Supplementary Material:
Dual Aggregation Transformer for Image Super-Resolution

Zheng Chen[1], Yulun Zhang[2*], Jinjin Gu[3,4], Linghe Kong[1*], Xiaokang Yang[1], Fisher Yu[2]
[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Shanghai AI Laboratory, [4]The University of Sydney

Overview

The supplementary material is organized as follows. In Sec. 1, we provide more discussions about AIM. Sec. 2 provides more versions of DAT. In Sec. 3, we provide further analyses, investigating the advantages of aggregating channel and spatial information. Sec. 4, we conduct experiments on the recent state-of-the-art method ELAN for fair comparisons with our method. Sec. 5 and Sec. 6 provide more quantitative and visual comparisons. Finally, in Sec. 7, we discuss the limitation and future work of our method.

1. More Discussions about AIM

Firstly, we further explain the motivation of the AIM. **Then**, we describe the design considerations of spatial-interaction (S-I) and channel-interaction (C-I) of AIM.

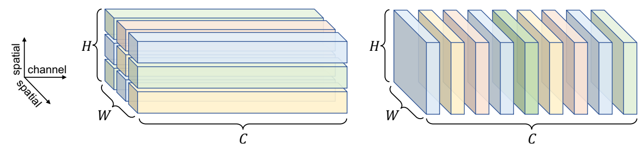
1.1. Motivation

In general, AIM is proposed to enhance the fusion of depth-wise convolution (DW-Conv) and self-attention (SA) branches, and to aggregate spatial and channel information in a single SA module. **Firstly**, considering the misalignment between local (DW-Conv) and global (SA) features [18, 24], the two branches cannot be fused effectively. **Secondly**, SA applies a weight-sharing mechanism, limiting its feature learning in shared dimensions [8, 3]. As shown in Fig. 1, SW-SA applies the same spatial attention map (dynamic weights, from the dot-product between $query$ and key) to each channel, namely, sharing weights on channel dimensions. Similarly, CW-SA shares weights on spatial dimensions. Since weight sharing, a single SA module cannot effectively aggregate both spatial and channel dimensions. **Finally**, to alleviate above issues, we adaptively adjust features through dynamic weights. Meanwhile, considering the parallel structure, the dynamic weights are generated from interactions between two branches.

1.2. Design Considerations

To realize the above purpose, we propose AIM, which consists of spatial-interaction (S-I) and channel-interaction (C-I). Specifically, we first introduce interaction from DW-Conv to SA. **For SW-SA**, as the above analysis, we generate a channel attention map (C-Map) to adjust channel dimension. **For CW-SA**, we generate a spatial attention map (S-Map). For C-I, we follow the design of the SE layer [10]. For S-I, considering that the convolution branch already extracts spatial information, we only utilize 1×1 convolution to compress the channel dimension without explicitly modeling spatial information in S-I.

Furthermore, from the perspective of duality, we also introduce interaction from SA to DW-Conv. Since SW-SA extracts strong spatial information, we utilize S-I to transfer it to the corresponding DW-Conv branch. Similarly, we apply C-I in CW-SA to strengthen the channel expression of convolution. Synthesizing the above designs, we propose the AIM to enhance branch fusion and achieve feature aggregation. The ablation study in Table 1(b, c) in the main paper demonstrates the effectiveness of our AIM.



(a) SW-SA (b) CW-SA

Figure 1: Illustration of the weight sharing mechanism. (a) SW-SA shares attention weights on the channel dimensions (same color along the channel dimension). (b) CW-SA shares attention weights on spatial dimensions (same color along the spatial dimension). For simplification, we depict the case that the attention is single-head. The feature size ($H\times W\times C$) here is $3\times3\times9$.

2. More DAT Variants

In this section, we provide more versions of DAT to demonstrate the effectiveness of our proposed method. **Firstly**, we provide DAT-2 with fewer Params (model parameters) and similar FLOPs (computational complexity) to SwinIR [13]. **Secondly**, we provide a light-weight model, DAT-light, for light-weight image SR. **Finally**, we provide DAT-3 with the same window size (8×8) as SwinIR.

