Supplementary File:
Second-order Attention Network for Single Image Super-resolution

Tao Dai[1,2,*,‡], Jianrui Cai[3,*], Yongbing Zhang[1], Shu-Tao Xia[1,2], Lei Zhang[3,4,§]
[1]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[2] PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4]DAMO Academy, Alibaba Group
{dait14, zhang.yongbing, xiast}@sz.tsinghua.edu.cn, {csjcai, cslzhang}@comp.polyu.edu.hk

Abstract

In this supplementary file, we first give more details about the forward propagation (FP) and backward propagatio (BP) of covariance normalization. Then we provide the convergence analysis of some very deep networks to verify the effectiveness of the proposed non-locally enhanced residual group (NLRG) structure. Meanwhile, we provide more comparison results with the state-of-the-art CNN-based SR methods under the Bicubic (BI) degradation model. All quantitative results are evaluated in terms of PSNR and SSIM metrics.

1. FP and BP of Covariance Normalization

FP of Newton-Schulz Iteration. Traditional convariance normalization relies heavily on EIG, which is however not well supported on GPU platform, thus leading to inefficient training. As explored in [7], we also relied on Newton-Schulz iteration to speed up the computation of covariance normalization. Given $\mathbf{Y}_0 = \mathbf{\Sigma}, \mathbf{Z_0} = \mathbf{I}$, for $n = 1, \cdots, N$, the Newton-Schulz iteration is then updated alternately as follows:

$$\begin{aligned} \mathbf{Y}_n &= \tfrac{1}{2}\mathbf{Y}_{n-1}(3\mathbf{I} - \mathbf{Z}_{n-1}\mathbf{Y}_{n-1}), \\ \mathbf{Z}_n &= \tfrac{1}{2}(3\mathbf{I} - \mathbf{Z}_{n-1}\mathbf{Y}_{n-1})\mathbf{Z}_{n-1}. \end{aligned} \quad (1)$$

After enough iterations, $\mathbf{Y}_n$ and $\mathbf{Z}_n$ quadratically converges to $\mathbf{Y}$ and $\mathbf{Y}^{-1}$. Such iterative operation is suitable for parallel implementation on GPU. In practice, one can achieve approximate solution with few iterations, *e.g.*, no more than 5 iterations in our method.

FP of Pre-normalization. Since Newton-Schulz iteration only converge locally, to guarantee the convergence, we pre-normalize $\mathbf{\Sigma}$ first via

$$\widehat{\mathbf{\Sigma}} = \frac{1}{\text{tr}(\mathbf{\Sigma})}\mathbf{\Sigma}, \quad (2)$$

where $\text{tr}(\mathbf{\Sigma}) = \sum_i^C \lambda_i$ denotes the trace of $\mathbf{\Sigma}$. In such case, it can be inferred that the $||\mathbf{\Sigma} - \mathbf{I}||_\mathbf{2}$ equals to the largest singular value of $(\mathbf{\Sigma} - \mathbf{I})$, i.e., $1 - \frac{\lambda_i}{\sum_i \lambda_i)}$ less than 1, which thus satisfies the convergence condition.

FP of Post-compensation. After Newton-Schulz iteration, we apply a post-compensation procedure to compensate the data magnitude caused by pre-normalization, thus producing the final normalized covariance matrix

$$\widehat{\mathbf{Y}} = \sqrt{\text{tr}(\mathbf{\Sigma})}\mathbf{Y}_N. \quad (3)$$

BP of Post-compensation. Given L the loss function and $\frac{\partial L}{\partial \widehat{\mathbf{Y}}}$, then the chain rule can be formulated as

$$\text{tr}((\frac{\partial L}{\partial \widehat{\mathbf{Y}}})^T \text{d}\widehat{\mathbf{Y}} = \text{tr}((\frac{\partial L}{\partial \mathbf{Y}_N})^T \text{d}\mathbf{Y}_N + (\frac{\partial L}{\partial \mathbf{\Sigma}})^T \text{d}\mathbf{\Sigma}), \quad (4)$$

where $\text{d}\widehat{\mathbf{Y}}$ is variation of $\widehat{\mathbf{Y}}$. After some simplifications, we can obtain

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{\Sigma}}|_{\text{post}} &= \frac{1}{2\sqrt{\text{tr}(\mathbf{\Sigma})}}\text{tr}\left(\left(\frac{\partial L}{\partial \widehat{\mathbf{Y}}}\right)^T \mathbf{Y}_N\right)\mathbf{I}, \\ \frac{\partial L}{\partial \mathbf{Y}_N} &= \sqrt{\text{tr}(\mathbf{\Sigma})}\frac{\partial L}{\partial \widehat{\mathbf{Y}}}. \end{aligned} \quad (5)$$

BP of Newton-Schulz iteration. The next step is to compute the derivatives of the loss function L w.r.t. $\frac{\partial L}{\partial \mathbf{Y}_n}$ and $\frac{\partial L}{\partial \mathbf{Z}_n}, n = N-1, \cdots, 1$, with $\frac{\partial L}{\partial \mathbf{Y}_N}$ obtained and $\frac{\partial L}{\partial \mathbf{Z}_N} = 0$.

