Activating More Pixels in Image Super-Resolution Transformer

Xiangyu Chen[1,2,3], Xintao Wang[4], Jiantao Zhou[1], Yu Qiao[2,3], Chao Dong[2,3]

[1]State Key Laboratory of Internet of Things for Smart City, University of Macau

[2]ShenZhen Key Lab of Computer Vision and Pattern Recognition,

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

[3]Shanghai Artificial Intelligence Laboratory [4]ARC Lab, Tencent PCG

Paper

Code

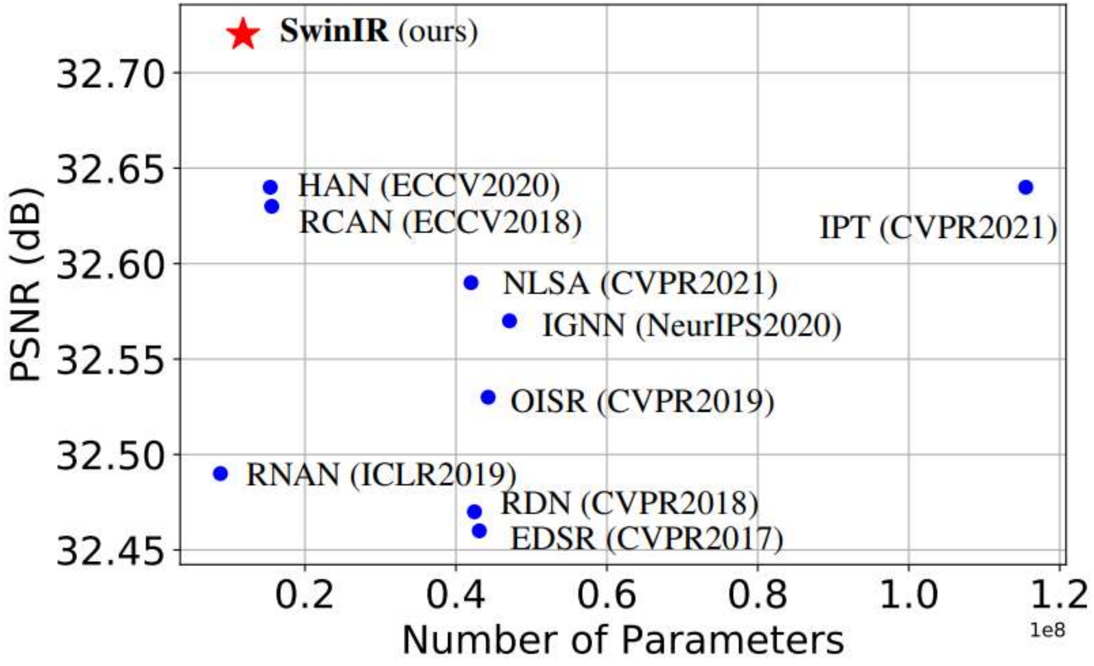
Overview

- We investigate the existing Transformer-based method for image super-resolution through attribution analysis and feature visualization.
- We propose a **H**ybrid **A**ttention **T**ransformer (**HAT**) that combines self-attention, channel attention and a novel overlapping cross-attention.
- We introduce the same-task pretraining strategy to exploit the potential of SR Transformer for further performance improvement.
- HAT achieves the **state-of-the-art** performance on image super-resolution that significantly outperforms existing methods.

Overview

×2 SR on Urban100
35.1
34.7
34.3
33.9
33.5
33.81
34.27
34.81
35.09
SwinIR EDT HAT HAT-L
×3 SR on Urban100
31.0
30.6
30.2
29.8
29.4
29.75
30.07
30.70
30.92
SwinIR EDT HAT HAT-L
×4 SR on Urban100
28.6
28.2
27.8
27.4
27.0
27.45
27.75
28.37
28.60
SwinIR EDT HAT HAT-L
×2 SR on Manga109
40.9
40.5
40.1
39.7
39.3
39.92
40.37
40.71
40.90
SwinIR EDT HAT HAT-L
×3 SR on Manga109
36.1
35.7
35.3
34.9
34.5
35.12
35.47
35.84
36.02
SwinIR EDT HAT HAT-L
×4 SR on Manga109
33.1
32.7
32.3
31.9
31.5
32.03
32.39
32.87
33.09
SwinIR EDT HAT HAT-L