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

Method	Scale	Params	FLOPs	Set5		Set14		B100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [13]	×2	11.75M	205.31G	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
DAT-2	×2	11.06M	206.93G	**38.58**	**0.9627**	**34.78**	**0.9268**	**32.60**	**0.9050**	**34.31**	**0.9457**	**40.29**	**0.9806**
SwinIR [13]	×3	11.94M	208.48G	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
DAT-2	×3	11.25M	210.09G	**35.14**	**0.9328**	**31.06**	**0.8547**	**29.53**	**0.8161**	**30.13**	**0.8878**	**35.49**	**0.9550**
SwinIR [13]	×4	11.90M	215.32G	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
DAT-2	×4	11.21M	216.93G	**33.01**	**0.9047**	**29.21**	**0.7965**	**27.98**	**0.7509**	**27.86**	**0.8341**	**32.41**	**0.9285**

Table 1: Quantitative comparison (PSNR/SSIM) between DAT-2 and SwinIR [13]. The input size is 3×128×128 to calculate FLOPs.

Method	Scale	Params	FLOPs	Set5		Set14		B100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light [13]	×2	910K	244.37G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN-light [29]	×2	621K	201.34G	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
DAT-light	×2	553K	194.26G	**38.24**	**0.9614**	**34.01**	**0.9214**	**32.34**	**0.9019**	**32.89**	**0.9346**	**39.49**	**0.9788**
SwinIR-light [13]	×3	930K	110.80G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
ELAN-light [29]	×3	629K	89.48G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
DAT-light	×3	561K	88.59G	**34.76**	**0.9299**	**30.63**	**0.8474**	**29.29**	**0.8103**	**28.89**	**0.8666**	**34.55**	**0.9501**
SwinIR-light [13]	×4	930K	63.59G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN-light [29]	×4	640K	53.72G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
DAT-light	×4	573K	49.69G	**32.57**	**0.8991**	**28.87**	**0.7879**	**27.74**	**0.7428**	**26.64**	**0.8033**	**31.37**	**0.9178**

Table 2: Quantitative comparison (PSNR/SSIM) for lightweight image SR. Output size is 3×1280×720 to calculate FLOPs. We re-test the Params and FLOPs with all official codes on SwinIR and ELAN.

Method	Scale	Params	FLOPs	Set5		Set14		B100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [13]	×2	11.75M	205.31G	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	**0.9427**	39.92	0.9797
DAT-3	×2	11.06M	186.54G	**38.52**	**0.9626**	**34.58**	**0.9261**	**32.54**	**0.9043**	**33.86**	0.9424	**40.12**	**0.9804**

Table 3: Quantitative comparison (PSNR/SSIM) between DAT-3 and SwinIR [13] (×2). The input size is 3×128×128 to calculate FLOPs.

2.1. Another Model: DAT-2

Implementation details. We provide another model variant of DAT: DAT-2. We set the residual group (RG) number, dual aggregation Transformer block (DATB) pair number, channel number, attention head number, and channel expansion factor as 6, 3, 180, 6, and 2, respectively. We set the window size as 8×32 for DAT-2.

Training Settings. We train DAT-2 on DIV2K [25] and Flickr2K [14]. We use five datasets for evaluation: Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20] with three factors: ×2, ×3, and ×4. The training settings are consistent with DAT-S and DAT.

Quantitative Results. We compare our DAT-2 with SwinIR [13]. The results are listed in Table 1. FLOPs are calculated when the input size is set as 3×128×128 for three scale factors. As we can see, our DAT-2 outperforms SwinIR on all datasets with all scale factors. Specially, our DAT-2 achieves 0.5 dB and 0.37 dB gains on Urban100 and Manga109 (×2), respectively. Meanwhile, DAT-2 has fewer Params and similar FLOPs to SwinIR.

2.2. Light-weight Model: DAT-light

Implementation details. We provide a light-weight model, DAT-light, for light-weight image SR. DAT-light only has 1 RG and 9 DATB pairs (9 DCTB and 9 DSTB). The channel number, attention head number, and channel expansion factor are set as 60, 6, and 2, respectively. The window size for DSTB is set as 8×32.

Training Settings. We train DAT-light on DIV2K [25] and Flickr2K [14], and test it on Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20]. The training settings are consistent with DAT-S and DAT.

Quantitative Results. We compare our DAT-light with recent state-of-the-art lightweight methods: SwinIR [13] and ELAN [29], in Table 2. FLOPs are calculated when the output size is set as 3×1280×720 for three scale factors. Please note that we re-test the Params and FLOPs of SwinIR and ELAN with their official codes. Our DAT-light achieves better performance with fewer Params and FLOPs for all scale factors, compared with SwinIR and ELAN.

