ACCURATE IMAGE RESTORATION WITH ATTENTION RETRACTABLE TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Transformer-based image restoration networks have achieved promising improvements over convolutional neural networks due to parameter-independent global interactions. To lower computational cost, existing works generally limit self-attention computation within non-overlapping windows. However, each group of tokens are always from a dense area of the image. This is considered as a dense attention strategy since the interactions of tokens are restrained in dense regions. Obviously, this strategy could result in restricted receptive fields. To address this issue, we propose **A**ttention **R**etractable **T**ransformer (ART) for image restoration, which presents both dense and sparse attention modules in the network. The sparse attention module allows tokens from sparse areas to interact and thus provides a wider receptive field. Furthermore, the alternating application of dense and sparse attention modules greatly enhances representation ability of Transformer while providing retractable attention on the input image.We conduct extensive experiments on image super-resolution, denoising, and JPEG compression artifact reduction tasks. Experimental results validate that our proposed ART outperforms state-of-the-art methods on various benchmark datasets both quantitatively and visually. We also provide code and models at an anonymous website [1].

1 INTRODUCTION

Image restoration aims to recover the high-quality image from its low-quality counterpart and includes a series of computer vision applications, such as image super-resolution (SR) and denoising. It is an ill-posed inverse problem since there are a huge amount of candidates for any original input. Recently, deep convolutional neural networks (CNNs) have been investigated to design various models Kim et al. (2016b); Zhang et al. (2020; 2021b) for image restoration. SRCNN Dong et al. (2014) firstly introduced deep CNN into image SR. Then several representative works utilized residual learning (e.g., EDSR Lim et al. (2017)) and attention mechanism (e.g., RCAN Zhang et al. (2018c)) to train very deep network in image SR. Meanwhile, a number of methods were also proposed for image denoising such as DnCNN Zhang et al. (2017a), RPCNN Xia & Chakrabarti (2020), and BRDNet Tian et al. (2020). These CNN-based networks have achieved remarkable performance.

However, due to parameter-dependent receptive field scaling and content-independent local interactions of convolutions, CNN has limited ability to model long-range dependencies. To overcome this limitation, recent works have begun to introduce self-attention into computer vision systems Hu et al. (2019); Ramachandran et al. (2019); Wang et al. (2020); Zhao et al. (2020). Since Transformer has been shown to achieve state-of-the-art performance in natural language processing Vaswani et al. (2017) and high-level vision tasks Dosovitskiy et al. (2021); Touvron et al. (2021); Wang et al. (2021); Zheng et al. (2021); Chu et al. (2021), researchers have been investigating Transformer-based image restoration networks Yang et al. (2020); Kumar et al. (2021); Wang et al. (2022b). Chen et al. proposed a pre-trained image processing Transformer named IPT Chen et al. (2021a). Liang et al. proposed a strong baseline model named SwinIR Liang et al. (2021) based on Swin Transformer Liu et al. (2021) for image restoration. Zamir et al. also proposed an efficient Transformer model using U-net structure named Restormer Zamir et al. (2022) and achieved state-of-the-art results on several image restoration tasks. In contrast, higher performance can be achieved when using Transformer.

[1] https://anonymous.4open.science/r/ART_attention_retractable_transformer

Despite showing outstanding performance, existing Transformer backbones for image restoration still suffer from serious defects. As we know, SwinIR Liang et al. (2021) takes advantage of shifted window scheme to limit self-attention computation within non-overlapping windows. On the other hand, IPT Chen et al. (2021a) directly splits features into $P \times P$ patches to shrink original feature map P^2 times, treating each patch as a token. In short, these methods compute self-attention with shorter token sequences and the tokens in each group are always from a dense area of the image. It is considered as a dense attention strategy, which obviously causes a restricted receptive field. To address this issue, the sparse attention strategy is employed. We extract each group of tokens from a sparse area of the image to provide interactions like previous studies (e.g., GG-Transformer Yu et al. (2021), MaxViT Tu et al. (2022b), CrossFormer Wang et al. (2022a)), but different from them. Our proposed sparse attention module focuses on equal-scale features. Besides, We pay more attention to pixel-level information than semantic-level information. Since the sparse attention has not been well proposed to solve the problems in low-level vision fields, our proposed method can bridge this gap.

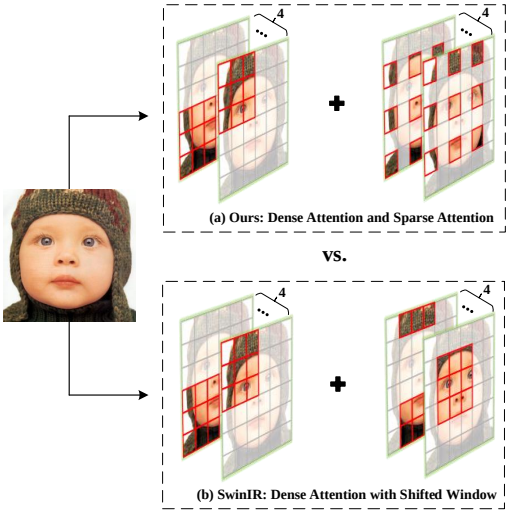


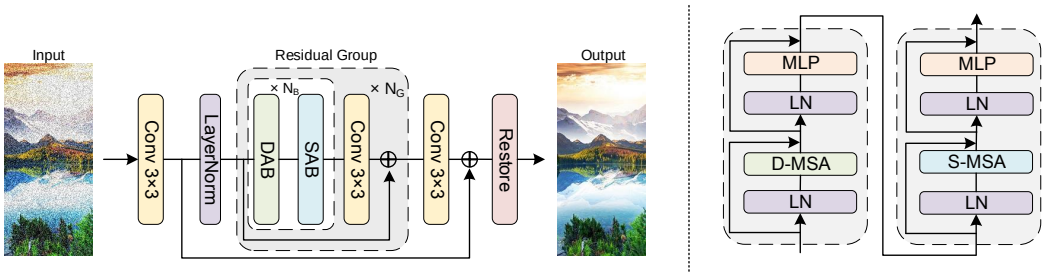
Figure 1: (**a**) Dense attention and sparse attention strategies of our ART. (**b**) Dense attention strategy with shifted window of SwinIR.

We further propose Attention Retractable Transformer named ART for image restoration. Following RCAN Zhang et al. (2018c) and SwinIR Liang et al. (2021), we reserve the residual in residual structure Zhang et al. (2018c) for model architecture. Based on joint dense and sparse attention strategies, we design two types of self-attention blocks. We utilize fixed non-overlapping local windows to obtain tokens for the first block named dense attention block (DAB) and sparse grids to obtain tokens for the second block named sparse attention block (SAB). To better understand the difference between our work and SwinIR, we show a visual comparison in Fig. 1. As we can see, the image is divided into four groups and tokens in each group interact with each other. Visibly, the token in our sparse attention block can learn relationships from farther tokens while the one in dense attention block of SwinIR cannot. At the same computational cost, the sparse attention block has stronger ability to compensate for the lack of global information. We consider our dense and sparse attention blocks as successive ones and apply them to extract deep feature. In practice, the alternating application of DAB and SAB can provide retractable attention for the model to capture both local and global receptive field. Our main contributions can be summarized as follows:

- We propose the sparse attention to compensate the defect of mainly using dense attention in existing Transformer-based image restoration networks. The interactions among tokens extracted from a sparse area of an image can bring a wider receptive field to the module.
- We further propose Attention Retractable Transformer (ART) for image restoration. Our ART offers two types of self-attention blocks to obtain retractable attention on the input feature. With the alternating application of dense and sparse attention blocks, the Transformer model can capture local and global receptive field simultaneously.
- We employ ART to train an effective Transformer-based network. We conduct extensive experiments on three image restoration tasks: image super-resolution, denoising, and JPEG compression artifact reduction. Our method achieves state-of-the-art performance.

2 Related Work

Image Restoration. With the rapid development of CNN, numerous works based on CNN have been proposed to solve image restoration problems Anwar & Barnes (2020); Dudhane et al. (2022); Zamir et al. (2020; 2021); Li et al. (2022); Chen et al. (2021b) and achieved superior performance over conventional restoration approaches Timofte et al. (2013); Michaeli & Irani (2013); He et al. (2010). The pioneering work SRCNN Dong et al. (2014) was firstly proposed for image SR. DnCNN Zhang et al. (2017a) was a representative image denoising method. Following these works, various model



(a) The architecture of ART for image restoration (b) Two successive attention blocks

Figure 2: (**a**) The architecture of our proposed ART for image restoration. (**b**) The inner structure of two successive attention blocks DAB and SAB with two attention modules D-MSA and S-MSA.

designs and improving techniques have been introduced into the basic CNN frameworks. These techniques include but not limit to the residual structure Kim et al. (2016a); Zhang et al. (2021a), skip connection Zhang et al. (2018c; 2020), dropout Kong et al. (2022), and attention mechanism Dai et al. (2019); Niu et al. (2020). Recently, due to the limited ability of CNN to model long-range dependencies, researchers have started to replace convolution operator with pure self-attention module for image restoration Yang et al. (2020); Liang et al. (2021); Zamir et al. (2022); Chen et al. (2021a).