Motivation

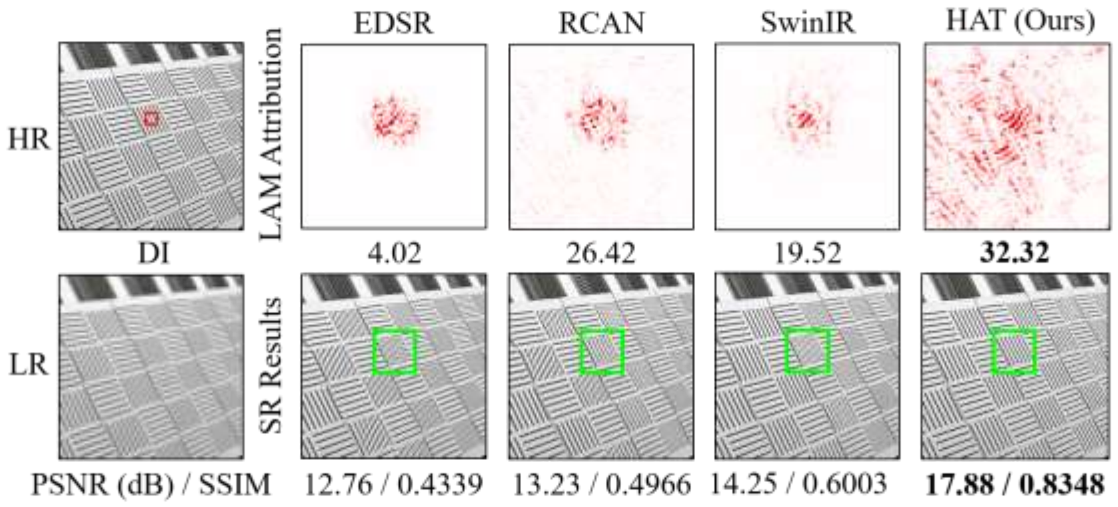


Since SwinIR obtains impressive performance on image SR, we want to know:

- ➢ Why does the Transformer-based model perform better than CNN-based methods?
- ➢ How to design a better SR Transformer to achieve greater performance breakthroughs?

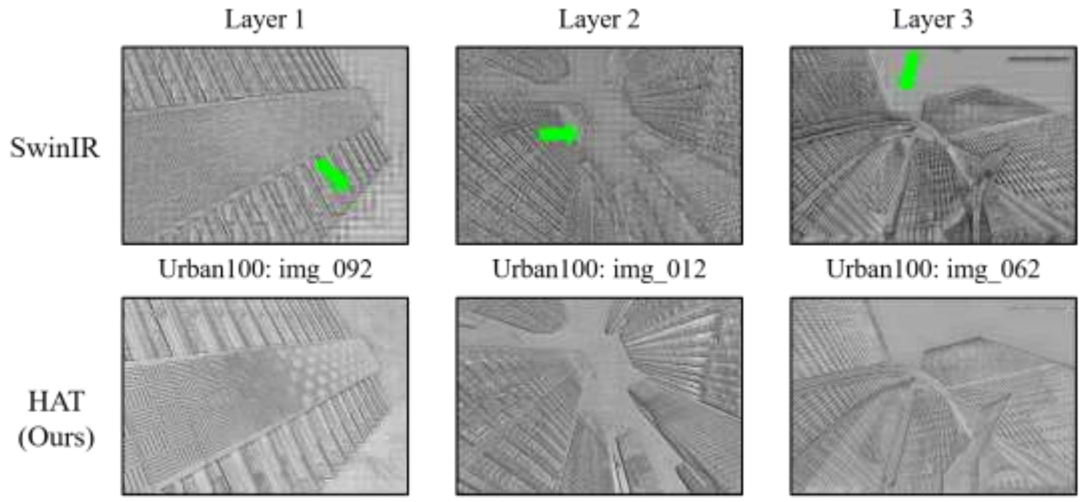
Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." ICCV. 2021.

Analysis



Thanks to the attribution analysis tool – **LAM**, we found that:

