Supplementary Materials of CiaoSR: Continuous Implicit Attention-in-Attention Network for Arbitrary-Scale Image Super-Resolution

Jiezhang Cao[1] Qin Wang[1] Yongqin Xian[1*] Yawei Li[1] Bingbing Ni[2] Zhiming Pi[2]
Kai Zhang[1†] Yulun Zhang[1†] Radu Timofte[1,3] Luc Van Gool[1,4]
[1]ETH Zürich [2]Huawei Inc. [3]University of Wurzburg [4]KU Leuven

https://github.com/caojiezhang/CiaoSR

Organization. In this paper, we organize our supplementary materials as follows. In Section A, we provide more implementation details and the network architectures of the proposed method. In Section B, we provide more results of our method. In Section C, we provide more visual comparisons. In Section D, the limitations and societal impacts of our proposed method are discussed.

A. More Training Details and Architectures

A.1. More Training Details

When testing on a large-scale image, we crop the image into multiple overlap patches of a smaller size (*e.g.*, 128 or 256), and then merge the SR results into the original large size. In the scale-aware non-local attention, if we use multi-scale features, we integrate the features via concatenation. For the real SR settings, we consider the following degradations: Gaussian blur, random resizing, random noise, JPEG compression. These degradations are the same as BSRGAN [9] and Real-ESRGAN [7]. In the training, we use the SwinIR as the backbone. To speed up the training, we remove the scale-aware non-local attention module.

A.2. Detailed Network Architectures

The decoding Query, Key and Value networks ϕ_q, ϕ_k and ϕ_v are a 5-layer MLP with ReLU activation and hidden dimensions of 256. We provide the detailed network architectures in Table A2. For the backbone RDN, we set the dimensions as follows. For the Query network ϕ_q, d_{in}=640, d_{out}=3. For the Key network ϕ_k, d_{in}=580, d_{out}=576. For the Value network ϕ_v, d_{in}=644, d_{out}=640. For the backbone SwinIR, we set the dimensions as follows. For the Query network ϕ_q, d_{in}=1800, d_{out}=3. For the Key network ϕ_k, d_{in}=1624, d_{out}=1620. For the Value network ϕ_v, d_{in}=1804, d_{out}=1800.

Next, we provide the detailed network architectures of the non-local attention in Table A1. For the backbone RDN, we set the dimensions as follows. For the Query and Key networks, φ_q, d_1=64, d_2=32. For the Value network φ_v, d_1=64, d_2=64. For the backbone SwinIR, we set the dimensions as follows. For the Query and Key networks, φ_q, d_1=180, d_2=90. For the Value network φ_v, d_1=180, d_2=180.

Table A1. The architecture of the Query ϕ_q, Key ϕ_k and Value ϕ_v networks for the local attention.

l-th layer	Layer information
0	Linear(d_{in}, 256, bias=True), ReLU( )
1	Linear(256, 256, bias=True), ReLU( )
2	Linear(256, 256, bias=True), ReLU( )
3	Linear(256, 256, bias=True), ReLU( )
4	Linear(256, d_{out}, bias=True), ReLU( )

Table A2. The architecture of the Query φ_q, Key φ_k and Value φ_v networks and the downsampling network for the scale-aware non-local attention.

l-th layer	Layer information
0	Conv2d(d_1, d_2, k=(1, 1), stride=(1, 1)
1	PReLU(num_parameters=1)
l-th layer	**Downsampling Layer information**
0	Linear(d_1, d_1, k = (3, 3), stride=(2, 2), padding=(1, 1)

B. More Results

Effect of local size. We investigate how the size of the local region affects the SR performance. We show the results with different local sizes in Figure B1. With a small size 1×1, the performance drops because the discontinuous pattern may appear within SR images. For a large size, all models are trained with sufficient iterations to ensure the convergence. Our model has comparable performance but introduces more computational cost. To trade-off the performance and computational cost, we set the local size as 2×2 in the experiment.

