NEURAL INFORMATION PROCESSING SYSTEMS

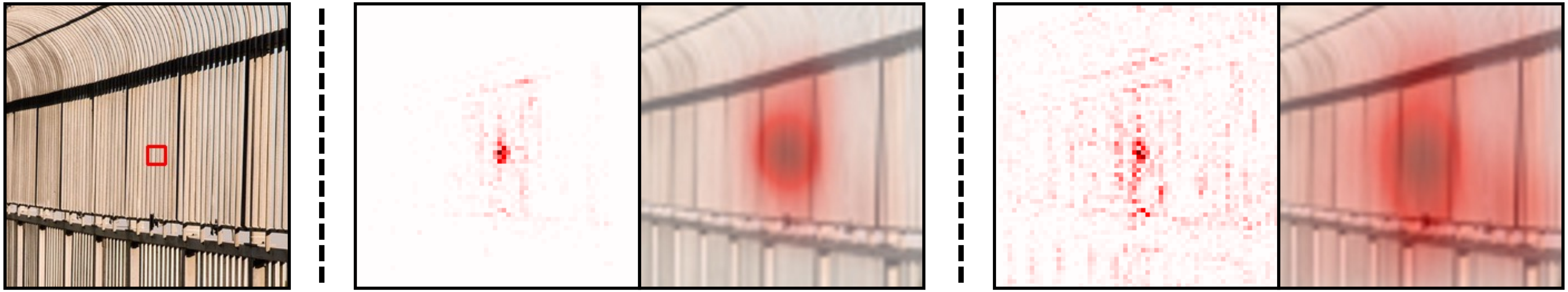
Cross Aggregation Transformer for Image Restoration

Zheng Chen[1], Yulun Zhang[2], Jinjin Gu[3,4], Yongbing Zhang[5], Linghe Kong[1*], Xin Yuan[6]

[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Shanghai AI Laboratory,
[4]The University of Sydney, [5]Harbin Institute of Technology (Shenzhen), [6]Westlake University

Introduction

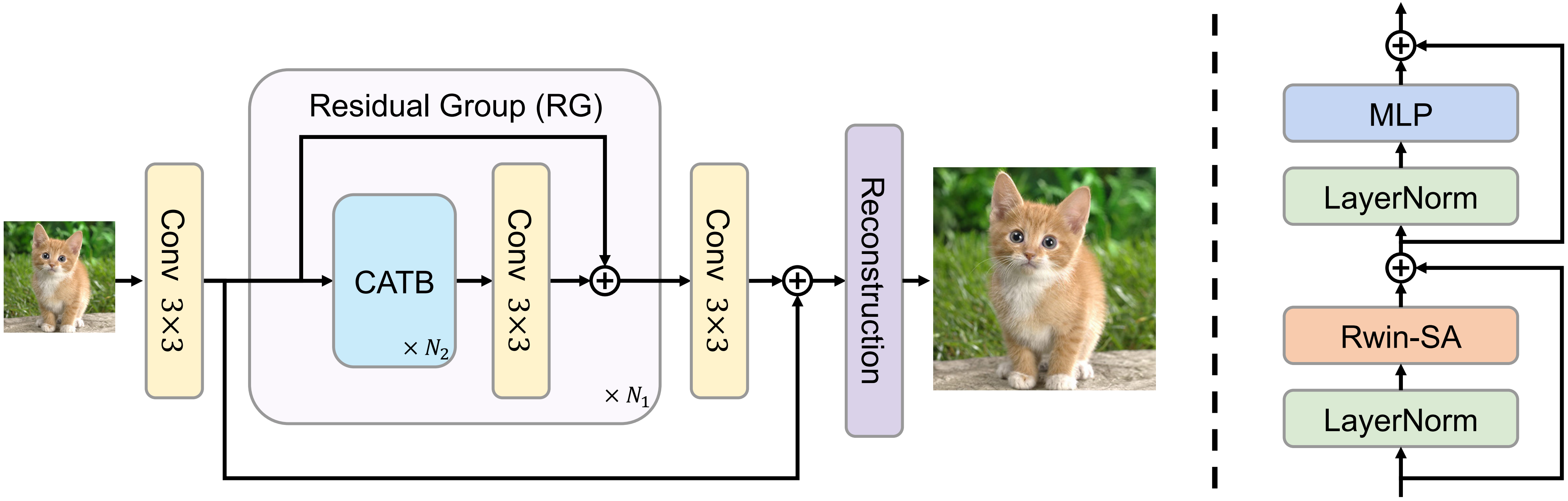
- Transformer achieves excellent performance in multiple vision tasks.
- Vanilla Transformer with quadratic complexity, $O(H^2W^2)$.
- Local square window attention to reduces Transformer complexity, but restricts the performance.
- We propose Cross Aggregation Transformer (CAT), utilizing the window self-attention and aggregating the features cross different windows.



