Supplementary File:
Image Super-Resolution with Non-Local Sparse Attention

Yiqun Mei, Yuchen Fan, Yuqian Zhou
University of Illinois at Urbana-Champaign

1. Comparison with Criss-Cross Attention

Criss-Cross Attention [6] is a widely used variant of non-local attention for high-level vision tasks, which also has a sparse attention pattern. Specifically, it sums over only pixels on the criss-cross path of the query point. Therefore, the sparse attention pattern is fixed and purely depends on locations. Every pixel can finally obtain full image dependencies by consecutively performing it two rounds.

Table 1. Comparison with Criss-Cross Attention on Set5 [1] (×2).

	baseline	Non-Local	Criss-Cross	NLSA
PSNR	37.78	37.86	37.83	**37.92**

Here we compare our Non-Local Sparse Attention (NLSA) with Criss-Cross Attention. Results are reported in Table 1. Although Criss-Cross Attention brings improvements over the baseline, both NLSA and standard Non-Local Attention outperform it. The best result is achieved by our approach. This shows that enforcing sparsity based on content similarity indeed makes better use of global information as compared to methods based on locations.

2. Running time and Memory Comparison

Here we present running time and peak memory consumption comparison with standard Non-Local Attention (NLA), following the settings in Table 5. Models are evaluated at two input sizes: 100×100 and 150×150. The running time is the average of 1K times on one RTX 2070. As shown in Table 2, NLSA significantly saves running time and memory consumption, demonstrating it is indeed a more efficient operation. In Table 3, we also provide an additional model-level comparison with previous state-of-the-art SAN.

3. Visualization of Attention Maps

To obtain a deeper understanding of our NLSA, we visualize the learned attention maps in Figure 1. For each example image, we select one point and show its corresponding correlation maps.

Table 2. Running time and memory comparison with NLA.

Method	100 × 100		150 × 150	
	Time (ms)	Mem (MB)	Time (ms)	Mem (MB)
NLA	9.6	730	47.3	3738
NLSA-r1	0.9	45	4.0	104
NLSA-r2	1.4	109	6.2	242
NLSA-r4	2.4	231	10.3	521
NLSA-r8	4.3	509	16.8	1074

Table 3. Efficiency comparison with SAN.

Method	100 × 100			150 × 150		
	Time (ms)	Mem (MB)	FLOPs (G)	Time (ms)	Mem (MB)	FLOPs (G)
SAN	320	1188	1120	1540	4289	5422
NLSN	**142**	**625**	**988**	**375**	**1199**	**2005**

As mentioned before, NLSA contains 4 rounds of independent attention and final responses take their weighted sum. Column 2 to 5 correspond to the maps of each attention round. It can be observed that each map keeps sparse but captures highly-correlated locations. The differences between maps is mainly because of the randomness in Spherical LSH. As shown in the last column, the final weighted attention map takes the union of most related locations and suppresses less correlated ones, resulting in a more robust and powerful operation.

4. More Visual Results

We present more qualitative results in Figure 2 and Figure 3 to demonstrate the superiority of our approach.

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 *Proceedings of the British Machine Vision Conference*, 2012. 1

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional net-

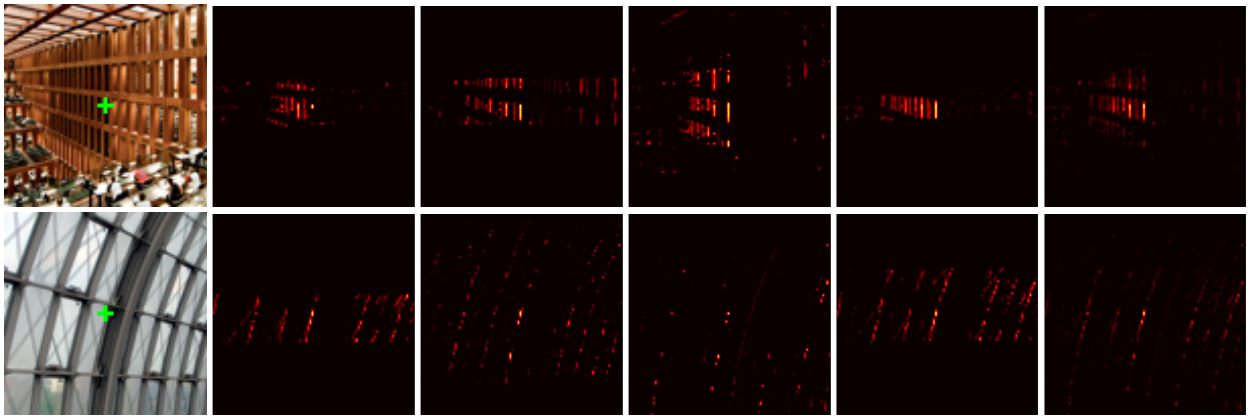

Figure 1. Visualization of attention maps of Non-Local Sparse Attention. Brighter color indicates higher engagement. Attention maps of each individual round are shown in column 2 to 5. The last column corresponds to the final weighted attention map. One can see each map keeps sparse and captures correlated regions. Further weighted summing them unions related locations while suppressing less related ones.