2.3. 8×8 Window Size Model: DAT-3

Implementation details. We provide DAT-3 with a window size of 8×8, the same as SwinIR [13]. Specifically, we set the RG number, DATB pair number, channel number, attention head number, and channel expansion factor as 6, 3, 180, 6, and 2. The window size for DSTB is set as 8×8.

Training Settings. We train DAT-3 on DIV2K [25] and Flickr2K [14]. The training settings are consistent with DAT-S and DAT. The main paper has more details.

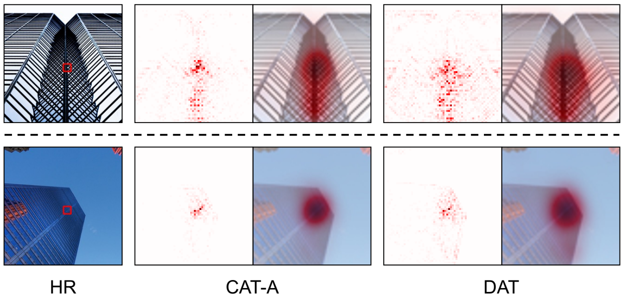


Figure 2: LAM [7] comparison between CAT-A [4] and DAT.

Quantitative Results. To demonstrate the effectiveness of our method, we build DAT-3 with the same window size as SwinIR [13]. Due to time issues, we only train DAT-3 on image SR (×2), and compare it with SwinIR. The results are shown in Table 3. FLOPs are calculated when the output size is set as 3×128×128. As we can see, with fewer Params and FLOPs, our DAT-3 outperforms SwinIR, except for the SSIM value on Urban100. Especially, our DAT-3 obtains a 0.2 dB PSNR gains on Manga109. All these results demonstrate the effectiveness of our methods.

3. Further Analyses

In this section, we provide more quantitative and visual analyses. **Firstly**, we apply the LAM [7] to visualize the range of information utilization. **Then**, we introduce several perceptual similarity metrics to evaluate our method. **Finally**, we plot the convergence curves during training for SwinIR and our models.

3.1. LAM Analyses

We apply the LAM [7] to analyze the performance of our DAT. LAM is a diagnostic tool designed for image super-resolution (SR). It can show the pixels that contribute most to the reconstruction of the selected region. The corresponding pixels are marked as red in related images. More marked pixels mean the model can utilize more information, thus resulting in better performance. Fig. 2 shows the LAM comparisons between CAT-A [4] and DAT. Comparing the second and fourth columns, we can find that the number of red marker points of our DAT is more than CAT-A. It indicates that our DAT has larger receptive fields and utilizes more global information to restore images. This is because our method has a stronger representation ability through aggregating spatial and channel features.

3.2. Perceptual Similarity Analyses

In the main paper, we quantitatively compare our method with current methods using metrics: PSNR/SSIM. However, the literature [2] reveals that the superiority of PSNR values does not always accord with better visual quality. Moreover, we also found that compared to CAT-A, our DAT has a lower PSNR value on Urban100 (×4), but has better visual results. To further evaluate our methods, we introduce two metrics: LPIPS [28] and DISTS [6]. Compared with PSNR, LPIPS and DISTS align more with human perception. The lower the value of LPIPS and DISTS, the more similar the two images. We compare our DAT with SwinIR and CAT-A on Urban100 and Manga109 with a scale factor of ×4. The results are listed in Table 4. Our DAT achieves the best performance (lowest value) on both datasets. This result demonstrates the superiority of our method. It is also consistent with the visual comparison in Figs. 4 and 5.

Method	Scale	Urban100		Manga109	
		LPIPS ↓	DISTS ↓	LPIPS ↓	DISTS ↓
SwinIR [13]	×4	0.1840	0.1533	0.0926	0.0766
CAT-A [4]	×4	0.1801	0.1502	0.0906	0.0753
DAT (ours)	×4	**0.1765**	**0.1487**	**0.0896**	**0.0735**

Table 4: Perceptual Similarity (LPIPS/DISTS) Comparison.