HR SwinIR [1] CAT

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

Method

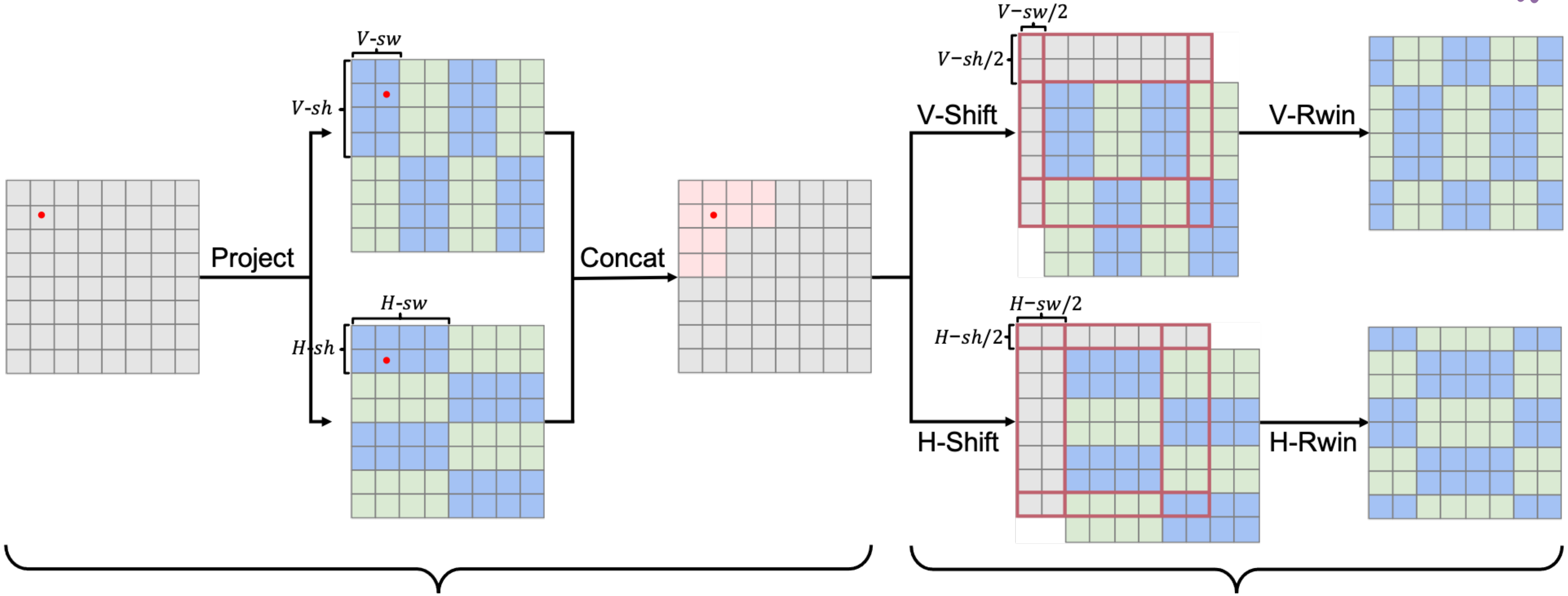


(a) Architecture of CAT (b) CATB

- Rectangle-window self-attention (Rwin-SA)
- Cross aggregation Transformer block (CATB)
- RCAN [2] backbone
- Cross Aggregation Transformer (CAT)

[2] 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.