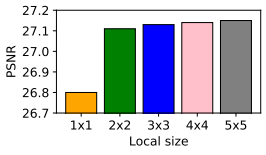

Figure B1. Impacts of local size.

*Currently with Google. This work was done at ETH Zürich.

†Corresponding Authors: Kai Zhang, cskaizhang@gmail.com; Yulun Zhang, yulun100@gmail.com

Computation cost. In Table B3, our proposed CiaoSR has the best performance, although it has large FLOPs. In addition, we also compare a variant of CiaoSR by removing the non-local attention, denoted by CiaoSR-L. It can trade-off the cost and the performance.

Table B3. Performance comparison on a 224×224 image.

RDN-Models	Meta-SR [3]	LIIF [2]	ITSRN [8]	LTE [5]	CiaoSR-L	CiaoSR
FLOPs (G)	560.56	722.52	1032.77	1415.21	960.88	2508.70
Memory (G)	10.54	4.73	6.90	6.18	8.83	19.13
PSNR (dB)	26.55	26.68	26.77	26.81	26.96	**27.11**

Performance of scales. In the main paper, we evaluate the effectiveness of our implicit model when trained with discrete (including single scale ($\{2\}/\{3\}/\{4\}$) and multiple scales $\{2, 3, 4\}$) and continuous scales $[1, 4]$. In Figure B2, we show the performance of the larger continuous scales $s\in[1, 8]$. The performance improvement is even larger when trained with scales $s\in[1, 8]$. For fair comparisons with other methods, we train the models with scales $s\in[1, 4]$.

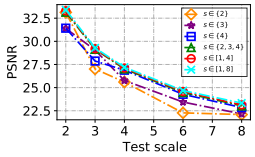

Figure B2. Impacts of scales.

Larger network. We construct a large network by removing the ensemble weights learning such that it has a comparable computational cost. Table B4 demonstrates the effectiveness of our proposed architecture.

Table B4. PSNR (dB) of large network and CiaoSR on Urban100.

Methods	×2	×3	×4	×6	×8	×12
Large network	33.07	29.00	26.85	24.30	22.89	21.25
CiaoSR	**33.30**	**29.17**	**27.11**	**24.58**	**23.13**	**21.42**

C. More Visual Results

We show more qualitative comparisons on Urban100 [4] and Manga109 [6] in Figure C3. Our model is able to synthesize the SR images with sharper textures than other methods. Taking the last line as an example, CiaoSR is able to restore the textures of the clothes buttons. Our results have sharper texture compared with other methods. Besides, we show more visual results on RealSRSet [9] in Figure C4. Our proposed method achieves comparable or better results than other methods.

D. Discussion on Limitations

In this paper, we propose a new continuous implicit attention-in-attention network, which achieves state-of-the-art performance. However, there may have some limitations, such as model efficiency. The main computational cost comes from searching the non-local attention features in the whole image. In the implementation, we reduce the searching regions. We leave the direction of boosting efficiency in the future.

References

[1] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *IEEE Conference on International Conference on Computer Vision*, pages 3086–3095, 2019. 4

[2] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. 2, 3

[3] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019. 2, 3

[4] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 2, 3

[5] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1929–1938, 2022. 2, 3

[6] 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*, 76(20):21811–21838, 2017. 2, 3

[7] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops*. 1, 4

[8] Jingyu Yang, Sheng Shen, Huanjing Yue, and Kun Li. Implicit transformer network for screen content image continuous super-resolution. *Advances in Neural Information Processing Systems*, 34:13304–13315, 2021. 2, 3

[9] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE Conference on International Conference on Computer Vision*, 2021. 1, 2, 4

[10] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 3

Figure C3. Visual comparison of different methods on benchmarks. "*" means the model first synthesizes twice to $\times 12$ images.

