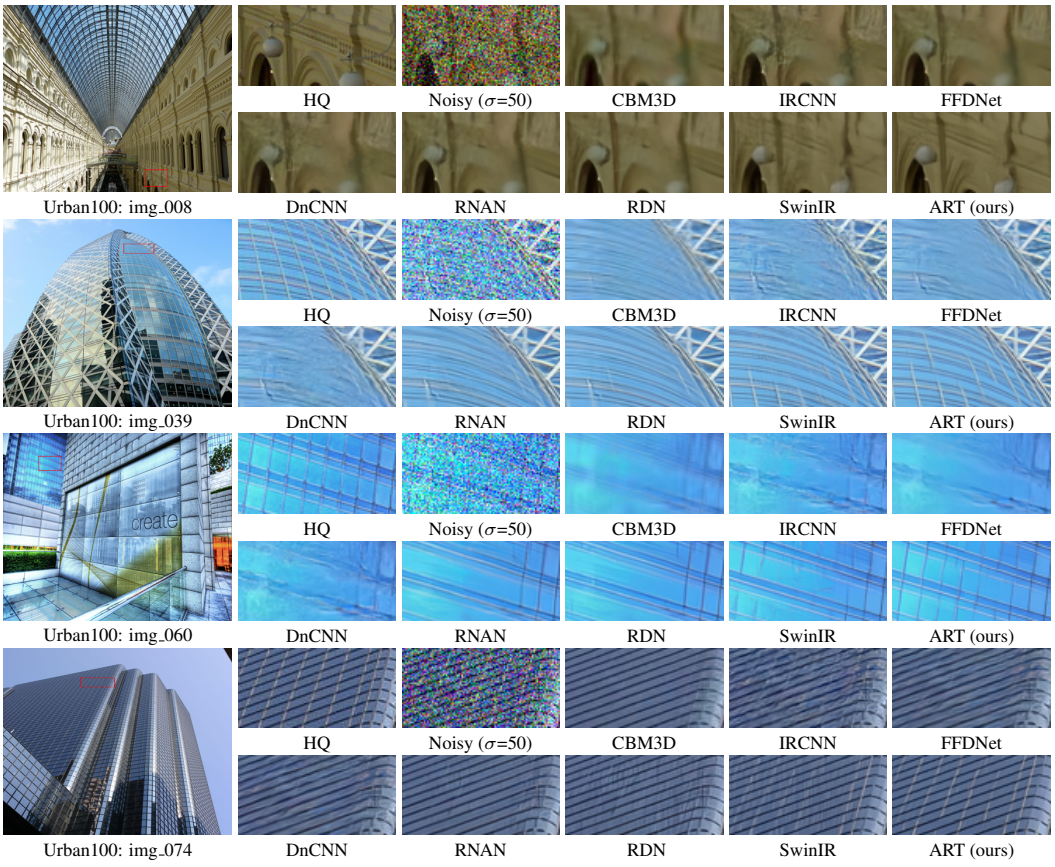

Figure 7: Visual comparison (σ=50) for Gaussian color image denoising on Urban100 dataset.