[*] The first two authors contribute equally to this work.

[§] Corresponding author: Lei Zhang

[‡] This work is supported in part by the National Natural Science Foundation of China under Grant 61771273, the R &D Program of Shenzhen under Grant JCYJ20180508152204044, and the research fund of PCL Future Regional Network Facilities for Large-scale Experiments and Applications (PCL2018KP001).

Since $\mathbf{\Sigma}$ is symmetric, it can be easily inferred that $\mathbf{Y}_n$ and $\mathbf{Z}_n$ are also symmetric. Based on chain rules of BP and after some simplifications, $n = N, \cdots, 2$, we can have

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{Y}_{n-1}} &= \frac{1}{2}(\frac{\partial L}{\partial \mathbf{Y}_n}(3\mathbf{I} - \mathbf{Y}_{n-1}\mathbf{Z}_{n-1}) - \mathbf{Z}_{n-1}Y_{n-1}\frac{\partial L}{\partial \mathbf{Y}_n} \\ &\quad - \mathbf{Z}_{n-1}\frac{\partial L}{\partial \mathbf{Z}_n}\mathbf{Z}_{n-1}) \\ \frac{\partial L}{\partial \mathbf{Z}_{n-1}} &= \frac{1}{2}((3\mathbf{I} - \mathbf{Y}_{n-1}\mathbf{Z}_{n-1})\frac{\partial L}{\partial \mathbf{Z}_n} - \frac{\partial L}{\partial \mathbf{Z}_n}\mathbf{Z}_{n-1}Y_{n-1} \\ &\quad - \mathbf{Y}_{n-1}\frac{\partial L}{\partial \mathbf{Y}_n}\mathbf{Y}_{n-1}). \end{aligned} \tag{6}$$

The last step of this layer is associated with the partial derivative w.r.t. $\frac{\partial L}{\partial \widehat{\mathbf{\Sigma}}}$, which can be formulated as

$$\frac{\partial L}{\partial \widehat{\mathbf{\Sigma}}} = \frac{1}{2}\left(\frac{\partial L}{\partial \mathbf{Y}_1}\left(3\mathbf{I} - \widehat{\mathbf{\Sigma}}\right) - \frac{\partial L}{\partial \mathbf{Z}_1} - \widehat{\mathbf{\Sigma}}\frac{\partial L}{\partial \mathbf{Y}_1}\right). \tag{7}$$

BP of Pre-normalization. From Eqn. (3), we can see that we also need to compute the gradient of the L w.r.t. $\mathbf{\Sigma}$, backpropagated from the post-compensation layer. Based on Eqn. (3), $\frac{\partial L}{\partial \mathbf{\Sigma}}$ can be easily inferred. More details can be seen in the supplementary file. $\mathbf{\Sigma} = \text{tr}(\mathbf{\Sigma})\widehat{\mathbf{\Sigma}}$, and after some manipulations and we can thus obtain similar formulations:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{\Sigma}} &= -\frac{1}{(\text{tr}(\mathbf{\Sigma}))^2}\text{tr}\left(\left(\frac{\partial L}{\partial \widehat{\mathbf{\Sigma}}}\right)^T \mathbf{\Sigma}\right)\mathbf{I} + \frac{1}{\text{tr}(\mathbf{\Sigma})}\frac{\partial L}{\partial \widehat{\mathbf{\Sigma}}} \\ &\quad + \frac{\partial L}{\partial \mathbf{\Sigma}}|_{\text{post}}. \end{aligned} \tag{8}$$

Based on $\frac{\partial L}{\partial \mathbf{\Sigma}}$ obtained, the gradient of the loss function L w.r.t. the input $\mathbf{X}$ can be easily derived as follows:

$$\frac{\partial L}{\partial \mathbf{X}} = \bar{\mathbf{I}}\mathbf{X}\left(\frac{\partial L}{\partial \mathbf{\Sigma}} + \left(\frac{\partial L}{\partial \mathbf{\Sigma}}\right)^T\right). \tag{9}$$