3.3. Convergence Analyses

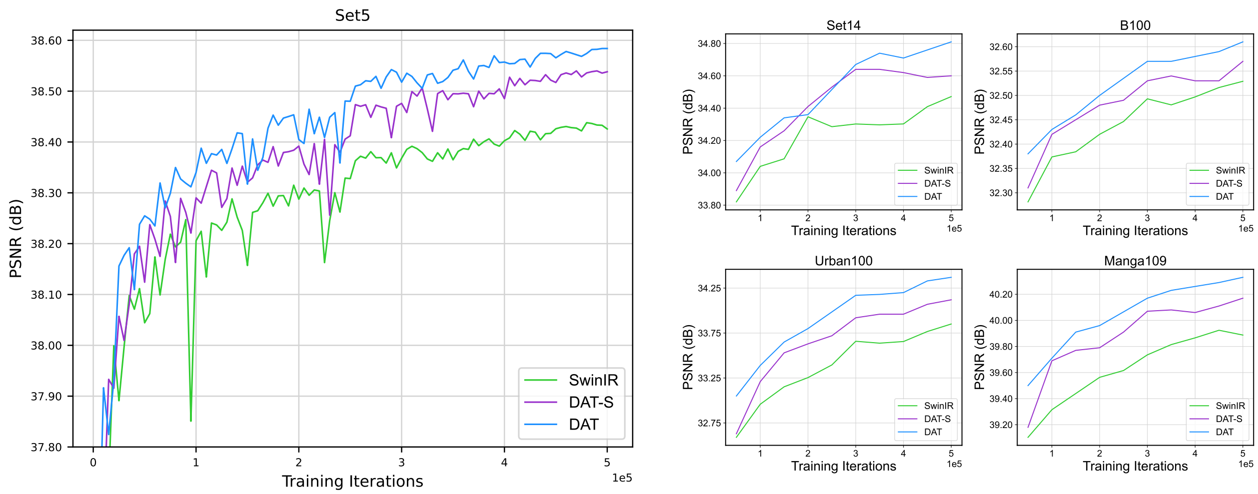
The convergence curves for SwinIR, DAT-S, and DAT are shown in Fig. 3. PSNR values are tested on Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20] (×2). For fair comparisons, we train SiwinIR under the official code with the same training settings as our methods. The total training iterations are 5×10^5. We sample every 5×10^3 iterations on Set5, while every 5×10^4 iterations on other datasets. We can observe that both DAT-S and DAT converge faster than SwinIR on all datasets. These results are in accord with quantitative comparisons in Table 6, further demonstrating the effectiveness of our method.

4. Recent Method: ELAN

Recently, Zhang *et al*. [29] proposed a new image SR Transformer model, named efficient long-range attention network (ELAN). ELAN improves the efficiency of Transformer in SR tasks and outperforms SwinIR [13] in some cases. However, ELAN is trained on DIV2K [25], while our DAT-S and DAT are on DIV2K and Flickr2K [14]. Meanwhile, the model size and computational complexity of ELAN are smaller than our model. For fair comparisons, we increase the efficient long-range attention block (ELAB) number in ELAN, thus constructing a new variant of ELAN, denoted as ELAN-2. Then we re-train ELAN and ELAN-2 on DIV2K and Flickr2K.

4.1. Experimental Settings

Implementation Details. For ELAN, we adopt the settings in the official paper [29]. Specifically, the ELAB number is 36, and the channel number is 180. The GMSA module contains three window sizes: 4×4, 8×8, and 16×16. For ELAN-2, we only increase the ELAB number from 36 to 48, while other settings are the same as ELAN.



(a) Sampled every 5×10^3 iterations.

(d) Sampled every 5×10^4 iterations.

Figure 3: Convergence comparison on SwinIR [13], DAT-S, and DAT.

Method	Training Dataset	Params	FLOPs	Set5		Set14		B100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ELAN [29]	DIV2K	8.43M	161.24G	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
ELAN* [29]	DIV2K+Flickr2K	8.43M	161.24G	38.38	0.9619	34.41	0.9247	32.48	0.9035	33.54	0.9410	39.74	0.9795
ELAN-2* [29]	DIV2K+Flickr2K	11.23M	214.86G	38.44	0.9623	34.43	0.9245	32.51	0.9040	33.76	0.9423	39.90	0.9800
SwinIR [13]	DIV2K+Flickr2K	11.75M	205.31G	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
DAT-S (ours)	DIV2K+Flickr2K	11.06M	193.34G	**38.54**	**0.9627**	**34.60**	**0.9258**	**32.57**	**0.9047**	**34.12**	**0.9444**	**40.17**	**0.9804**

Table 5: Quantitative comparison ($\times$2) about variants of ELAN [29], SwinIR [13], and our DAT-S. The input size is $3\times128\times128$ to calculate FLOPs. "*" indicates that we re-train the model on DIV2K [25] and Flickr2K [14].