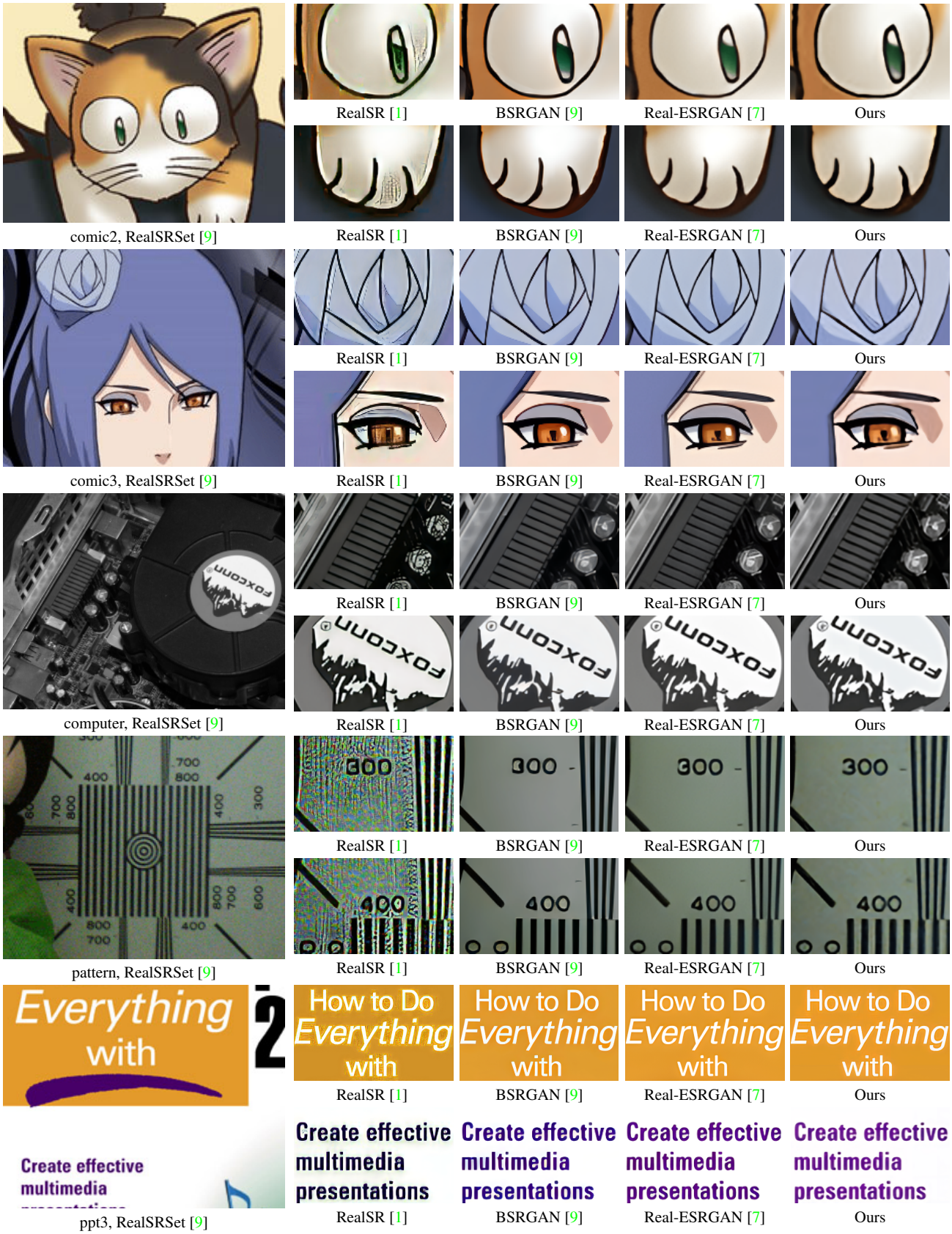

Figure C4. Visual comparison of different methods on the RealSRSet dataset [9] ($\times$16).