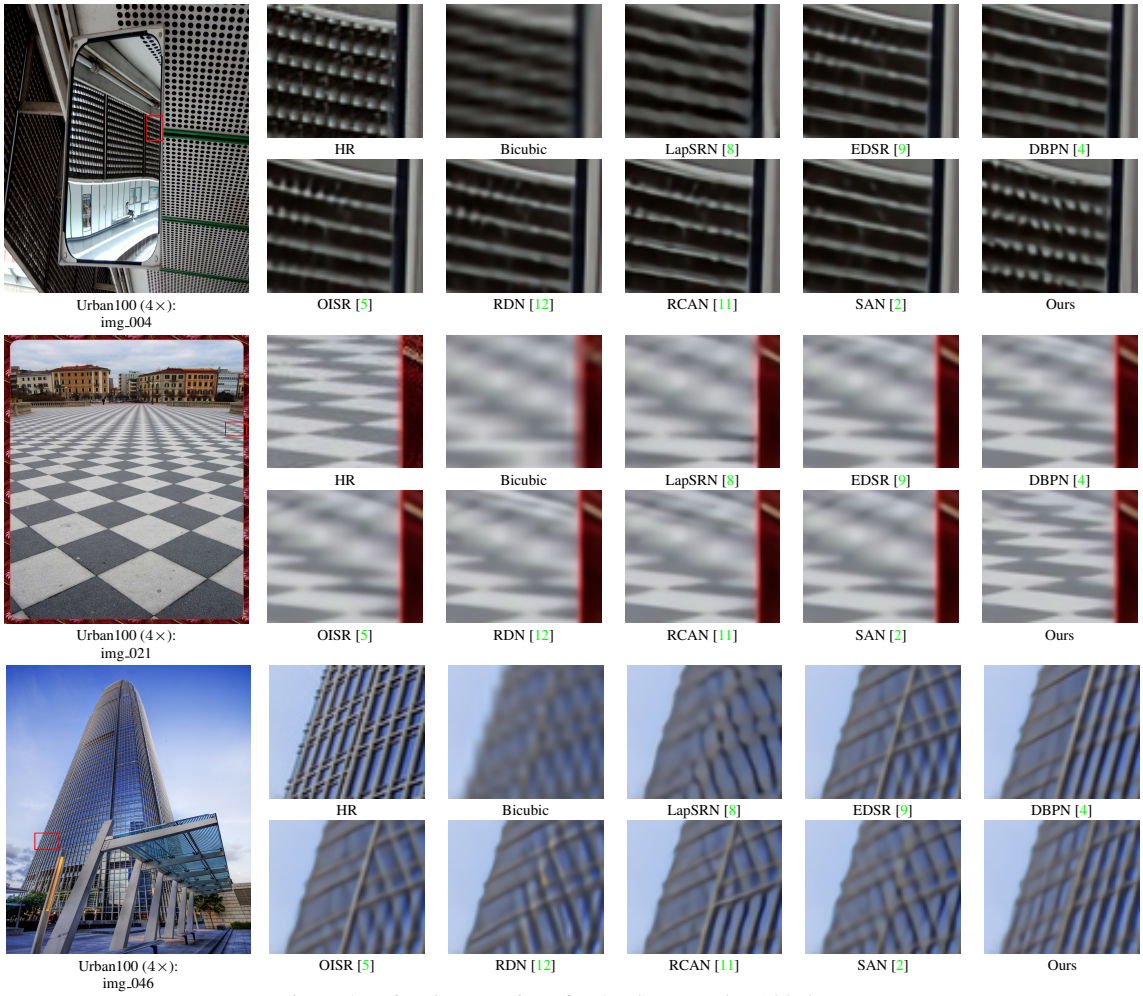

Figure 2. Visual comparison for $4\times$ SR on Urban100 dataset

works. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[4] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 2, 3

Figure 3. Visual comparison for 4× SR on Manga109 dataset

[5] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 2

[6] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3

[8] 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. 2, 3

[9] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 3

[10] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 3

[11] 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. 2, 3

[12] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2