Training Settings. We re-train ELAN and ELAN-2 on DIV2K [25] and Flickr2K [14] with the official code. We train two models with patch size 64×64 and batch size 32. Following the training details in the official paper [29], we use Adam optimizer with β_1=0.9, β_2=0.999, and ϵ=10^{-8}. The total training epochs are 500. The initial learning rate is set as 2×10^{-4} and reduced by half at epochs [250,400,425,450,475]. Moreover, we adopt random rotation and flips for data augmentation.

4.2. Model Comparisons

We compare re-trained ELAN and ELAN-2 with official ELAN [29], SwinIR [13], and our DAT-S on five benchmark datasets: Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20] with scale factor $\times$2. The performance (PSNR/SSIM), parameters, and FLOPs are reported. FLOPs are calculated when the input size is $3\times128\times128$. The results are listed in Table 5. We can observe that using DIV2K and Flickr2K to train ELAN can improve the performance of the model. Compared with training on DIV2K, ELAN obtains 0.1 dB and 0.12 dB gains on Urban100 and Manga109. Meanwhile, increasing the number of efficient long-range attention blocks in ELAN can also advance performance. Furthermore, comparing ELAN-2, SwinIR, and our DAT-S, we can find that our DAT-S outperforms the other two models with fewer parameters and FLOPs. Especially the three models are trained on the same dataset.

5. More Quantitative Results

We compare our models: DAT-S, DAT-2, and DAT, with state-of-the-art methods: EDSR [14], SRMDNF [27], RDN [32], OISR [9], RCAN [30], NLRN [15], RNAN [31], SRFBN [12], SAN [5], RFANet [16], IGNN [34], HAN [23], CSNLN [22], NLSA [21], CRAN [33], ELAN [29], DFSA [17], SwinIR [13], and CAT-A [4]. We use self-ensemble strategy in testing and mark the model with a symbol "+". Quantitative results are shown in Table 6. Our DAT outperforms other methods in all cases, except for the PSNR value (CAT-A) on Urban100 ($\times$4). Additionally, our DAT-S and DAT-2 achieve comparable or better performance than previous methods.

6. More Visual Results

We provide more visual comparisons in Figs. 4 and 5 as the supplement of the visualization in the main paper. As we can see, most compared methods suffer from blurring artifacts and cannot recover some details in some challenging cases. In contrast, our DAT can alleviate the blurring artifact to some degree and recover sharp textures. For instance, in img_021, our DAT recovers more textures and patterns than other methods. Similar observations are shown in other images. These visual comparisons further demonstrate that our method has powerful modeling capability by aggregating spatial and channel features.

7. Limitations and Future Work

In this work, we propose the dual aggregation Transformer (DAT), for image SR. Our DAT can aggregate spatial and channel information via the inter-block and intra-block manner, thus obtaining powerful representation ability. Our DAT outperforms recent state-of-the-art image SR methods. Nevertheless, we for more types of image SR tasks (*e.g.*, blind and real-world image SR), we have not explored. We will apply our DAT to more kinds of image SR tasks in the future to further demonstrate the effectiveness of our proposed method. In addition, we mainly focus on designing the Transformer block to aggregate spatial and channel information. For the network architecture, we have not investigated it. In future work, we will explore other network structures, such as parallel or multi-scale architectures.

References

[1] 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. 2, 3, 4

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 3

[3] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *CVPR*, 2022. 1

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 3, 4, 7, 8, 9

[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. 4, 7, 8, 9

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 2020. 3

[7] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, 2021. 3

[8] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *ICLR*, 2022. 1

[9] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *CVPR*, 2019. 4, 7

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1

[11] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 2, 3, 4

[12] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. 4, 7

[13] 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, 7, 8, 9

[14] 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, 4, 7

[15] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, 2018. 4, 7

[16] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 4, 7, 8, 9

[17] 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. 4, 7

[18] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. In *NeurIPS*, 2021. 1

[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. 2, 3, 4

[20] 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. 2, 3, 4

[21] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 4, 7

[22] 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. 4, 7, 8, 9

[23] 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. 4, 7, 8, 9

[24] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, 2021. 1

[25] 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, 4

[26] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. 7th Int. Conf. Curves Surf.*, 2010. 2, 3, 4

[27] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 4, 7

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[29] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022. 2, 3, 4, 7