Method



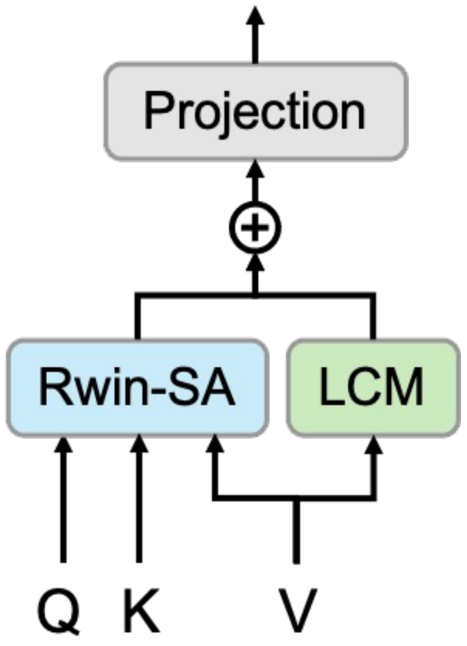
(a) Rectangle-Window Self-Attention

(b) Axial-Shift Operation

- Rwin (V-Rwin / H-Rwin), Axial-Shift (V-Shift / H-Shift)
- $4HWC^2 + 2(sw \times sh)HWC, O(HW)$
- Aggregate features across different windows
- Capture different features in horizontal and vertical directions

Method

- Axial-Rwin
- Cross aggregation (vertical + horizontal)
- $(sl, H), (W, sl)$
- $4HWC^2 + sl \times (H + W)HWC,$
  $O(HW(H + W))$



- Locality complementary module (LCM)
- (Depth-Wise) Convolution
- Operate on V without partition
- Global (self-attention) + Local (convolution)

Experiments

- Ablation study

Network	PSNR	SSIM	FLOPs
Sq. w/o shift	32.50	0.9325	281.8G
Sq. w/ shift	32.75	0.9347	281.8G
Re. w/o axial	32.66	0.9334	281.8G
Re. w/ axial	32.91	0.9360	281.8G

(a) Rectangle-Window Self-Attention

Network	PSNR	SSIM	FLOPs
C-R w/o LCM	32.91	0.9360	281.8G
C-R w/ LCM	32.98	0.9361	282.7G
C-A w/o LCM	33.01	0.9354	349.7G
C-A w/ LCM	33.11	0.9363	350.7G

(b) Locality Complementary Module

Network	PSNR	SSIM	FLOPs
C-R	32.98	0.9361	282.7G
C-A-1	32.97	0.9353	323.5G
C-A-2	33.11	0.9363	350.7G
C-A-3	33.20	0.9376	377.9G

(c) Window Size Impact

- (a) Rectangle window better than square window
- (b) LCM improves the performance
- (c) Larger window size, larger FLOPs, and better performance

Experiments

- Image SR

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [18]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [40]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [7]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IGNN [43]	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN [25]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
CSNLN [24]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [23]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
IPT [4]	×2	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [17]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
CAT-R (ours)	×2	38.48	0.9625	34.53	0.9251	32.56	0.9045	34.08	0.9443	40.09	0.9804
CAT-A (ours)	×2	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
CAT-R+ (ours)	×2	38.52	0.9627	34.59	0.9257	32.58	0.9047	34.19	0.9450	40.18	0.9805
CAT-A+ (ours)	×2	38.55	0.9628	34.81	0.9267	32.60	0.9048	34.34	0.9445	40.18	0.9806