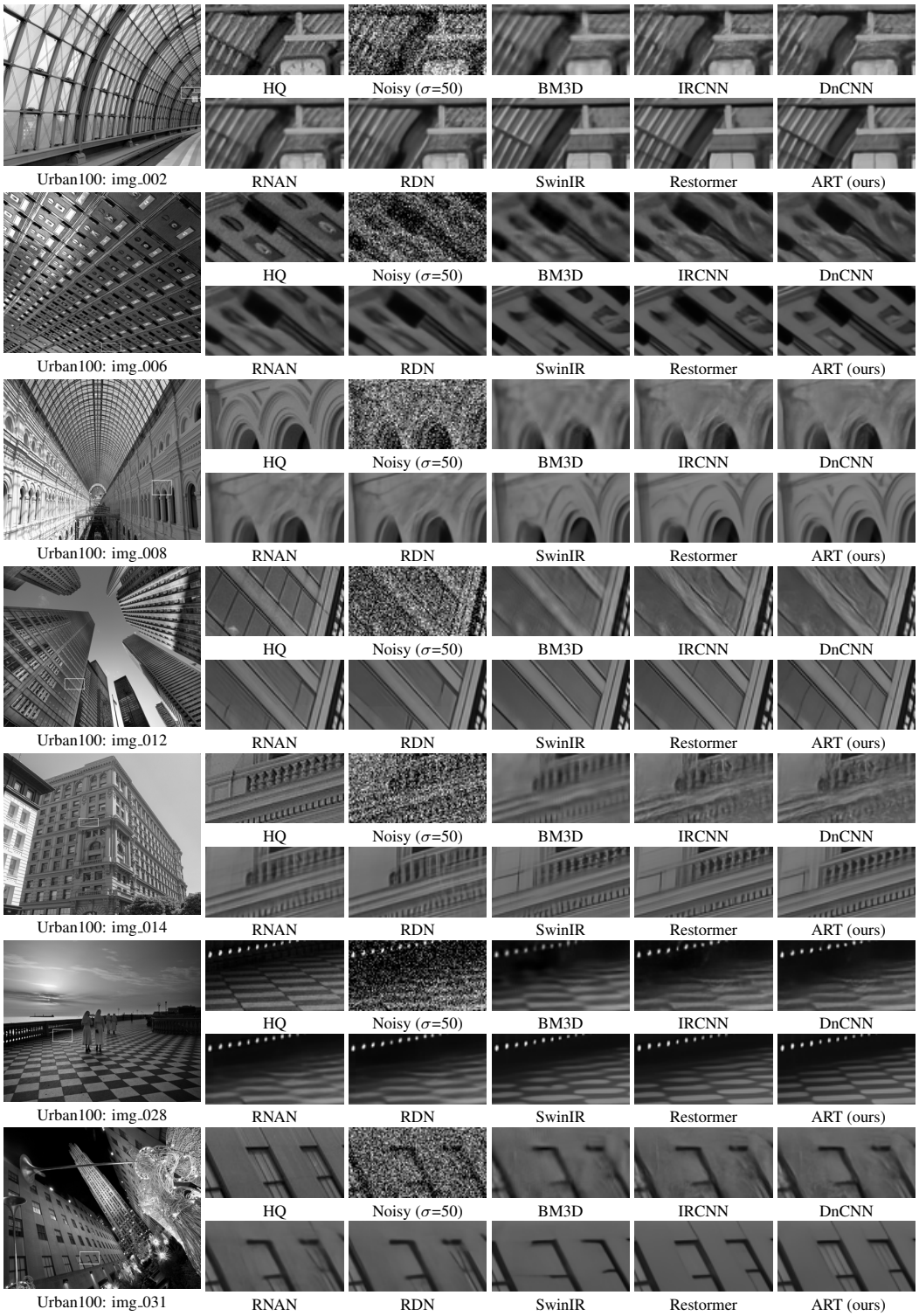

Figure 8: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.