Vision Transformer. Transformer has been achieving impressive performance in machine translation tasks Vaswani et al. (2017). Due to the content-dependent global receptive field, it has been introduced to improve computer vision systems in recent years. Dosovitskiy et al. Dosovitskiy et al. (2021) proposed ViT and introduced Transformer into image recognition by projecting large image patches into token sequences. Tu et al. proposed MaxViT Tu et al. (2022b) as an efficient Vision Transformer while introducing multi-axis attention. Wang et al. proposed CrossFormer Wang et al. (2022a) to build the interactions among long and short distance tokens. Yu et al. proposed GG-Transformer Yu et al. (2021), which performed self-attention on the adaptively-dilated partitions of the input. Inspired by the strong ability to learn long-range dependencies, researches have also investigated the usage of Transformer for low-level vision tasks Yang et al. (2020); Chen et al. (2021a); Liang et al. (2021); Zamir et al. (2022); Wang et al. (2022b). However, existing works still suffer from restricted receptive fields due to mainly using dense attention strategy. Very recently, Tu et al. proposed a MLP-based network named MAXIM Tu et al. (2022a) to introduce dilated spatial communications into image processing. It further demonstrates that the sparse interactions of visual elements are important for solving low-level problems. In our proposed method, we use dense and sparse attention strategies to build network, which can capture wider global interactions. As the sparse attention has not been well proposed to solve the low-level vision problems, our proposed method can bridge this gap.

3 Proposed Method

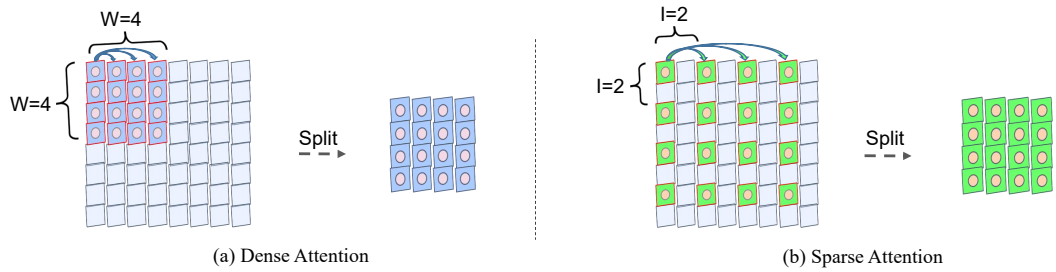
3.1 Overall Architecture

The overall architecture of our ART is shown in Fig. 2. Following RCAN Zhang et al. (2018c), ART employs residual in residual structure to construct a deep feature extraction module. Given a degraded image $I_{LQ} \in \mathbb{R}^{H\times D\times C_{in}}$ (H, D, and C_{in} are the height, width, and input channels of the input), ART firstly applies a 3×3 convolutional layer (Conv) to obtain shallow feature $F_0 \in \mathbb{R}^{H\times D\times C}$, where C is the dimension size of the new feature embedding. Next, the shallow feature is normalized and fed into the residual groups, which consist of core Transformer attention blocks. The deep feature is extracted and then passes through another 3×3 Conv to get further feature embeddings F_1. Then we use element-wise sum to obtain the final feature map $F_R = F_0 + F_1$. Finally, we employ the restoration module to generate the high-quality image I_{HQ} from the feature map F_R.

Residual Group. We use N_G successive residual groups to extract the deep feature. Each residual group consists of N_B pairs of attention blocks. We design two successive attention blocks shown in Fig. 2(b). The input feature x_{l-1} passes through layer normalization (LN) and multi-head self-attention (MSA). After adding the shortcut, the output x'_l is fed into the multi-layer perception (MLP). x_l is the final output at the l-th block. The process is formulated as

$$\begin{aligned} x'_l &= \mathrm{MSA}(\mathrm{LN}(x_{l-1})) + x_{l-1}, \\ x_l &= \mathrm{MLP}(\mathrm{LN}(x'_l)) + x'_l. \end{aligned} \tag{1}$$

Lastly, we also apply a 3×3 convolutional layer to refine the feature embeddings. As shown in Fig 2(a), a residual connection is employed to obtain the final output in each residual group module.



(a) Dense Attention (b) Sparse Attention

Figure 3: (**a**) Dense attention strategy. Tokens of each group are from a dense area of the image. (**b**) Sparse attention strategy. Tokens of each group are from a sparse area of the image.

Restoration Module. The restoration module is applied as the last stage of the framework to obtain the reconstructed image. As we know, image restoration tasks can be divided into two categories according to the usage of upsampling. For image super-resolution, we take advantage of the sub-pixel convolutional layer Shi et al. (2016) to upsample final feature map F_R. Next, we use a convolutional layer to get the final reconstructed image I_{HQ}. The whole process is formulated as

$$I_{HQ} = \mathrm{Conv}(\mathrm{Upsample}(F_R)). \tag{2}$$

For tasks without upsampling, such as image denoising, we directly use a convolutional layer to reconstruct the high-quality image. Besides, we add the original image to the last output of restoration module for better performance. We formulate the whole process as

$$I_{HQ} = \mathrm{Conv}(F_R) + I_{LQ}. \tag{3}$$

Loss Function. We optimize our ART with two types of loss functions. There are various well-studied loss functions, such as L_2 loss Dong et al. (2016); Sajjadi et al. (2017); Tai et al. (2017), L_1 loss Lai et al. (2017); Zhang et al. (2020), and Charbonnier loss Charbonnier et al. (1994). Same with previous works Zhang et al. (2018c); Liang et al. (2021), we utilize L_1 loss for image super-resolution (SR) and Charbonnier loss for image denoising and compression artifact reduction. For image SR, the goal of training ART is to minimize the L_1 loss function, which is formulated as

$$\mathcal{L} = \|I_{HQ} - I_G\|_1, \tag{4}$$

where I_{HQ} is the output of ART and I_G is the ground-truth image. For image denoising and JPEG compression artifact reduction, we utilize Charbonnier loss with super-parameter ε as 10^{-3}, which is

$$\mathcal{L} = \sqrt{\|I_{HQ} - I_G\|^2 + \varepsilon^2}. \tag{5}$$

3.2 Attention Retractable Transformer

We elaborate the details about our proposed two types of self-attention blocks in this section. As plotted in Fig. 2(b), the interactions of tokens are concentrated on the multi-head self-attention module (MSA). We formulate the calculation process in MSA as

$$\mathrm{MSA}(X) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{C}})V, \tag{6}$$

where $Q, K, V \in \mathbb{R}^{N \times C}$ are respectively the query, key, and value from the linear projecting of input $X \in \mathbb{R}^{N \times C}$. N is the length of token sequence, and C is the dimension size of each token. Here we assume that the number of heads is 1 to transfer MSA to singe-head self-attention for simplification.

Multi-head Self Attention. Given an image with size $H \times D$, vision Transformer firstly splits the raw image into numerous patches. These patches are projected by convolutions with stride size P. The new projected feature map $\hat{X} \in \mathbb{R}^{h \times w \times C}$ is prepared with $h = \frac{H}{P}$ and $w = \frac{D}{P}$. Common MSA uses all the tokens extracted from the whole feature map and sends them to self-attention module to learn relationships between each other. It will suffer from high computational cost, which is

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C. \tag{7}$$

To lower the computational cost, existing works generally utilize non-overlapping windows to obtain shorter token sequences. However, they mainly consider the tokens from a dense area of an image. Different from them, we propose the retractable attention strategies, which provide interactions of tokens from not only dense areas but also sparse areas of an image to obtain a wider receptive field.