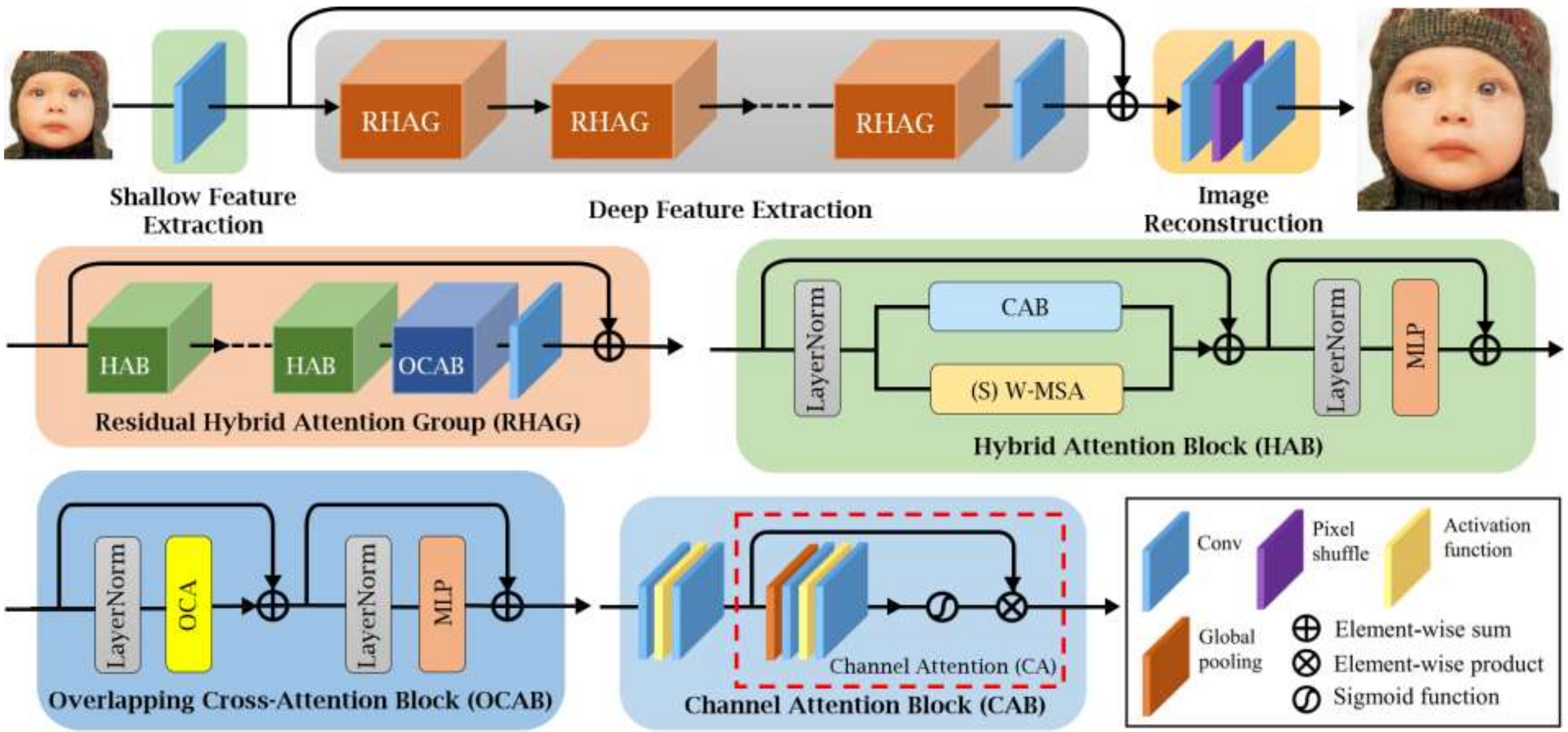
- SwinIR achieves better performance by utilizing fewer pixels, indicating that it has stronger local representation ability.
- SwinIR still restore wrong textures while RCAN obtain the correct results, suggesting that using more pixels may help.



We further observe the blocking artifacts in the intermediate features of SwinIR due to the window partition mechanism. We think that the cross-window interaction should be enhanced.

Gu, Jinjin, and Chao Dong. "Interpreting super-resolution networks with local attribution maps." CVPR. 2021.