[30] 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*, 2018. 4, 7, 8, 9

[31] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 4, 7

[32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 4, 7

[33] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *ICCV*, 2021. 4, 7

[34] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, 2020. 4, 7

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [14]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
SRMDNF [27]	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
RDN [32]	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
OISR [9]	×2	38.21	0.9612	33.94	0.9206	32.36	0.9019	33.03	0.9365	-	-
RCAN [30]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
NLRN [15]	×2	38.00	0.9603	33.46	0.9159	32.19	0.8992	31.81	0.9249	-	-
RNAN [31]	×2	38.17	0.9611	33.87	0.9207	32.31	0.9014	32.73	0.9340	39.23	0.9785
SRFBN [12]	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
SAN [5]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
RFANet [16]	×2	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
IGNN [34]	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN [23]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
CSNLN [22]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [21]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
CRAN [33]	×2	38.31	0.9617	34.22	0.9232	32.44	0.9029	33.43	0.9394	39.75	0.9793
ELAN [29]	×2	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
DFSA [17]	×2	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
SwinIR [13]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
CAT-A [4]	×2	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
DAT-S (ours)	×2	38.54	0.9627	34.60	0.9258	32.57	0.9047	34.12	0.9444	40.17	0.9804
DAT-2 (ours)	×2	38.58	0.9627	34.78	0.9268	32.60	0.9050	34.31	0.9457	40.29	0.9806
DAT (ours)	×2	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458	40.33	0.9807
DAT-S+ (ours)	×2	38.57	0.9627	34.72	0.9267	32.60	0.9049	34.25	0.9451	40.26	0.9807
DAT2+ (ours)	×2	38.61	0.9629	34.81	0.9271	32.62	0.9052	34.44	0.9463	40.38	0.9808
DAT+ (ours)	×2	38.63	0.9631	34.86	0.9274	32.63	0.9053	34.47	0.9465	40.43	0.9809
EDSR [14]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
SRMDNF [27]	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN [32]	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
OISR [9]	×3	34.72	0.9297	30.57	0.8470	29.29	0.8103	28.95	0.8680	-	-
RCAN [30]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
NLRN [15]	×3	34.27	0.9266	30.16	0.8374	29.06	0.8026	27.93	0.8453	-	-
RNAN [31]	×3	34.66	0.9290	30.53	0.8463	29.26	0.8090	28.75	0.8646	34.25	0.9483
SRFBN [12]	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
SAN [5]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
RFANet [16]	×3	34.79	0.9300	30.67	0.8487	29.34	0.8115	29.15	0.8720	34.59	0.9506
IGNN [34]	×3	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
HAN [23]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
CSNLN [22]	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
NLSA [21]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
CRAN [33]	×3	34.80	0.9304	30.73	0.8498	29.38	0.8124	29.33	0.8745	34.84	0.9515
ELAN [29]	×3	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
DFSA [17]	×3	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
SwinIR [13]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
CAT-A [4]	×3	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
DAT-S (ours)	×3	35.12	0.9327	31.04	0.8543	29.51	0.8157	29.98	0.8846	35.41	0.9546
DAT-2 (ours)	×3	35.14	0.9328	31.06	0.8547	29.53	0.8161	30.13	0.8878	35.49	0.955
DAT (ours)	×3	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
DAT-S+ (ours)	×3	35.16	0.9329	31.10	0.8550	29.54	0.8161	30.10	0.8861	35.54	0.9551
DAT-2+ (ours)	×3	35.17	0.9331	31.13	0.8555	29.56	0.8166	30.28	0.8896	35.63	0.9555
DAT+ (ours)	×3	35.19	0.9334	31.17	0.8558	29.58	0.8173	30.30	0.8902	35.72	0.9559
EDSR [14]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
SRMDNF [27]	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
RDN [32]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
OISR [9]	×4	32.53	0.8992	28.86	0.7878	27.75	0.7428	26.79	0.8068	-	-
RCAN [30]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
NLRN [15]	×3	31.92	0.8916	28.36	0.7745	27.48	0.7306	25.79	0.7729	-	-
RNAN [31]	×3	32.43	0.8977	28.83	0.7871	27.72	0.7410	26.61	0.8023	31.09	0.9149
SRFBN [12]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
SAN [5]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
RFANet [16]	×4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.918
IGNN [34]	×4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN [23]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
CSNLN [22]	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
NLSA [21]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
CRAN [33]	×4	32.72	0.9012	29.01	0.7918	27.86	0.7460	27.13	0.8167	31.75	0.9219
ELAN [29]	×4	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
DFSA [17]	×4	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	0.9266
SwinIR [13]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
CAT-A [4]	×4	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
DAT-S (ours)	×4	33.00	0.9047	29.20	0.7962	27.97	0.7502	27.68	0.8300	32.33	0.9278
DAT-2 (ours)	×4	33.01	0.9047	29.21	0.7965	27.98	0.7509	27.86	0.8341	32.41	0.9285
DAT (ours)	×4	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
DAT-S+ (ours)	×4	33.06	0.9052	29.25	0.7968	27.99	0.7507	27.78	0.8316	32.50	0.9289
DAT-2+ (ours)	×4	33.07	0.9053	29.26	0.7973	28.01	0.7515	27.97	0.8358	32.60	0.9296
DAT+ (ours)	×4	33.15	0.9062	29.29	0.7983	28.03	0.7518	27.99	0.8365	32.67	0.9301