Dense Attention. As shown in Fig. 3(a), dense attention allows each token to interact with a smaller number of tokens from the neighborhood position of a non-overlapping $W \times W$ window. All tokens

Methods	Solving problems	Structure	Interval of extracted tokens	Representation of tokens	Using long-distance residual connection
GG-Transformer Yu et al. (2021)	High-level	Pyramid	Changed	Semantic-level	No
MaxViT Tu et al. (2022b)	High-level	Pyramid	Changed	Semantic-level	No
CrossFormer Wang et al. (2022a)	High-level	Pyramid	Changed	Semantic-level	No
ART (Ours)	Low-level	Isotropic	Unchanged	Pixel-level	Yes

Table 1: Comparison to related works. The differences between our ART with other works.

are split into several groups and each group has $W \times W$ tokens. We apply these groups to compute self-attention for $\frac{h}{W} \times \frac{w}{W}$ times and the computational cost of new module named D-MSA is

$$\Omega(\text{D-MSA}) = (4W^2C^2 + 2W^4C) \times \frac{h}{W} \times \frac{w}{W} = 4hwC^2 + 2W^2hwC. \tag{8}$$

Sparse Attention. Meanwhile, as shown in Fig. 3(b), we propose sparse attention to allow each token to interact with a smaller number of tokens, which are from sparse positions with interval size I. After that, the updates of all tokens are also split into several groups and each group has $\frac{h}{I} \times \frac{w}{I}$ tokens. We further utilize these groups to compute self-attention for $I \times I$ times. We name the new multi-head self-attention module as S-MSA and the corresponding computational cost is

$$\Omega(\text{S-MSA}) = (4\frac{h}{I} \times \frac{w}{I}C^2 + 2(\frac{h}{I} \times \frac{w}{I})^2C) \times I \times I = 4hwC^2 + 2\frac{h}{I}\frac{w}{I}hwC. \tag{9}$$

By contrast, our proposed D-MSA and S-MSA modules have lower computational cost since $W^2 \ll hw$ and $\frac{h}{I}\frac{w}{I} < hw$. After computing all groups, the outputs are further merged to form original-size feature map. In practice, we apply these two attention strategies to design two types of self-attention blocks named as dense attention block (DAB) and sparse attention block (SAB) as plotted in Fig. 2.

Successive Attention Blocks. We propose the alternating application of these two blocks. As the local interactions have higher priority, we fix the order of DAB in front of SAB. Besides, we provide the long-distance residual connection between each three pairs of blocks. We show the effectiveness of this joint application with residual connection in Appendix Sec. A and Sec. B.

Attention Retractable Transformer. We demonstrate that the application of these two blocks enables our model to capture local and global receptive field simultaneously. We treat the successive attention blocks as a whole and get a new type of Transformer named Attention Retractable Transformer, which can provide interactions for both local dense tokens and global sparse tokens.

3.3 Differences to Related Works

We summarize the differences between our proposed approach, ART with the closely related works in Tab. 1. We conclude them as three points. **(1) Different tasks.** GG-Transformer Yu et al. (2021), MaxViT Tu et al. (2022b) and CrossFormer Wang et al. (2022a) are proposed to solve high-level vision problems. Our ART is the only one to employ the sparse attention in low-level vision fields. **(2) Different designs of sparse attention.** In the part of attention, GG-Transformer utilizes the adaptively-dilated partitions, MaxViT utilizes the fixed-size grid attention and CrossFormer utilizes the cross-scale long-distance attention. As the layers get deeper, the interval of tokens from sparse attention becomes smaller and the channels of tokens become larger. Therefore, each token learns more semantic-level information. In contrast, the interval and the channel dimension of tokens in our ART keep unchanged and each token represents the accurate pixel-level information. **(3) Different model structures.** Different from these works using Pyramid model structure, our proposed ART enjoys an Isotropic structure. Besides, we provide the long-distance residual connection between several Transformer encoders, which enables the feature of deep layers to reserve more low-frequency information from shallow layers. More detailed discussion can be found in the Appendix Sec. B.

3.4 Implementation Details

Some details about how to apply our ART to construct image restoration model are introduced here. Firstly, the residual group number, DAB number, and SAB number in each group are set as 6, 3, and 3. Secondly, all the convolutional layers are equipped with 3×3 kernel, 1-length stride, and 1-length padding, so the height and width of feature map remain unchanged. In practice, we treat 1×1 patch as a token. Besides, we set the channel dimension as 180 for most layers except for the shallow feature extraction and the image reconstruction process. Thirdly, the window size in DAB is set as 8 and the interval size in SAB is adjustable according to different tasks, which is discussed in Sec. 4.2. Lastly, to adjust the division of windows and sparse grids, we use padding and mask strategies to the input feature map of self-attention, so that the number of division is always an integer.

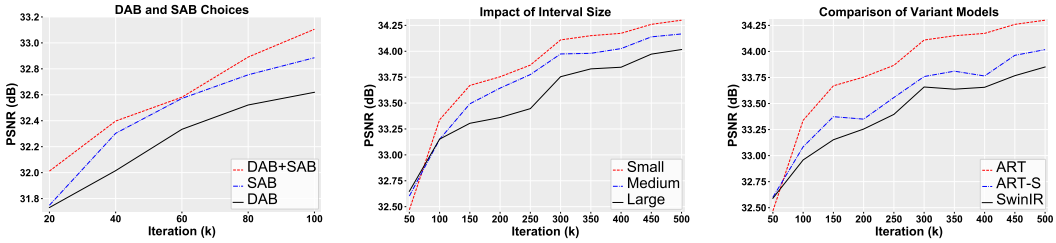


Figure 4: **Left:** PSNR (dB) comparison of our ART using all dense attention block (DAB), using all sparse attention block (SAB), and using alternating DAB and SAB. **Middle:** PSNR (dB) comparison of our ART using large interval size in sparse attention block which is $(8, 8, 8, 8, 8, 8)$ for six residual 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.

4 Experimental Results

4.1 Experimental Settings

Data and Evaluation. We conduct experiments on three image restoration tasks, including image SR, denoising, and JPEG Compression Artifact Reduction (CAR). For image SR, following previous works Zhang et al. (2018c); Haris et al. (2018), 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 test data. For image denoising and JPEG CAR, same as SwinIR Liang et al. (2021), we use training data: DIV2K, Flickr2K, BSD500 Arbelaez et al. (2010), and WED Ma et al. (2016). We use BSD68 Martin et al. (2001), Kodak24 Franzen (1999), McMaster Zhang et al. (2011), and Urban100 as test data of image denoising. Classic5 Foi et al. (2007) and LIVE1 Sheikh et al. (2006) are test data of JPEG CAR. Note that we crop large-size input image into 200×200 partitions with overlapping pixels during inference. Following Lim et al. (2017), we adopt the self-ensemble strategy to further improve the performance of our ART and name it as ART+. We evaluate experimental results with PSNR and SSIM Wang et al. (2004) values on Y channel of images transformed to YCbCr space.

Training Settings. Data augmentation is performed on the training data through horizontal flip and random rotation of $90°$, $180°$, and $270°$. Besides, we crop the original images into 64×64 patches as the basic training inputs for image SR, 128×128 patches for image denoising, and 126×126 patches for JPEG CAR. We resize the training batch to 32 for image SR, and 8 for image denoising and JPEG CAR in order to make a fair comparison. We choose ADAM Kingma & Ba (2015) to optimize our ART model with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and zero weight decay. The initial learning rate is set as 2×10^{-4} and is reduced by half as the training iteration reaches a certain number. Taking image SR as an example, we train ART for total 500k iterations and adjust learning rate to half when training iterations reach 250k, 400k, 450k, and 475k, where 1k means one thousand. Our ART is implemented on PyTorch Paszke et al. (2017) with 4 NVIDIA RTX8000 GPUs.

4.2 Ablation Study

For ablation experiments, we train our models for image super-resolution ($\times 2$) based on DIV2K and Flicke2K datasets. The results are evaluated on Urban100 benchmark dataset.

Design Choices for DAB and SAB. We demonstrate the necessity for simultaneous usage of dense attention block (DAB) and sparse attention block (SAB) by conducting ablation study. We set three different experiment conditions, which are using 6 DABs, 6 SABs, and 3 pairs of alternating DAB and SAB. We keep the rest of experiment environment the same and train all models within 100k iterations. The experimental results are shown in Fig. 4(Left). As we can see, only using DAB or SAB suffers from poor performance, because they lack either global receptive field or local receptive field. On the other hand, the structure of SAB following DAB brings higher performance. It validates that both local contextual interactions and global sparse interactions are important for improving strong representation ability of Transformer by obtaining retractable attention on the input feature.

Impact of Interval Size. The interval size in sparse attention block has a vital impact on the performance of our ART. In fact, if the interval size is set as 1, it will be transferred to full attention. Generally, a smaller interval means wider receptive fields but higher computational cost. We compare the experimental results under different interval settings in Fig. 4(Middle). As we can see, smaller

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
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
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
IPT Chen et al. (2021a)	×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
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
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
IPT Chen et al. (2021a)	×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
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
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
IPT Chen et al. (2021a)	×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 2: PSNR (dB)/SSIM comparisons for image super-resolution on five benchmark datasets. We color best and second best results in red and blue.

Method	EDSR	RCAN	SRFBN	HAN	CSNLN	SwiIR	ART-S (ours)	ART (ours)
Params (M)	43.09	15.59	3.63	16.07	7.16	11.90	11.87	16.55
Mult-Adds (G)	1,286	407	498	420	103,640	336	392	782
PSNR on Urban100 (dB)	26.64	26.82	26.60	26.85	27.22	27.45	27.54	27.77
PSNR on Manga109 (dB)	31.02	31.22	31.15	31.42	31.43	32.03	32.13	32.31

Table 3: Model size comparisons (×4 SR). Output size is 3×640×640 for Mult-Adds calculation.

intervals bring more performance gains. To keep the balance between accuracy and complexity, we set the interval size of 6 residual groups as $(4, 4, 4, 4, 4, 4)$ for image SR, $(16, 16, 12, 12, 8, 8)$ for image denoising, and $(18, 18, 13, 13, 7, 7)$ for JPEG CAR in the following comparative experiments.