- CAT-R (regular-Rwin), CAT-A (axial-Rwin)
- Obtain 0.45 dB gain over SwinIR

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

- JPEG Compression Artifacts Reduction

Dataset	q	RNAN [41]		RDN [42]		DRUNet [38]		SwinIR [17]		CAT (ours)		CAT+ (ours)	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LIVE1	10	29.63	0.8239	29.67	0.8247	29.79	0.8278	29.86	0.8287	29.89	0.8295	29.92	0.8299
	20	32.03	0.8877	32.07	0.8882	32.17	0.8899	32.25	0.8909	32.30	0.8913	32.32	0.8915
	30	33.45	0.9149	33.51	0.9153	33.59	0.9166	33.69	0.9174	33.73	0.9177	33.75	0.9179
	40	34.47	0.9299	34.51	0.9302	34.58	0.9312	34.67	0.9317	34.72	0.9320	34.74	0.9322
Classic5	10	29.96	0.8178	30.00	0.8188	30.16	0.8234	30.27	0.8249	30.26	0.8250	30.30	0.8257
	20	32.11	0.8693	32.15	0.8699	32.39	0.8734	32.52	0.8748	32.57	0.8754	32.60	0.8756
	30	33.38	0.8924	33.43	0.8930	33.59	0.8949	33.73	0.8961	33.77	0.8964	33.80	0.8966
	40	34.27	0.9061	34.27	0.9061	34.41	0.9075	34.52	0.9082	34.58	0.9087	34.60	0.9088

Method	q=10		q=20		q=30		q=40	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [11]	30.55	0.8841	33.12	0.9252	34.58	0.9418	35.50	0.9508
CAT	30.80	0.8875	33.38	0.9274	34.81	0.9432	35.73	0.9520
CAT+	30.89	0.8885	33.46	0.9280	34.88	0.9436	35.81	0.9523

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

- Real Image Denoising

Dataset		DANet+ [43]	CycleISP [45]	MIRNet [46]	MPRNet [47]	Uformer [39]	Restormer [44]	CAT (ours)	CAT+ (ours)
Parameters (M)		9.15	2.83	31.79	15.74	50.88	26.11	25.77	25.77
SIDD*	PSNR	39.47	39.52	39.72	39.71	39.89	40.02	40.01	40.05
	SSIM	0.9570	0.9571	0.9586	0.9586	0.9594	0.9603	0.9600	0.9602
DND	PSNR	39.58	39.56	39.88	39.82	39.98	40.03	40.05	40.08
	SSIM	0.9545	0.9564	0.9563	0.9540	0.9554	0.9564	0.9561	0.9563

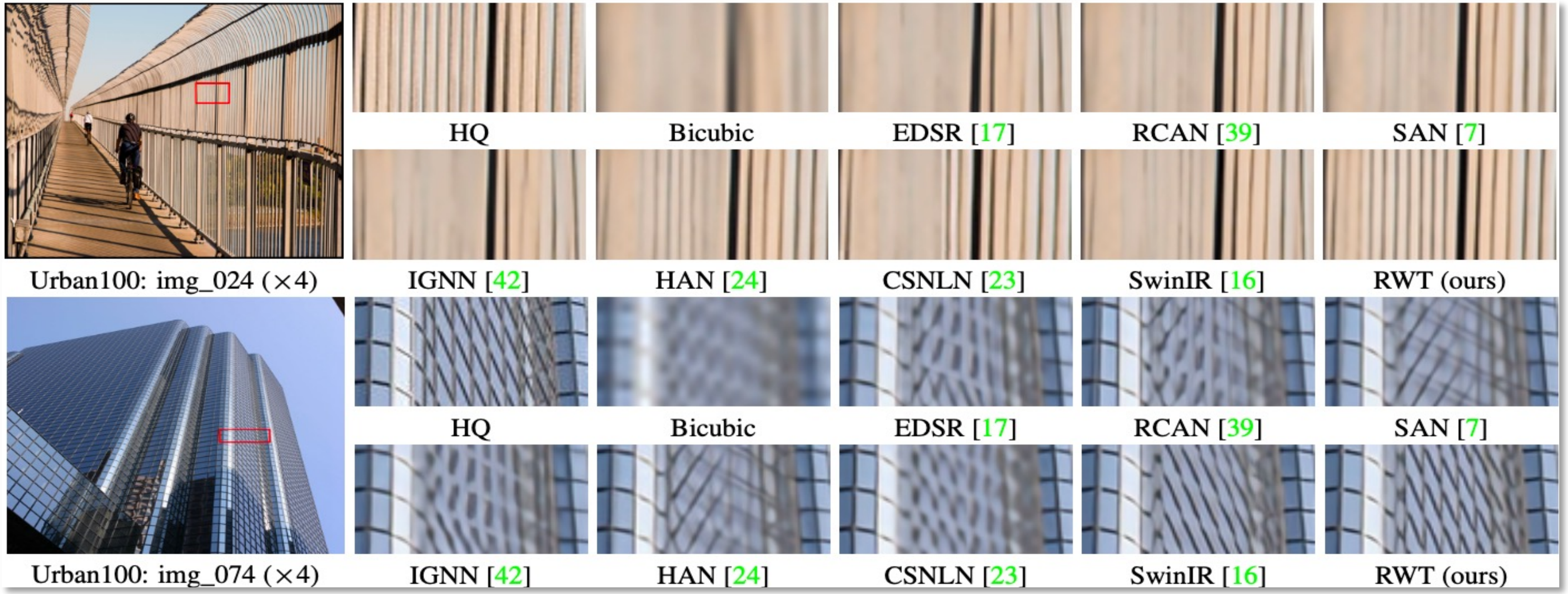
- Apply CATB to the U-Net architecture, following Restormer [3]
- Re-test the SIDD with all official pre-trained models
- Comparable performance with Restormer, fewer parameters