Table 6: Quantitative comparison with state-of-the-art methods. The best and second-best results are coloured red and blue.

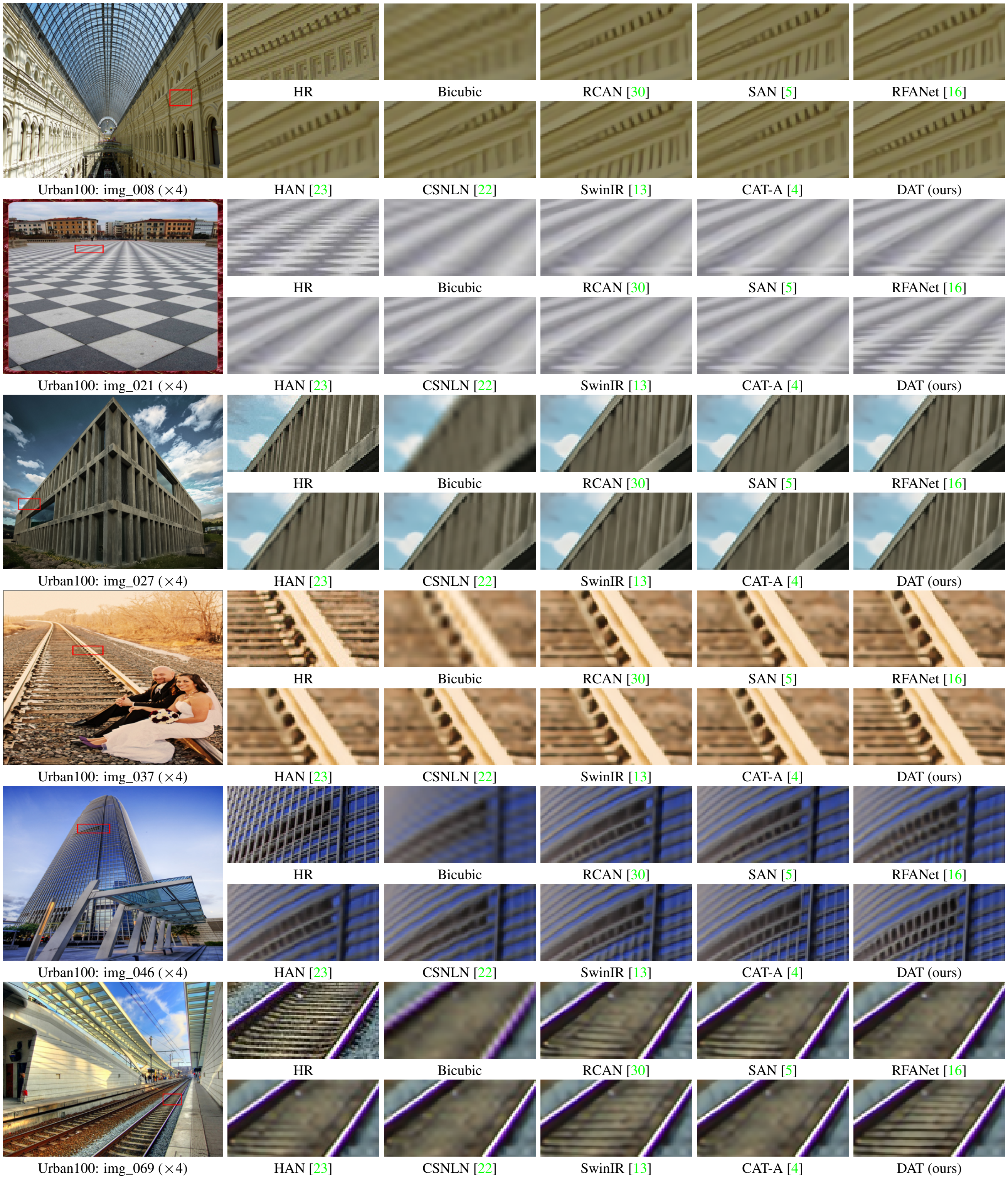


Figure 4: Visual comparison for image SR ($\times 4$) in some challenging cases.