Comparison 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 (4 in ART) and the interval size is set to 8. We demonstrate that ART-S has comparable model size with SwinIR. We provide the PSNR comparison results in Fig. 4(Right). As we can see, our ART-S achieves better performance than SwinIR. More comparative results can be found in following experiment parts.

4.3 Image Super-Resolution

We provide comparisons of our proposed ART with representative image SR methods, including CNN-based networks: EDSR Lim et al. (2017), 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), and Transformer-based networks: IPT Chen et al. (2021a) and SwinIR Liang et al. (2021). Note that IPT is a pre-trained model, which is trained on ImageNet benchmark dataset. All the results are provided by publicly available code and data. Quantitative and visual comparisons are provided in Tab. 2 and Fig. 5.

Quantitative Comparisons. We present PSNR/SSIM comparison results for ×2, ×3, and ×4 image SR in Tab. 2. As we can see, our ART achieves the best PSNR/SSIM performance on all five benchmark datasets. Using self-ensemble, ART+ gains even better results. Compared with existing state-of-the-art method SwinIR, our ART obtains better gains across all scale factors, indicating that

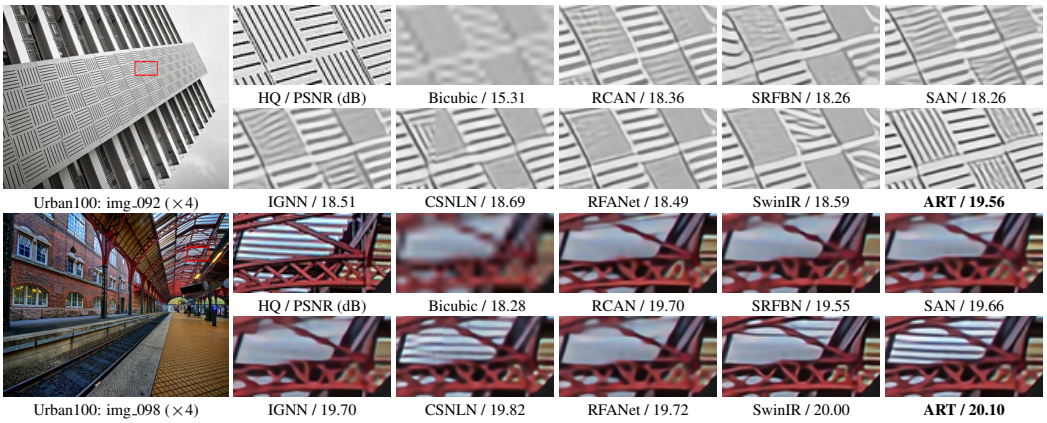


Figure 5: Visual comparison with challenging examples on image super-resolution ($\times 4$).

our proposed joint dense and sparse attention blocks enable Transformer stronger representation ability. Despite showing better performance than CNN-based networks, another Transformer-based network IPT is not as good as ours. It is validated that our proposed ART becomes a new promising Transformer-based network for image SR.

Retractable vs. Dense Attention. We further show a typical visual comparison with SwinIR in Fig. 6. As SwinIR mainly utilizes dense attention strategy, it restores wrong texture structures under the influence of close patches with mainly vertical lines. However, our ART can reconstruct the right texture, thanks to the wider receptive field provided by sparse attention strategy. Visibly, the patch is able to interact with farther patches with similar horizontal lines so that it can be reconstructed clearly. This comparison demonstrates the advantage of retractable attention and its strong ability to restore high-quality outputs.

Figure 6: Visual comparison ($\times 4$) of SwinIR and Ours.

Model Size Comparisons. Table 3 provides comparisons of parameters number and Mult-Adds of different networks, which include existing state-of-the-art methods. We calculate the Mult-Adds assuming that the output size is $3\times640\times640$ under $\times 4$ image SR. Compared with previous CNN-based networks, our ART has comparable parameter number and Mult-Adds but achieves high performance. Besides, we can see that our ART-S has less parameters and Mult-Adds than most of the compared methods. The model size of ART-S is similar with SwinIR. However, ART-S still achieves better performance gains than all compared methods except our ART. It indicates that our method is able to achieve promising performance at an acceptable computational and memory cost.

Visual Comparisons. We also provide some challenging examples for visual comparison ($\times 4$) in Fig. 5. We can see that our ART is able to alleviate heavy blurring artifacts while restoring detailed edges and textures. Compared with other methods, ART obtains visually pleasing results by recovering more high-frequency details. It indicates that ART preforms better for image SR.

4.4 Image Denoising

We show color image denoising results to compare our ART with representative methods in Tab. 4. These methods are CBM3D 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), IPT Chen et al. (2021a), DRUNet Zhang et al. (2021a), P3AN Hu et al. (2021), SwinIR Liang et al. (2021), and Restormer Zamir et al. (2022). Following most recent works, we set the noise level to 15, 25, and 50. We also shows visual comparisons of challenging examples in Fig. 7.

Quantitative Comparisons. Table 4 shows PSNR results of color image denoising. As we can see, our ART achieves the highest performance across all compared methods on three datasets except Kodak24. Even better results are obtained by ART+ using self-ensemble. Particularly, it obtains better gains than the state-of-the-art model Restormer Zamir et al. (2022) by up to 0.25dB on Urban100. Restormer also has restricted receptive fields and thus has difficulty in some challenging cases. In conclusion, these comparisons indicate that our ART also has strong ability in image denoising.

Method	BSD68			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
CBM3D Dabov et al. (2007)	N/A	N/A	27.38	N/A	N/A	28.63	N/A	N/A	N/A	N/A	N/A	27.94
IRCNN Zhang et al. (2017b)	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
FFDNet Zhang et al. (2018a)	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
DnCNN Zhang et al. (2017a)	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
RNAN Zhang et al. (2019)	N/A	N/A	28.27	N/A	N/A	29.58	N/A	N/A	29.72	N/A	N/A	29.08
RDN Zhang et al. (2020)	N/A	N/A	28.31	N/A	N/A	29.66	N/A	N/A	N/A	N/A	N/A	29.38
IPT Chen et al. (2021a)	N/A	N/A	28.39	N/A	N/A	29.64	N/A	N/A	29.98	N/A	N/A	29.71
DRUNet Zhang et al. (2021a)	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
P3AN Hu et al. (2021)	N/A	N/A	28.37	N/A	N/A	29.69	N/A	N/A	N/A	N/A	N/A	29.51
SwinIR Liang et al. (2021)	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Restormer Zamir et al. (2022)	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	35.13	32.96	30.02
ART (ours)	34.46	31.84	28.63	35.39	32.95	29.87	35.68	33.41	30.31	35.29	33.14	30.19
ART+ (ours)	34.47	31.85	28.65	35.41	32.98	29.89	35.71	33.44	30.35	35.34	33.20	30.27

Table 4: PSNR (dB) comparisons. The best and second best results are in red and blue.

Dataset	q	RNAN		RDN		DRUNet		SwinIR		ART (ours)		ART+ (ours)	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Classic5	10	29.96	0.8178	30.00	0.8188	30.16	0.8234	30.27	0.8249	30.27	0.8258	30.32	0.8263
	30	33.38	0.8924	33.43	0.8930	33.59	0.8949	33.73	0.8961	33.74	0.8964	33.78	0.8967
	40	34.27	0.9061	34.27	0.9061	34.41	0.9075	34.52	0.9082	34.55	0.9086	34.58	0.9089
LIVE1	10	29.63	0.8239	29.67	0.8247	29.79	0.8278	29.86	0.8287	29.89	0.8300	29.92	0.8305
	30	33.45	0.9149	33.51	0.9153	33.59	0.9166	33.69	0.9174	33.71	0.9178	33.74	0.9181
	40	34.47	0.9299	34.51	0.9302	34.58	0.9312	34.67	0.9317	34.70	0.9322	34.73	0.9324

Table 5: PSNR (dB)/SSIM comparisons. The best and second best results are in red and blue.

Visual Comparisons. The visual comparison for color image denoising of different methods is shown in Fig. 7. Our ART can preserve detailed textures and high-frequency components and remove heavy noise corruption. Compared with other methods, it has better performance to restore clean and crisp images. It demonstrates that our ART is also suitable for image denoising.