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

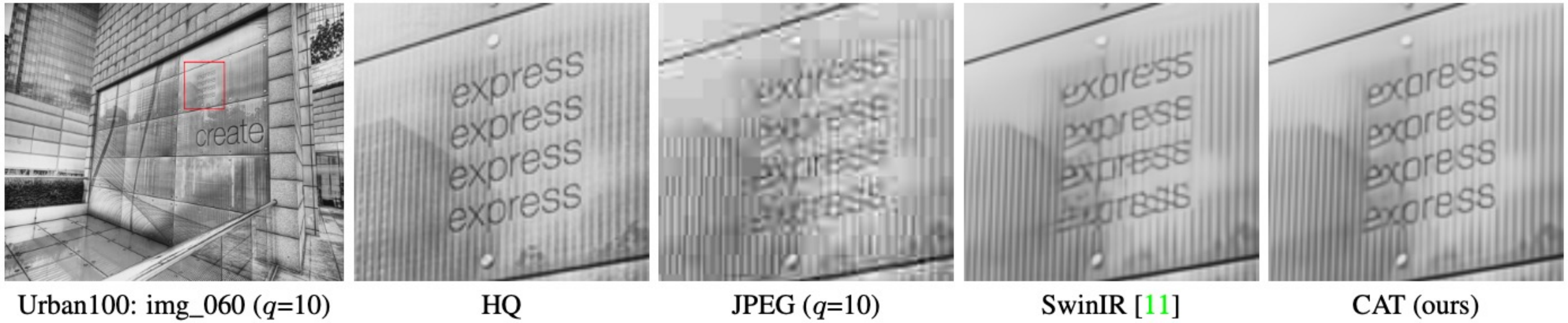
NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

- Image SR



- JPEG Compression Artifacts Reduction



Experiments

- Model Size Analyses

Method	EDSR [22]	RCAN [51]	HAN [31]	CSNLN [30]	SwinIR [21]	CAT-R (ours)	CAT-A (ours)	CAT-R-2 (ours)
PSNR (dB)	26.64	26.82	26.85	27.22	27.45	27.62	27.89	27.59
FLOPs (G)	823.3	261.0	269.1	84,155.2	215.3	292.7	360.7	216.3
Parameters (M)	43.09	15.59	16.07	6.57	11.90	16.60	16.60	11.93

- CAT-R, CAT-A, CAT-R-2 outperform other methods
- CAT-R-2 with similar computational complexity and parameters to SwinIR

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

- We propose a new Transformer model named cross aggregation Transformer (CAT) for image restoration.
- We propose a novel self-attention mechanism, named Rwin-SA, with axial-shift operation and the locality complementary module.
- SOTA performance on three classic image restoration tasks: image super-resolution, JPEG compression artifact reduction, and real image denoising.

Thanks

The code and models are available at: https://github.com/zhengchen1999/CAT