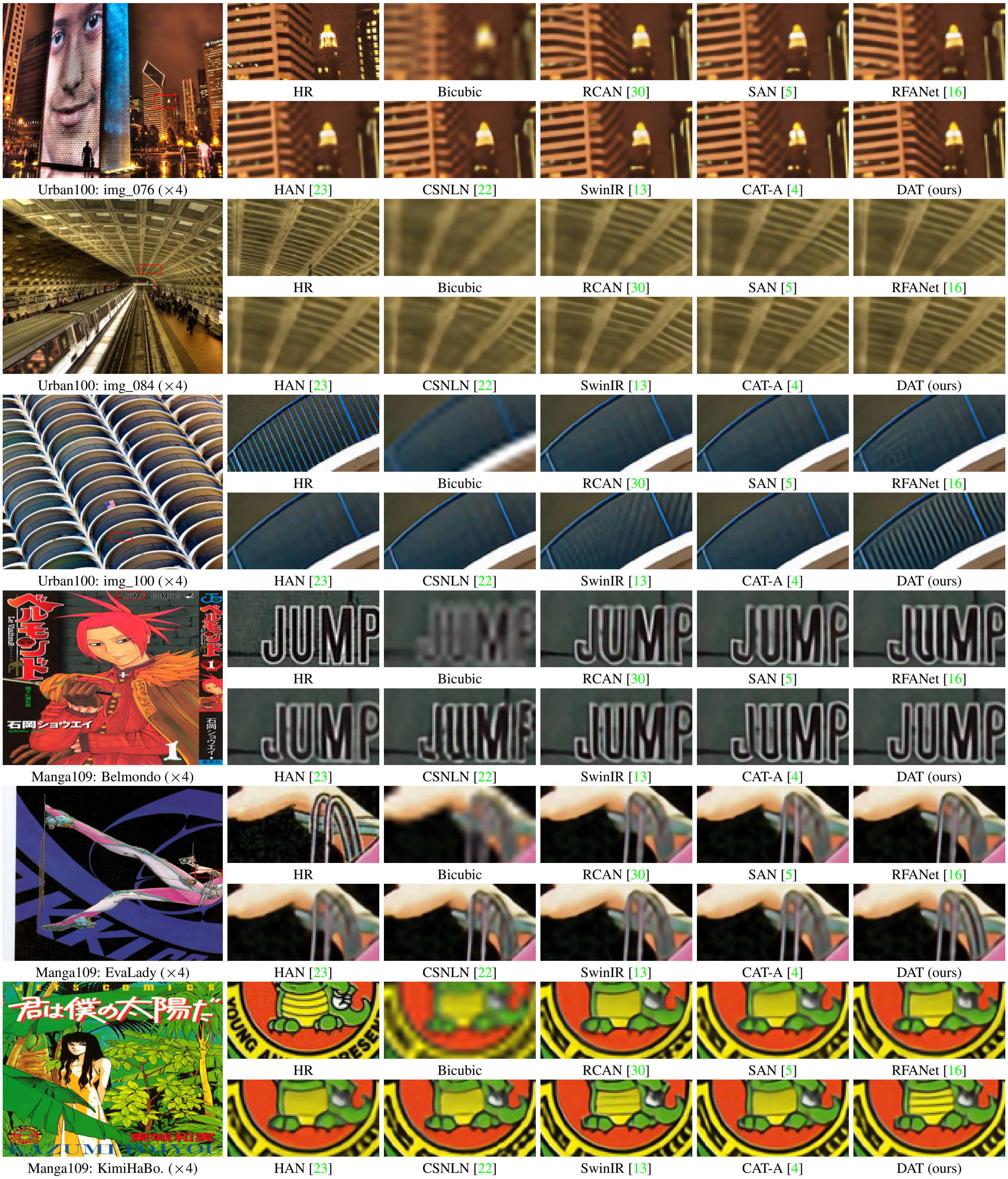


Figure 5: Visual comparison for image SR (×4) in some challenging cases.