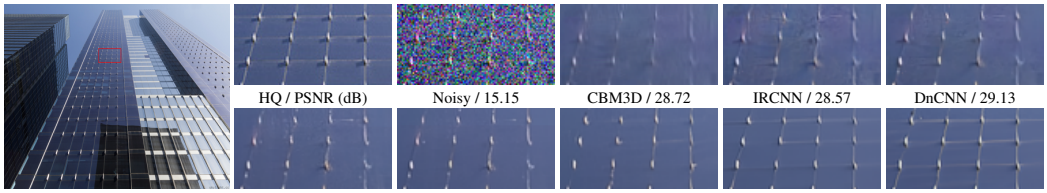


Figure 7: Visual comparison with challenging examples on color image denoising (σ=50).

4.5 JPEG Compression Artifact Reduction

We compare our ART with state-of-the-art JPEG CAR methods: RNAN Zhang et al. (2019), RDN Zhang et al. (2020), DRUNet Zhang et al. (2021a), and SwinIR Liang et al. (2021). Following most recent works, we set the compression quality factors of original images to 40, 30, and 10. We provide the PSNR and SSIM comparison results in Table 5.

Quantitative Comparisons. Table 5 shows the PSNR/SSIM comparisons of our ART with existing state-of-the-art methods. We can see that our proposed method has the best performance. Better results are achieved by ART+ using self-ensemble. These results indicate that our ART also performs outstandingly when solving image compression artifact reduction problems.

5 Conclusion

In this work, we propose Attention Retractable Transformer for image restoration named ART, which offers two types of self-attention blocks to enhance the Transformer representation ability. Most previous image restoration Transformer backbones mainly utilize dense attention modules to alleviate self-attention computation within non-overlapping regions and thus suffer from restricted receptive fields. Without introducing additional computational cost, we employ the sparse attention mechanism to enable tokens from sparse areas of the image to interact with each other. In practice, the alternating application of dense and sparse attention modules is able to provide retractable attention for the model and bring promising improvement. Experiments on image SR, denoising, and JPEG CAR tasks validate that our method achieves state-of-the-art results on various benchmark datasets both quantitatively and visually. In future work, we will try to apply our proposed method to more image restoration tasks, like image deraining, deblurring, dehazing, and so on. We will further explore the potential of sparse attention in solving low-level vision problems.

REPRODUCIBILITY STATEMENT

We provide the reproducibility statement of our proposed method in this section. We introduce the model architecture and core dense and sparse attention modules in Sec. 3. Besides, we also give the implementation details. In Sec. 4.1, we provide the detailed experiment settings. To ensure the reproducibility, we provide the source code and pre-trained models at an anonymous website[2]. We also anonymize our code and put it in a .zip file for submission as supplementary materials. Everyone can run our code to check the training and testing process according to the given instructions. At the website, the pre-trained models are provided to verify the validity of corresponding results. More details please refer to the website or the submitted supplementary materials.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 16

Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *TPAMI*, 2020. 2

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2010. 6

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. 6, 15

Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, 1994. 4

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*, 2021a. 1, 2, 3, 7, 8, 9, 18, 19

Haoyu Chen, Jinjin Gu, and Zhi Zhang. Attention in attention network for image super-resolution. *arXiv preprint arXiv:2104.09497*, 2021b. 2

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, 2021. 1

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. 8, 9, 17

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 3, 7, 18, 19

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1, 2

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016. 4

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. 1, 3

Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022. 2

Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *TIP*, May 2007. 6

Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 1999. 6

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 6, 18, 19

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 2010. 2

[2]https://anonymous.4open.science/r/ART_attention_retractable_transformer

Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, 2019. 1

Xiaowan Hu, Ruijun Ma, Zhihong Liu, Yuanhao Cai, Xiaole Zhao, Yulun Zhang, and Haoqian Wang. Pseudo 3d auto-correlation network for real image denoising. In *CVPR*, 2021. 8, 9

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 6, 15, 17

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016a. 3

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016b. 1

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *CVPR*, pp. 6002–6012, 2022. 3

Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *ICLR*, 2021. 1

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 4

Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. 7, 18, 19

Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *CVPR*, pp. 833–843, 2022. 2

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 1, 2, 3, 4, 6, 7, 8, 9, 15, 17, 18, 19

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1, 6, 7, 14, 15, 18, 19

Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 7, 18, 19

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. 1, 15

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 16

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. 6

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. 18, 19

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. 6, 15, 17

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. 6, 15

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. 7, 18, 19

Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 7, 18, 19

Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *ICCV*, 2013. 2

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. 3, 7, 18, 19

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 16

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 1

Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 4

Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *TIP*, 2006. 6

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 4

Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017. 4

Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 2020. 1

Radu Timofte, Vincent De, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013. 2

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. 6, 14, 15

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022a. 3

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022b. 2, 3, 5

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 3

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 1

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 1

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, 3, 5, 15

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. 1, 3, 16

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

Zhihao Xia and Ayan Chakrabarti. Identifying recurring patterns with deep neural networks for natural image denoising. In *WACV*, 2020. 1

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 1, 3

Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. In *NeurIPS*, 2021. 2, 3, 5

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. 2, 16

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. 2

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. 1, 3, 8, 9, 16, 17, 18

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. 7th Int. Conf. Curves Surf.*, 2010. 6, 15

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. 1, 2, 8, 9, 17

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017b. 8, 9, 17

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *TIP*, 2018a. 8, 9, 17

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018b. 18, 19

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. 3, 8, 9, 17

Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *J Electron Imaging*, 2011. 6

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. 1, 2, 3, 4, 6, 7, 18, 19

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 8, 9, 17

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 2020. 1, 3, 4, 8, 9, 17, 18, 19

Yulun Zhang, Huan Wang, Can Qin, and Yun Fu. Aligned structured sparsity learning for efficient image super-resolution. In *NeurIPS*, 2021b. 1

Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *ICCV*, 2021c. 18, 19

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020. 1

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 1

Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, 2020. 7, 18, 19

APPENDIX

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

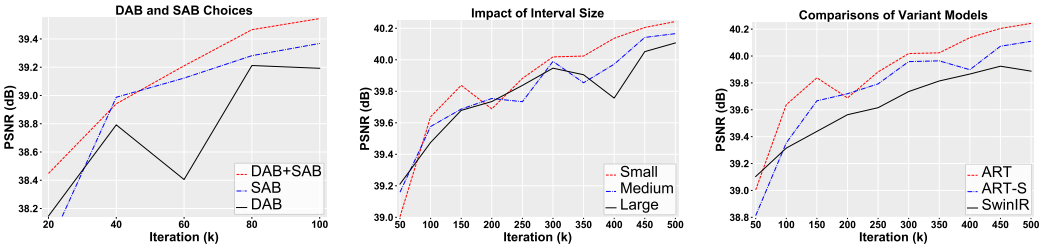


Figure 8: **Left:** PSNR (dB) comparison of our ART using all dense attention block (DAB), using all sparse attention block (SAB), and using alternating DAB and SAB. **Middle:** PSNR (dB) comparison of our ART using large interval size in sparse attention block which is $(8, 8, 8, 8, 8, 8)$ for six residual 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.

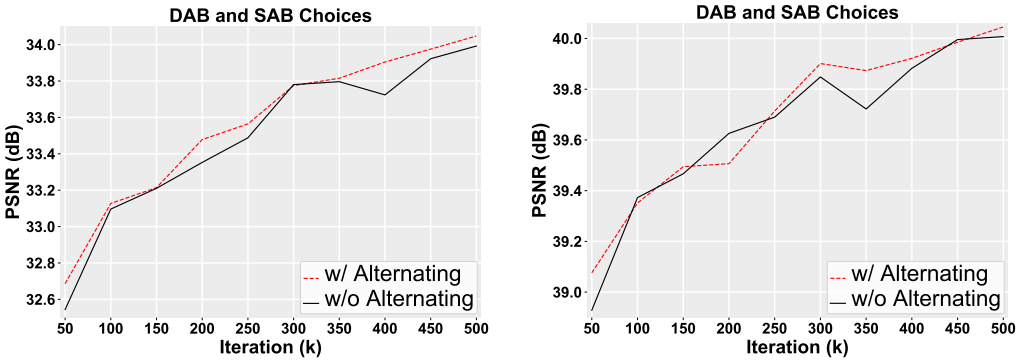


Figure 9: PSNR (dB) comparison of our ART using 3 pairs of alternating dense attention block (DAB) and sparse attention block (SAB), and using 3 successive SABs following 3 successive CABs. The former is with ("w/") alternating and the latter is without ("w/o") alternating. The **left** figure shows the testing results on Urban100 and the **right** one shows the testing results on Manga109.

A 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. 9 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. 8(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 Sec. 4.2 on Manga109. The results are shown in Fig. 8(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. Our purpose is to

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 6: 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

keep the same computation complexity with SwinIR in training phase. It is known that the input training image size is 64×64 and the window size of SwinIR is 8. Therefore, our sparse attention module will not introduce additional computational cost by extracting the same 8×8 tokens with SwinIR. In practice, our ART-S has the same FLOPs with SwinIR (e.g., 51.3G FLOPs) when training for ($\times 2$) image SR. We show PSNR comparison on Manga109 datasets as the training iterations increase in Fig. 8(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.