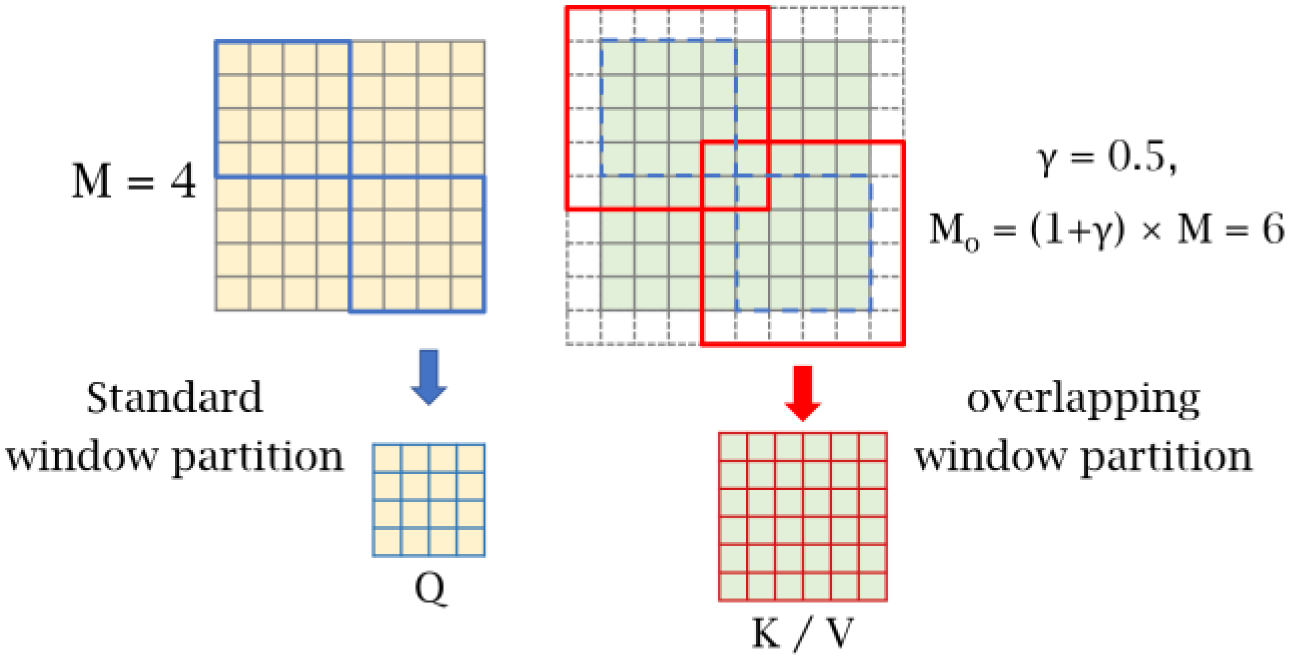
Proposed Method



The overall architecture of the proposed HAT.

Proposed Method

The window partition for the proposed overlapping cross-attention.