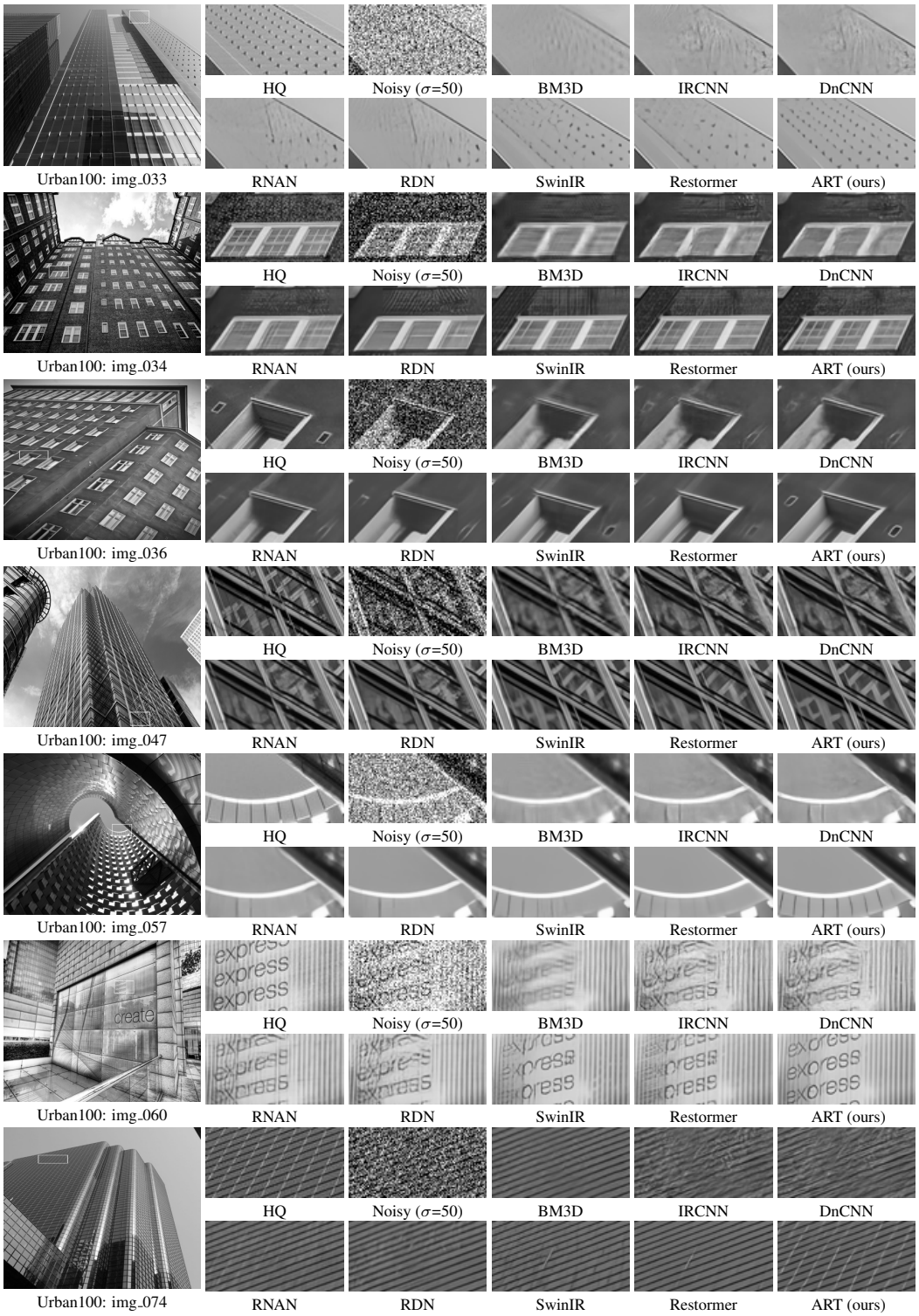

Figure 9: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.

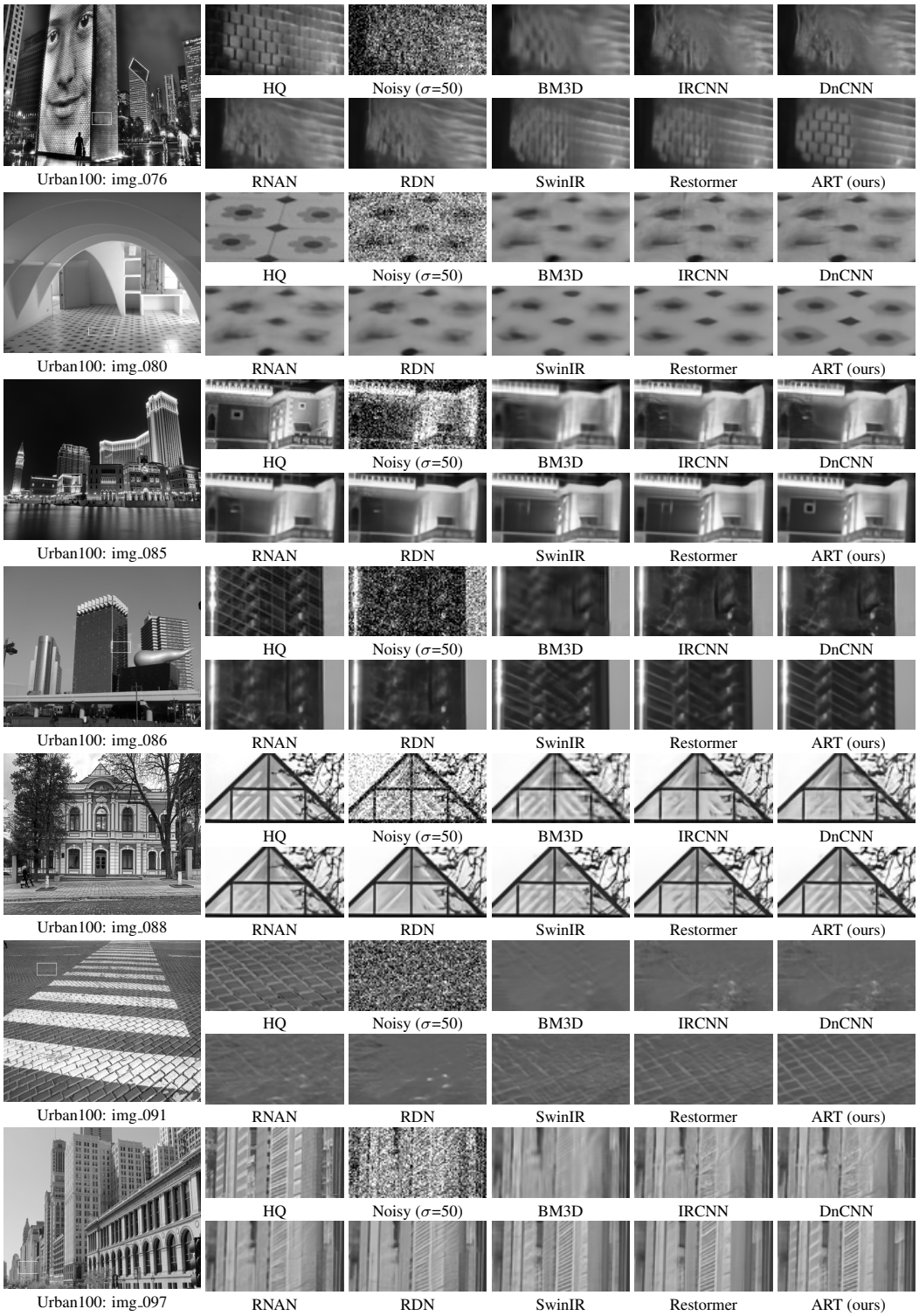

Figure 10: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.