B 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.

B.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 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. From this point of view, 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.

B.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. 6 and Fig. 10(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.

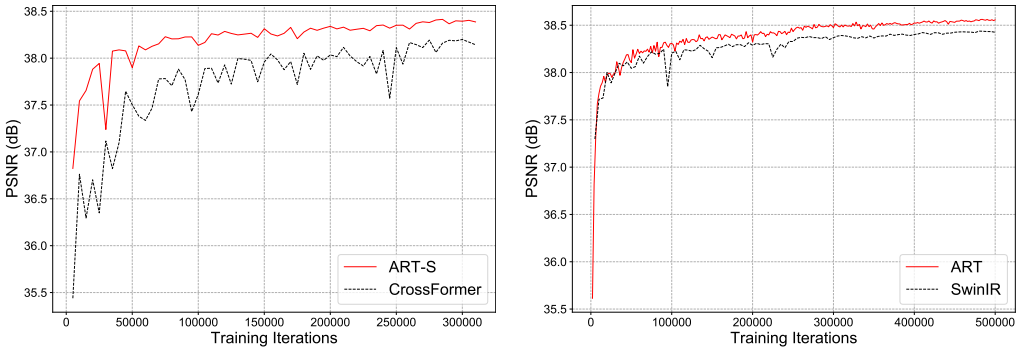


Figure 10: Convergence Analyses on $\times 2$ SR. **Left:** comparisons of CrossFormer and ART-S with training 300k iterations. **Right:** comparisons of SwinIR and ART with training 500k iterations.

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 7: 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.

Convergence Analyses. We train these two models for close to 300k iterations. In Fig. 10(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 6 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.

C Experiments on Real Image Denoising

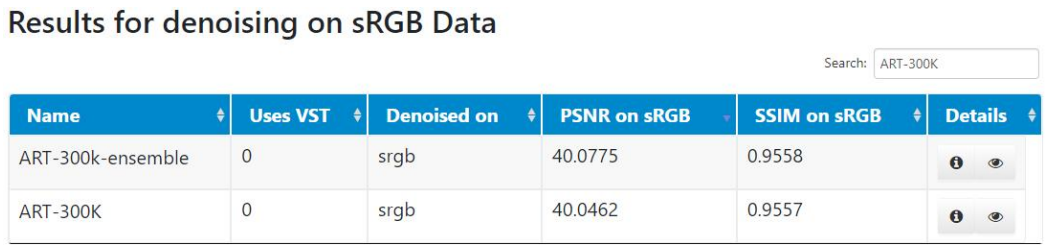
We conduct experiments on another image restoration task to show the superiority of our method. In practice, we employ our method to solve Real Image Denoising problems. We give detailed introduction as follows.

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 MIRNet Zamir et al. (2020), Uformer Wang et al. (2022b), and Restormer. In Tab. 7, 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. 11. In conclusion, our proposed method shows strong ability to restore high-quality images in real image denoising.



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

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

D 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 as introduced in Sec 4.1. We use commonly-used 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.

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. C. 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 8, 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 9 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$\times$128$\times$128. 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. 15-17. 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.

E Additional Experimental results

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

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 9: 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.

Quantitative Comparisons. Table 10 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. (2021c), DFSA Magid et al. (2021), IPT Chen et al. (2021a), 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. 10(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 $\times2$ 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 12 and 13 show some challenging examples for ($\times4$) 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 14 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 15-17 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.

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. (2021c)	×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. (2021a)	×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. (2021c)	×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. (2021a)	×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. (2021c)	×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. (2021a)	×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 10: PSNR (dB)/SSIM comparisons for image super-resolution on five benchmark datasets. We color best and second best results in red and blue.

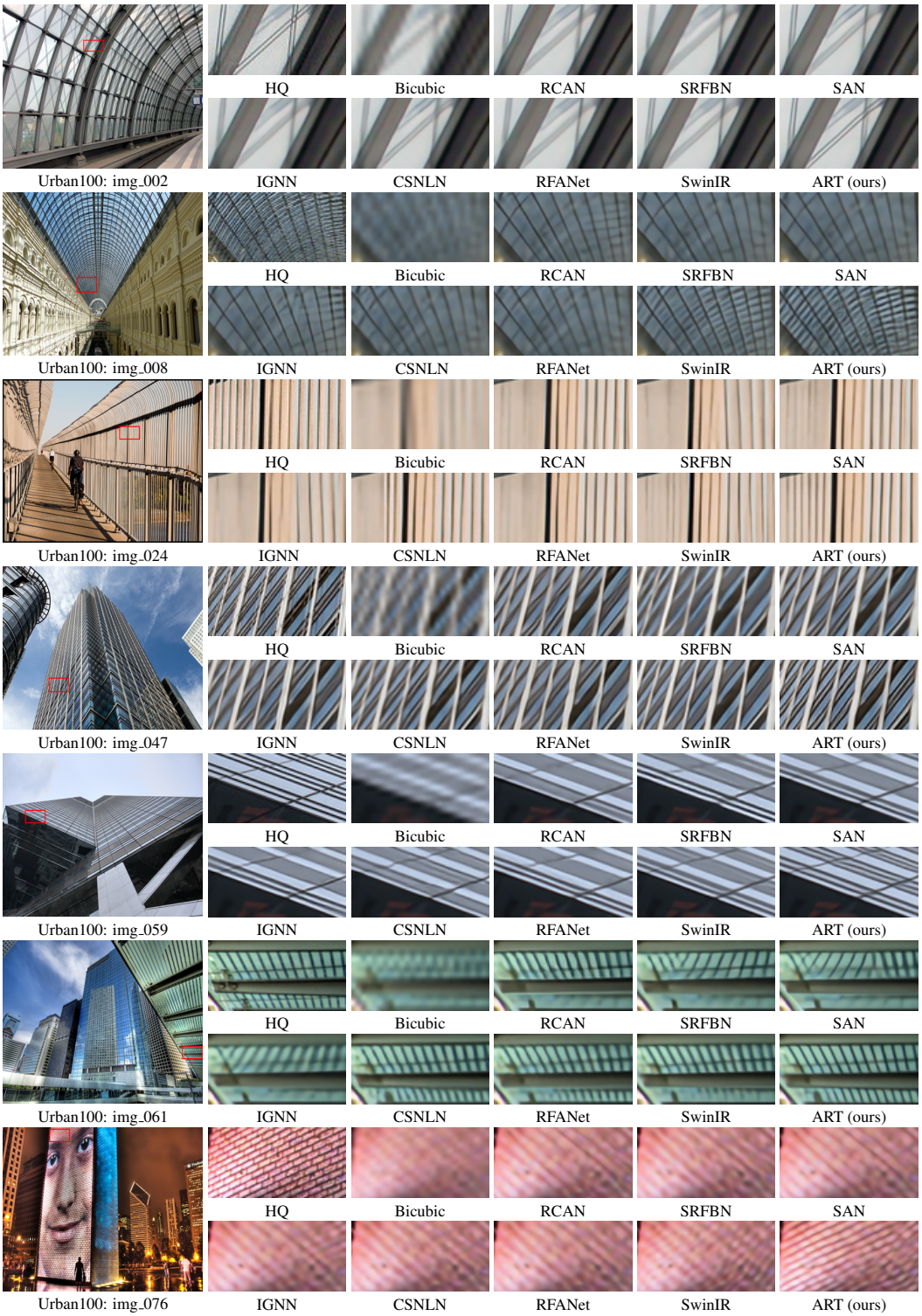


Figure 12: Visual comparison ($\times 4$) with image SR networks on Urban100 dataset.