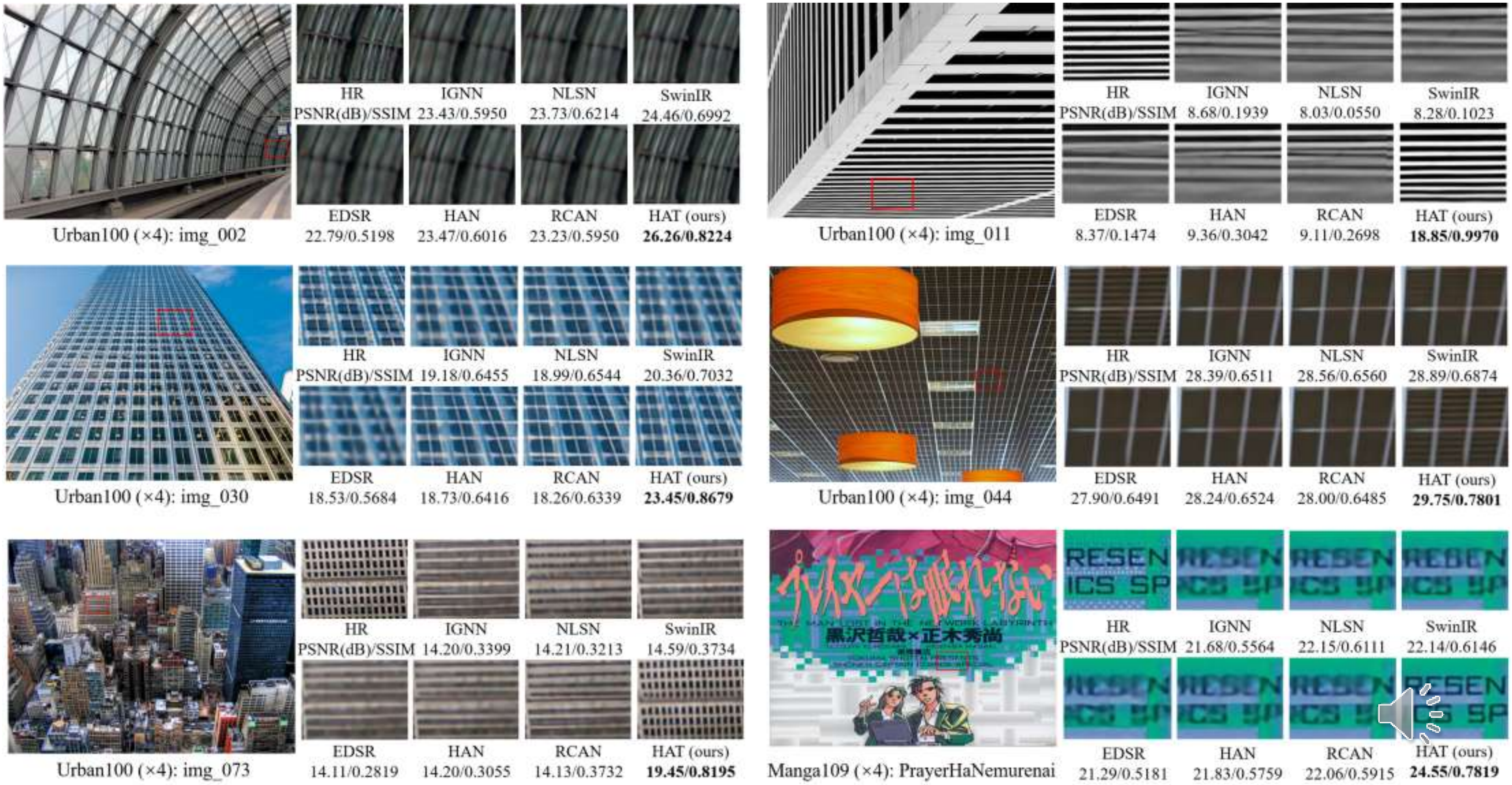
OCA computes **K**ey/**V**alue from a larger field where more information can be utilized for the **Q**uery directly.