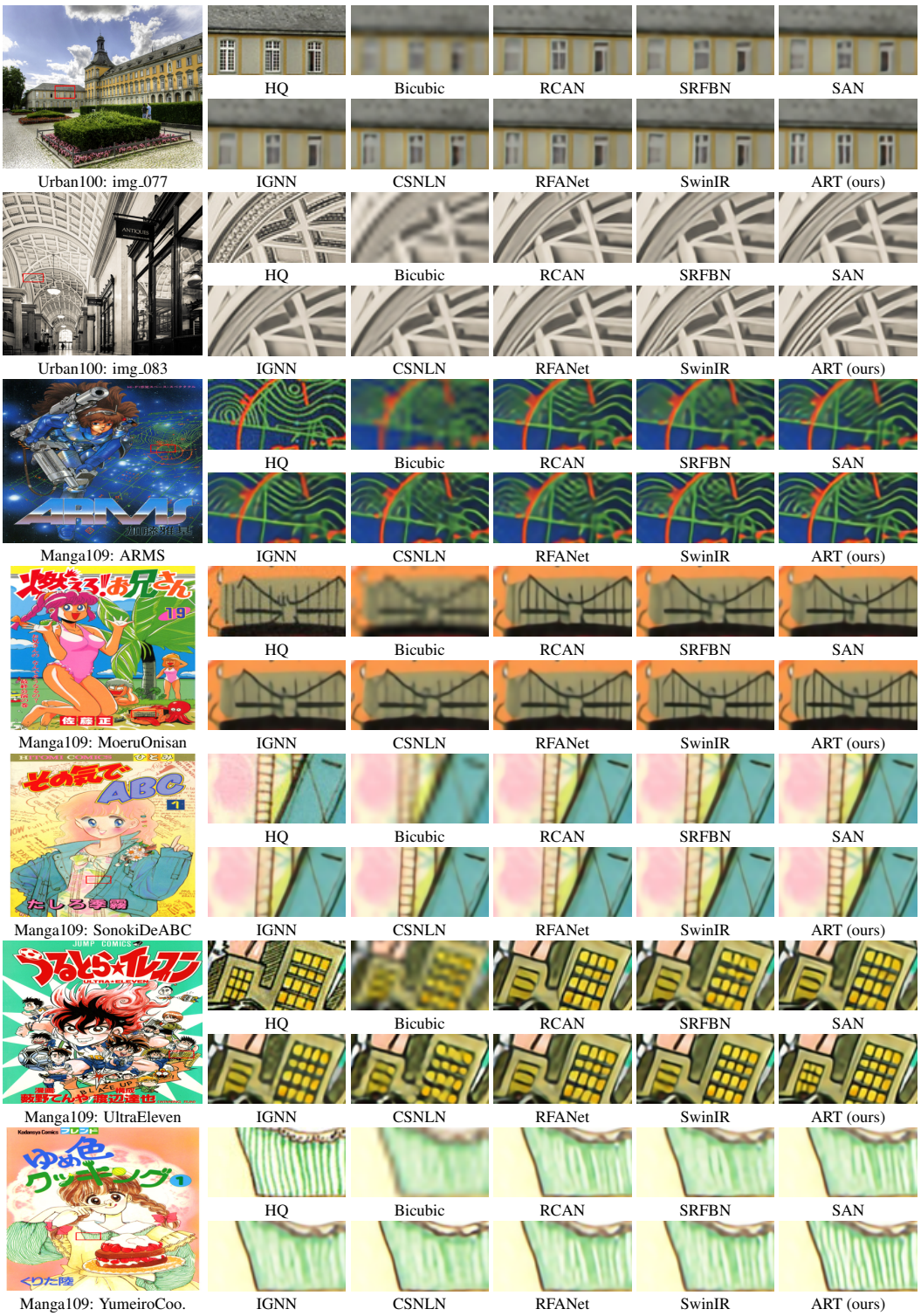


Figure 13: Visual comparison ($\times 4$) with image SR networks on Urban100 and Manga109 dataset.

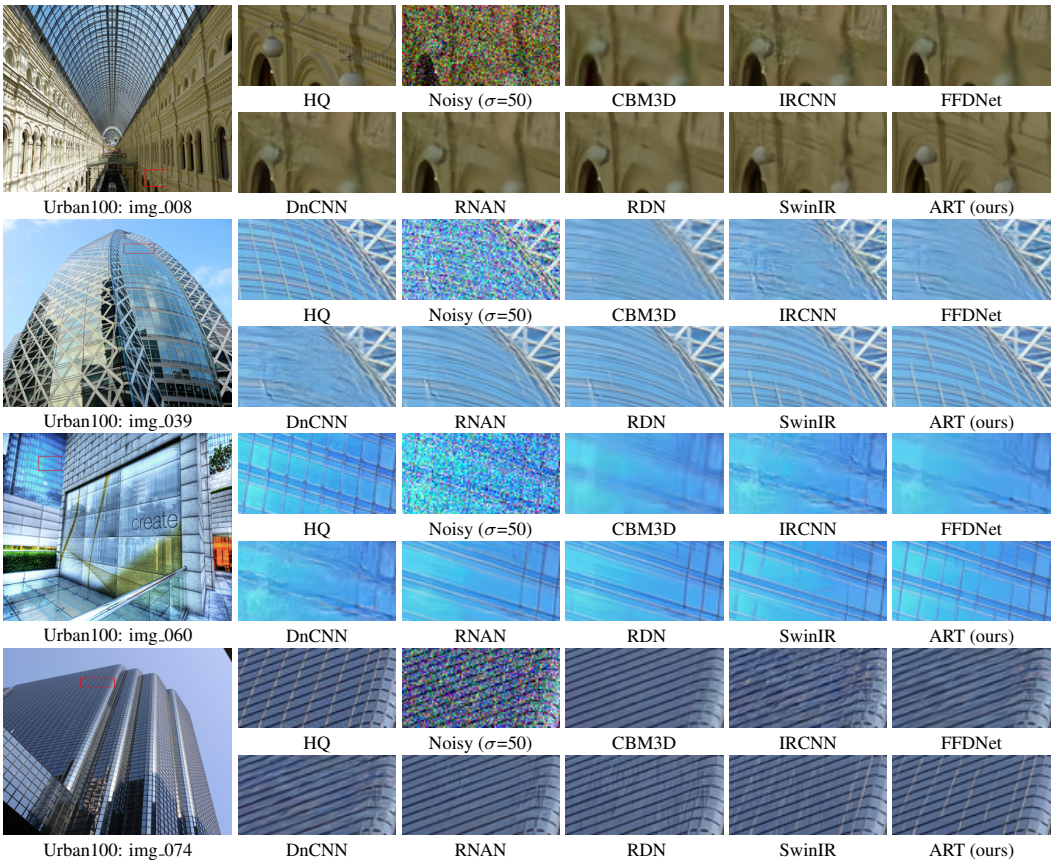


Figure 14: Visual comparison (σ=50) for Gaussian color image denoising on Urban100 dataset.

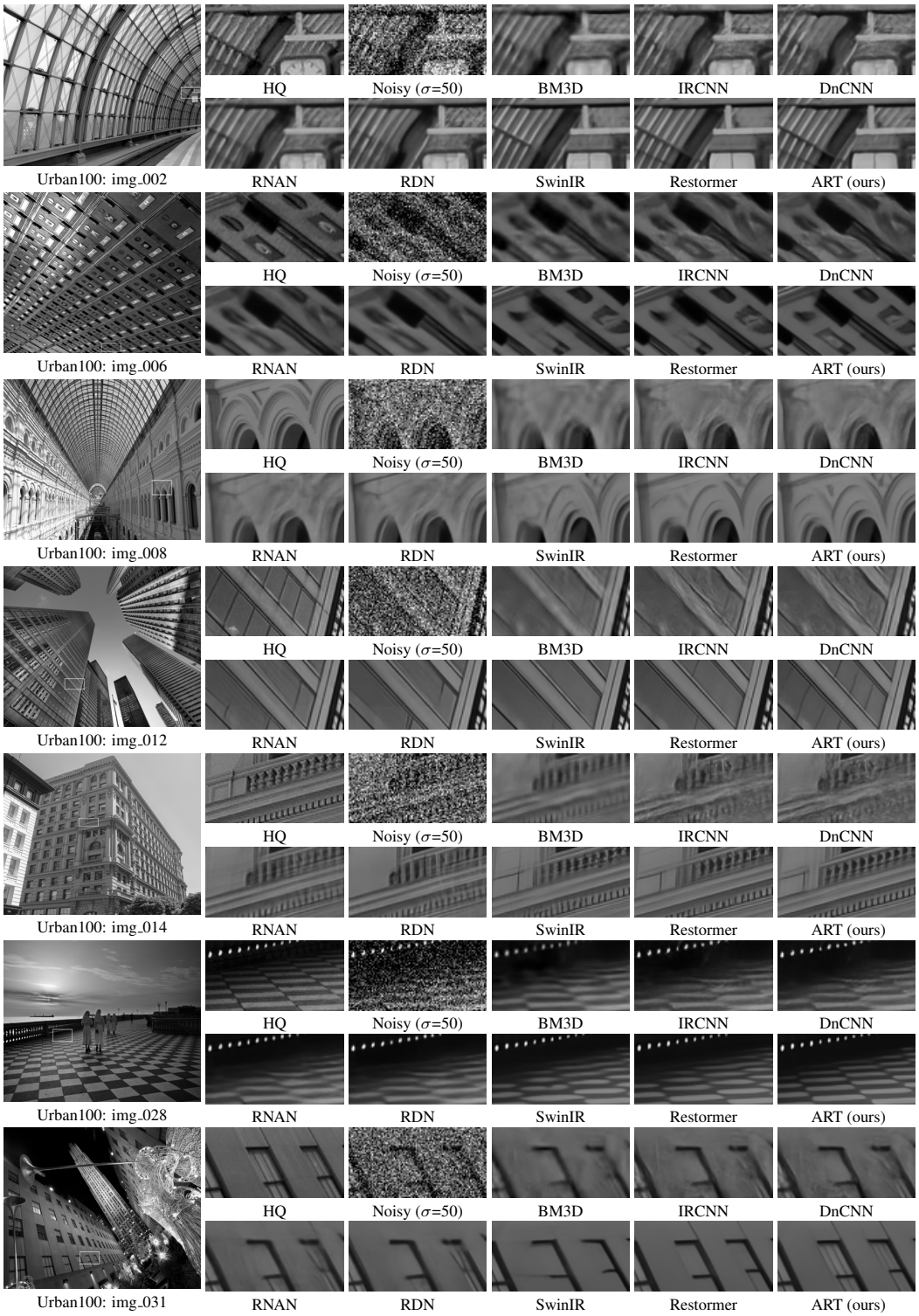


Figure 15: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.

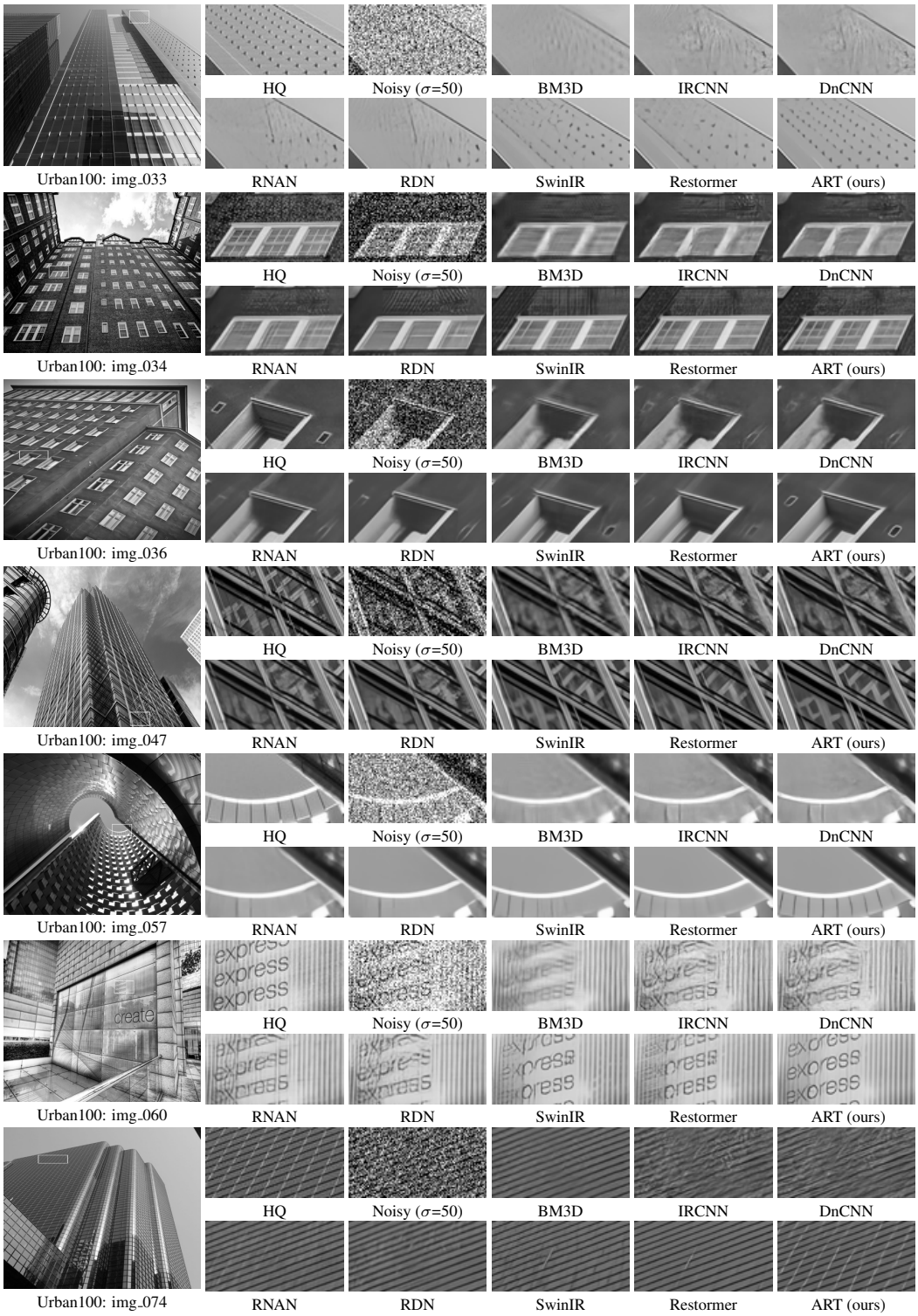


Figure 16: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.

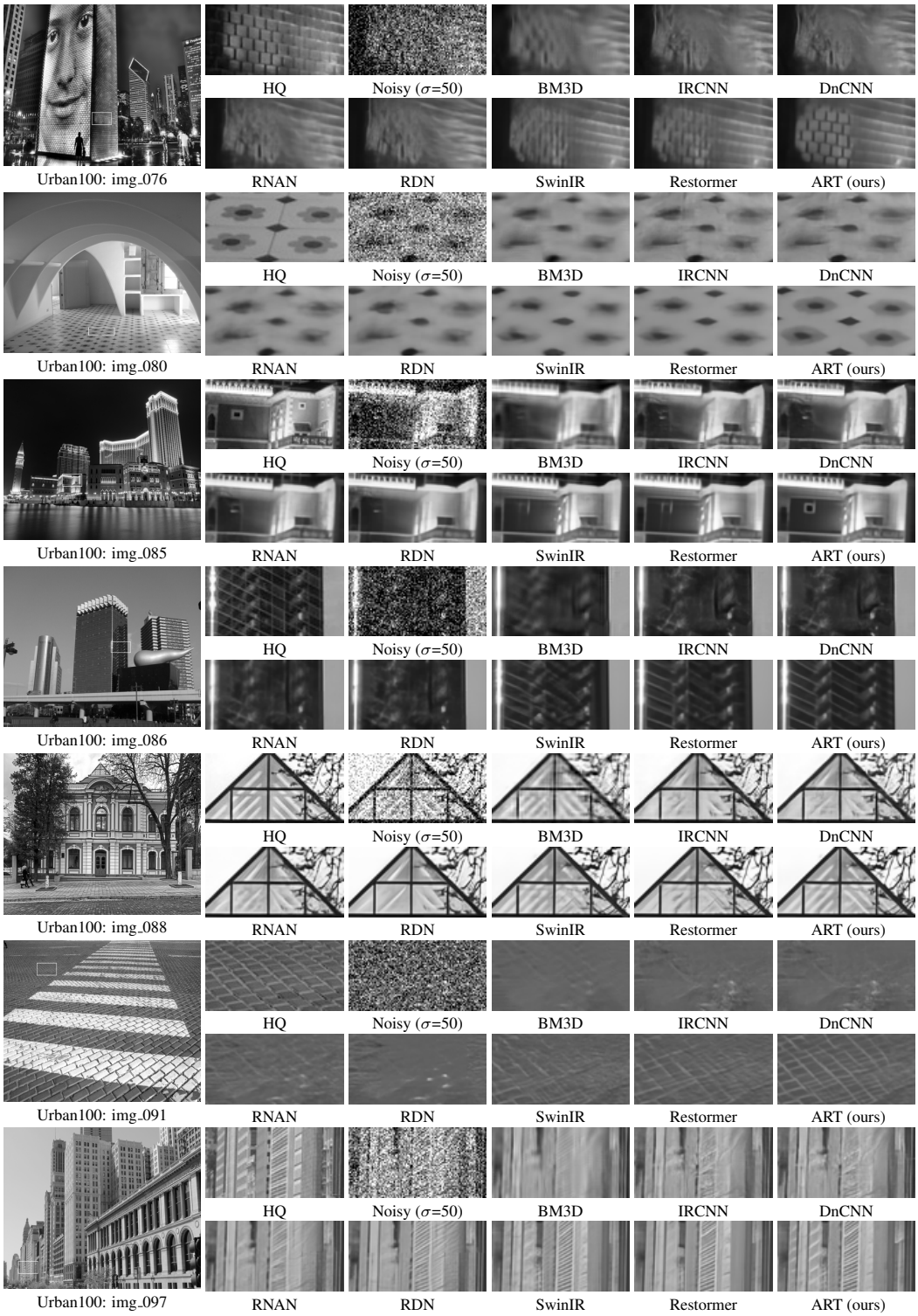


Figure 17: Visual comparison (σ=50) for Gaussian grayscale image denoising on Urban100 dataset.