We also introduce the same-task pre-training strategy by using large-scale dataset to further exploit the potential of SR Transformer.

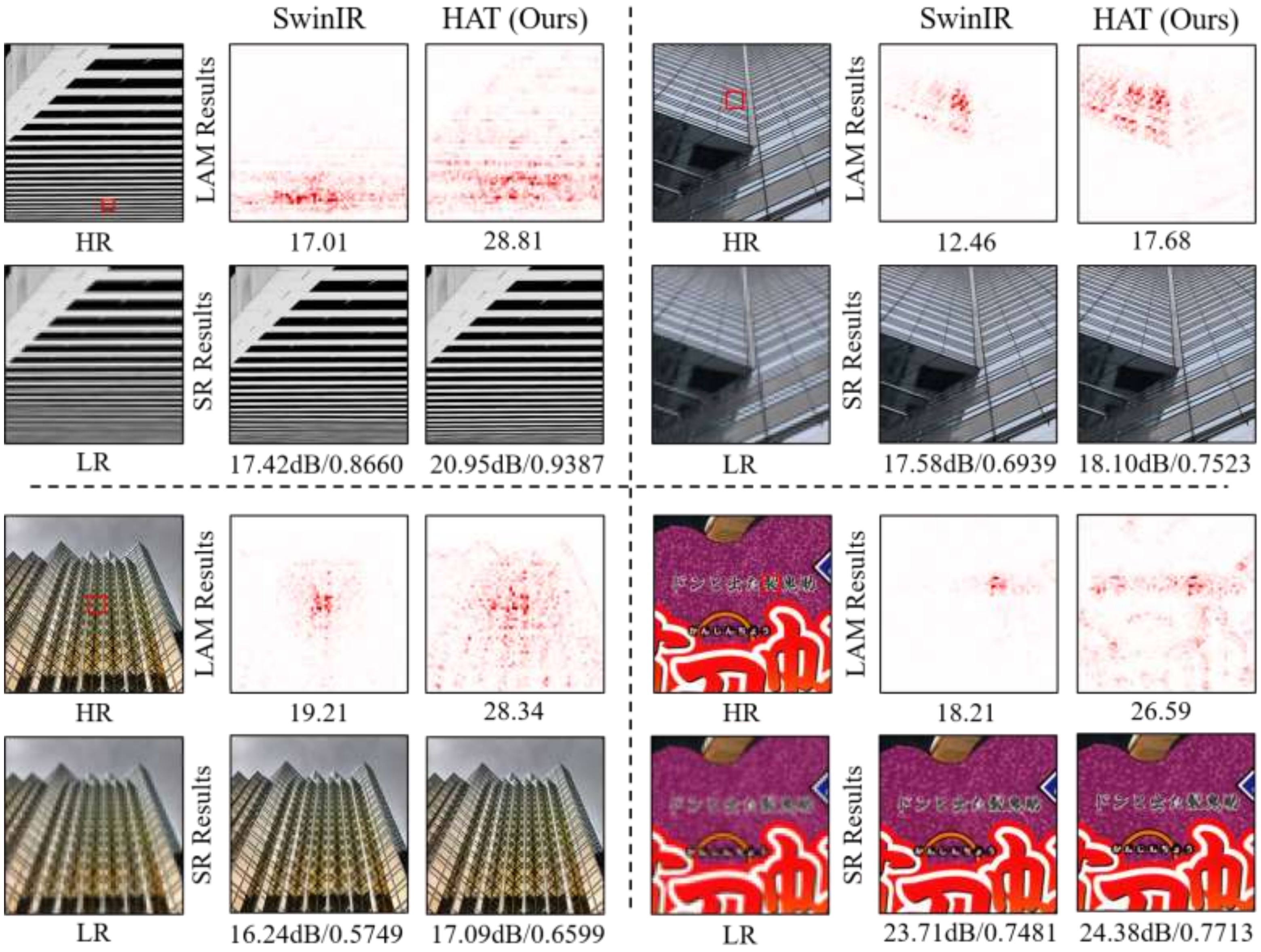
Quantitative Comparison

Method	Scale	Training Dataset	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	×2	DIV2K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN	×2	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN	×2	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IGNN	×2	DIV2K	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN	×2	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSN	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
RCAN-it	×2	DF2K	38.37	0.9620	34.49	0.9250	32.48	0.9034	33.62	0.9410	39.88	0.9799
SwinIR	×2	DF2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
EDT	×2	DF2K	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
HAT-S (ours)	×2	DF2K	38.58	0.9628	34.70	0.9261	32.59	0.9050	34.31	0.9459	40.14	0.9805
HAT (ours)	×2	DF2K	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
IPT†	×2	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
EDT†	×2	DF2K	38.63	0.9632	34.80	0.9273	32.62	0.9052	34.27	0.9456	40.37	0.9811
HAT† (ours)	×2	DF2K	38.73	0.9637	35.13	0.9282	32.69	0.9060	34.81	0.9489	40.71	0.9819
HAT-L† (ours)	×2	DF2K	38.91	0.9646	35.29	0.9293	32.74	0.9066	35.09	0.9505	41.01	0.9831
EDSR	×4	DIV2K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN	×4	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN	×4	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN	×4	DIV2K	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN	×4	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSN	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
RRDB	×4	DF2K	32.73	0.9011	28.99	0.7917	27.85	0.7455	27.03	0.8153	31.66	0.9196
RCAN-it	×4	DF2K	32.69	0.9007	28.99	0.7922	27.87	0.7459	27.16	0.8168	31.78	0.9217
SwinIR	×4	DF2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT	×4	DF2K	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
HAT-S (ours)	×4	DF2K	32.92	0.9047	29.15	0.7958	27.97	0.7505	27.87	0.8346	32.35	0.9283
HAT (ours)	×4	DF2K	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
IPT†	×4	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
EDT†	×4	DF2K	33.06	0.9055	29.23	0.7971	27.99	0.7510	27.75	0.8317	32.39	0.9283
HAT† (ours)	×4	DF2K	33.18	0.9073	29.38	0.8001	28.05	0.7534	28.37	0.8447	32.87	0.9319
HAT-L† (ours)	×4	DF2K	33.30	0.9083	29.47	0.8015	28.09	0.7551	28.60	0.8498	33.09	0.9335

Visual Comparison



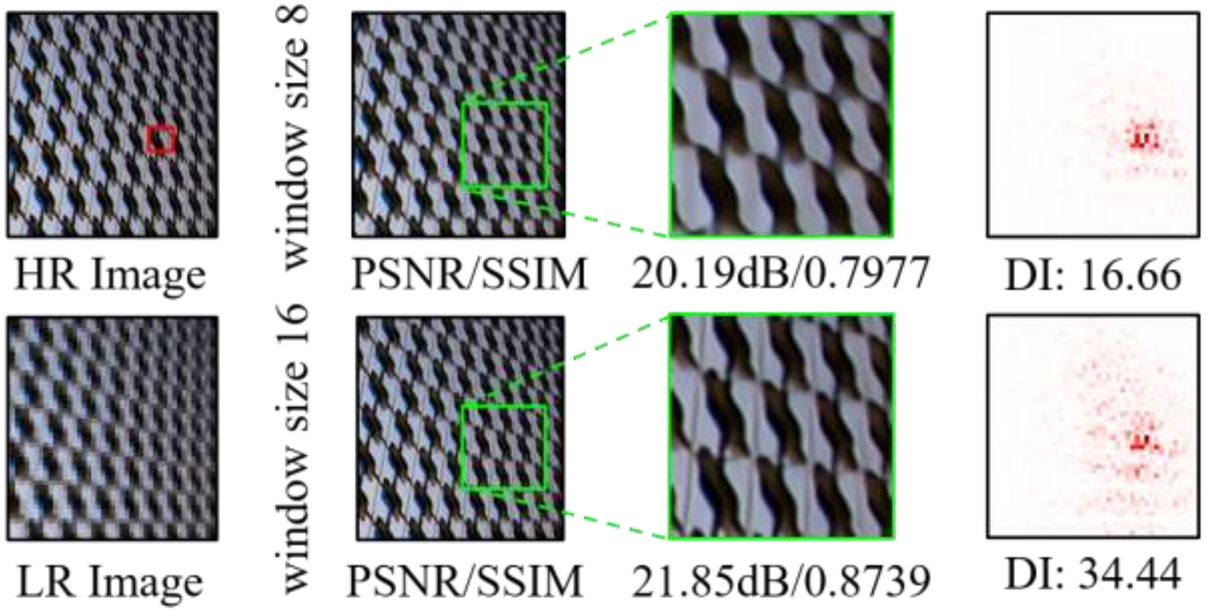
LAM Comparison



Ablation Study

Effects of different window sizes.

Size	Set5	Set14	BSD100	Urban100	Manga109
(8,8)	32.88	29.09	27.92	27.45	32.03
(16,16)	32.97	29.12	27.95	27.81	32.15



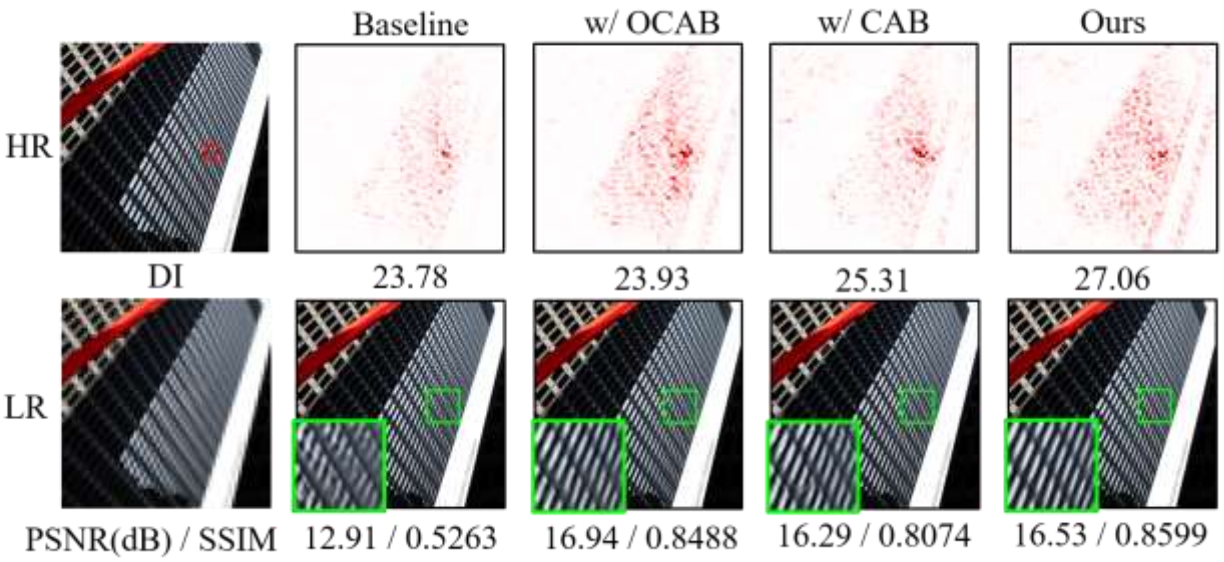
We investigate the effects of different window size on the performance and the utilized range of information.

We can observe that the model with larger window size has much better performance and activates more pixels for the reconstruction.

Ablation Study

Ablation study on OCAB and CAB.

	Baseline			
OCAB	X	✓	X	✓
CAB	X	X	✓	✓
PSNR	27.81dB	27.91dB	27.91dB	27.97dB



We investigate the effects of OCAB and CAB in HAT on the performance and the utilized range of information.

We can observe that both OCAB and CAN enlarge the utilized range of information and obtain great performance gains.

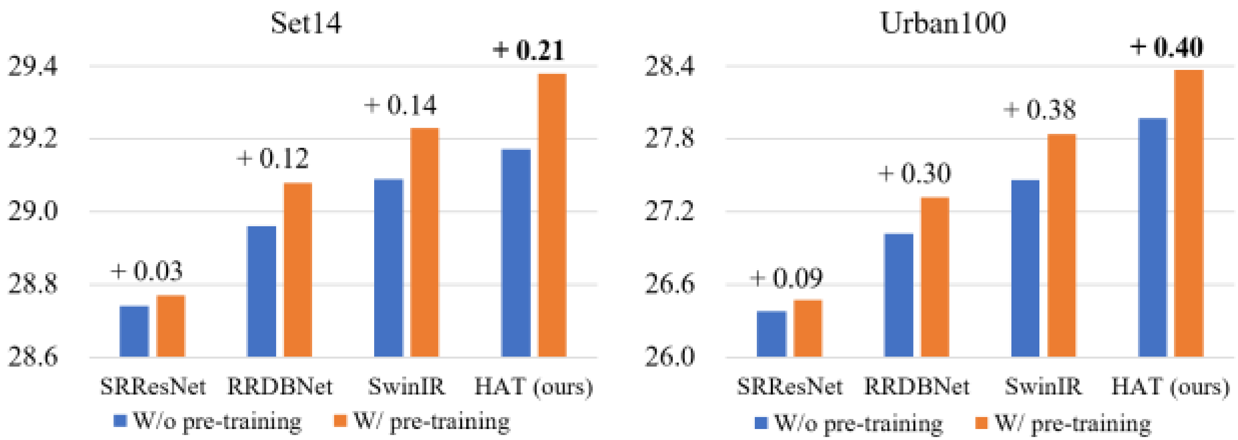
Study on Pre-training Strategy

Comparison between the multi-related-task pre-training in EDT and our proposed same-task pre-training.

Strategy	Stage	Set5	Set14	Urban100
Multi-related-task pre-training	pre-training	32.94	29.17	28.05
	fine-tuning	33.06	29.33	28.21
Same-task pre-training(ours)	pre-training	33.02	29.20	28.11
	fine-tuning	33.07	29.34	28.28

Compared to the multi-related-task pre-training, our same-task pre-training obtains better performance in the pre-training and the fine-tuning stages.

Effects of the pre-training for different networks.



All networks can benefit from the pre-training strategy. For the same type of network (i.e., CNN or Transformer), the larger the network capacity, the more performance gain.

Li, Wenbo, et al. "On efficient transformer and image pre-training for low-level vision." arXiv preprint arXiv:2112.10175. 2021.

Model Complexity Analysis

Model complexity comparison of window sizes.

window size	#Params.	#Multi-Adds.	PSNR
(8, 8)	11.9M	53.6G	27.45dB
(16, 16)	12.1M	63.8G	27.81dB

Enlarging window size can bring a large performance gain (+0.36dB) with a little increase in parameters and ∼%19 increase in Multi-Adds.

Model complexity comparison of window sizes.

Method	#Params.	#Multi-Adds.	PSNR
Baseline	12.1M	63.8G	27.81dB
w/ OCAB	13.7M	74.7G	27.91dB
w/ CAB	19.2M	92.8G	27.91dB
Ours	20.8M	103.7G	27.97dB

The proposed OCAB can bring a noticeable performance improvement with limited computation increase.

Model complexity comparison of SwinIR and HAT.

Method	#Params.	#Multi-Adds.	PSNR
SwinIR	11.9M	53.6G	27.45dB
HAT-S (ours)	9.6M	54.9G	27.80dB
SwinIR-L1	24.0M	104.4G	27.53dB
SwinIR-L2	23.1M	102.4G	27.58dB
HAT (ours)	20.8M	103.7G	27.97dB

HAT-S achieves much better performance than SwinIR with fewer params and similar computations. Simply Enlarging SwinIR cannot obtain comparable performance to our proposed HAT.

Conclusion

- We propose a novel Hybrid Attention Transformer HAT for image super-resolution.
- HAT combines channel attention and self-attention to activate more pixels for reconstruction.
- We introduce an overlapping cross-attention module to enhance the cross-window interaction.
- We further provide a same-task pre-training strategy to exploit the potential of SR Transformer.
- HAT achieves the state-of-the-art performance that significantly outperforms existing methods.

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

中国科学院深圳先进技术研究院
SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

腾讯 Tencent

ARC Applied Research Center

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Activating More Pixels in Image Super-Resolution Transformer

Xiangyu Chen[1,2,3], Xintao Wang[4], Jiantao Zhou[1], Yu Qiao[2,3], Chao Dong[2,3]

[1]State Key Laboratory of Internet of Things for Smart City, University of Macau

[2]ShenZhen Key Lab of Computer Vision and Pattern Recognition,

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

[3]Shanghai Artificial Intelligence Laboratory [4]ARC Lab, Tencent PCG

Paper

Code

Group

XPixel