2. Experiments

2.1. Convergence Analysis

We conduct experiments about convergence analysis of our non-locally enhanced residual group (NLRG). As shown in Fig. 1, The green line (NLRG_Base) denotes the NLRG structure with only one skip connection at the tail of each RG. In contrast, the blue lines (NLRG_SSC) denotes the each RG is connected through share-source skip connections (SSC). Based on NLRG_SSC, the **black** line (NLRG_FOCA) denotes the NLRG contains first-order channel attention (FOCA) in each RG. The red line (NLRG_SOCA) denotes the our proposed NLRG with second-order channel attention (SOCA) in each RG. All these four networks are trained from scratch. From Fig. 1 we can have some observations:

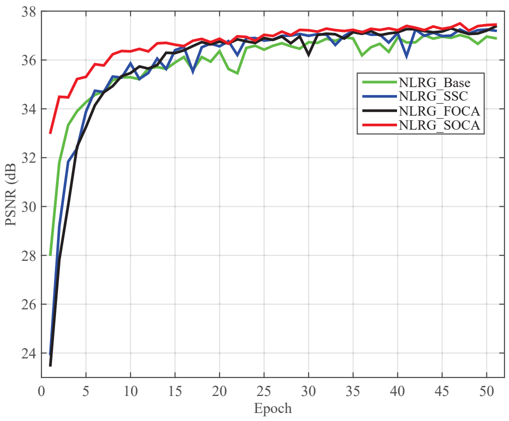

Figure 1. Convergence analysis on four variants of our proposed non-locally enhanced residual group (NLRG). All the four networks contain 20 residual groups (RG) and 10 residual blocks in each RG, thus producing over 400 residual blocks. The green line (NLRG_Base) denotes the NLRG structure with skip connections at the tail of each RG. In contrast, the blue lines (NLRG_SSC) denotes the each RG is connected through share-source skip connections (SSC). Based on NLRG_SSC, the black line (NLRG_FOCA) denotes the NLRG contains first-order channel attention (FOCA) in each RG. The red line (NLRG_SOCA) denotes the the NLRG contains our proposed second-order channel attention (SOCA) in each RG. The curves report PSNR values on Set5 (2×) in 50 epochs.

(1). Share-source skip connection plays a key role in the training of very deep CNN-based SR methods. We can see that NLRG_Base and NLRG_SSC both start at a relatively low performance, and gradually converge to the stable performance. With the increase of training epochs, NLRG_SSC would outperform NLRG_Base. This is mainly because the NLRG_SSC could allow more abundant LR information to be bypassed through share-source skip connections, which shows the effectiveness of the share-source skip connections.

(2). Channel attention is also important for further improving better SR performance. NLRG with first-order channel attention (NLRG_FOCA) or with second-order channel attention (NLRG_SOCA) both produce more stable PSNR curves and outperform NLRG_Base after some training epochs. The main reason is that channel attention exploits channel statistics among channels, thus enhancing the discriminative ability of the network.

(3). Second-order channel attention (SOCA) plays a more significant role in the training of deep networks. We can observe that the proposed NLRG_SOCA obtains the best performance. Specifically, it starts a relatively high and stable performance and converges faster than NLRG_FOCA, NLRG_SSC and NLRG_Base. These improvements mainly come from our proposed SOCA module, which exploits channel-wise feature statistics higher than first-order.

In summary, to build a deep trainable network for image SR, our NLRG structure plus share-source skip connections and second-order channel attention is a proper choice. The following experiments further demonstrate the effectiveness of our proposed non-locally enhanced attention networks (SAN).

2.2. Visual Results with Bicubic Degradation (BI)

For visual quality, We compare our SAN with several state-of-the-art CNN-based SR methods: SRCNN [2], FSRCNN [3], VDSR [5], LapSRN [6], EDSR [8], SRMD [9], DBPN [4], RDN [11] and RCAN [10]. As shown in Figs. 2-3, we can see that the early developed methods, such as SRCNN, VDSR, and LapSRN fail to restore the main structures and even produce blurry outputs, while RCAN and our SAN can output more faithful results. Compared with RCAN, our SAN can recover more image details.

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[3] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*. Springer, 2016.

[4] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep backprojection networks for super-resolution. In *CVPR*, 2018.

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[6] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, 2017.

[7] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *CVPR*, 2018.

[8] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

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

[10] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. 2018.

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

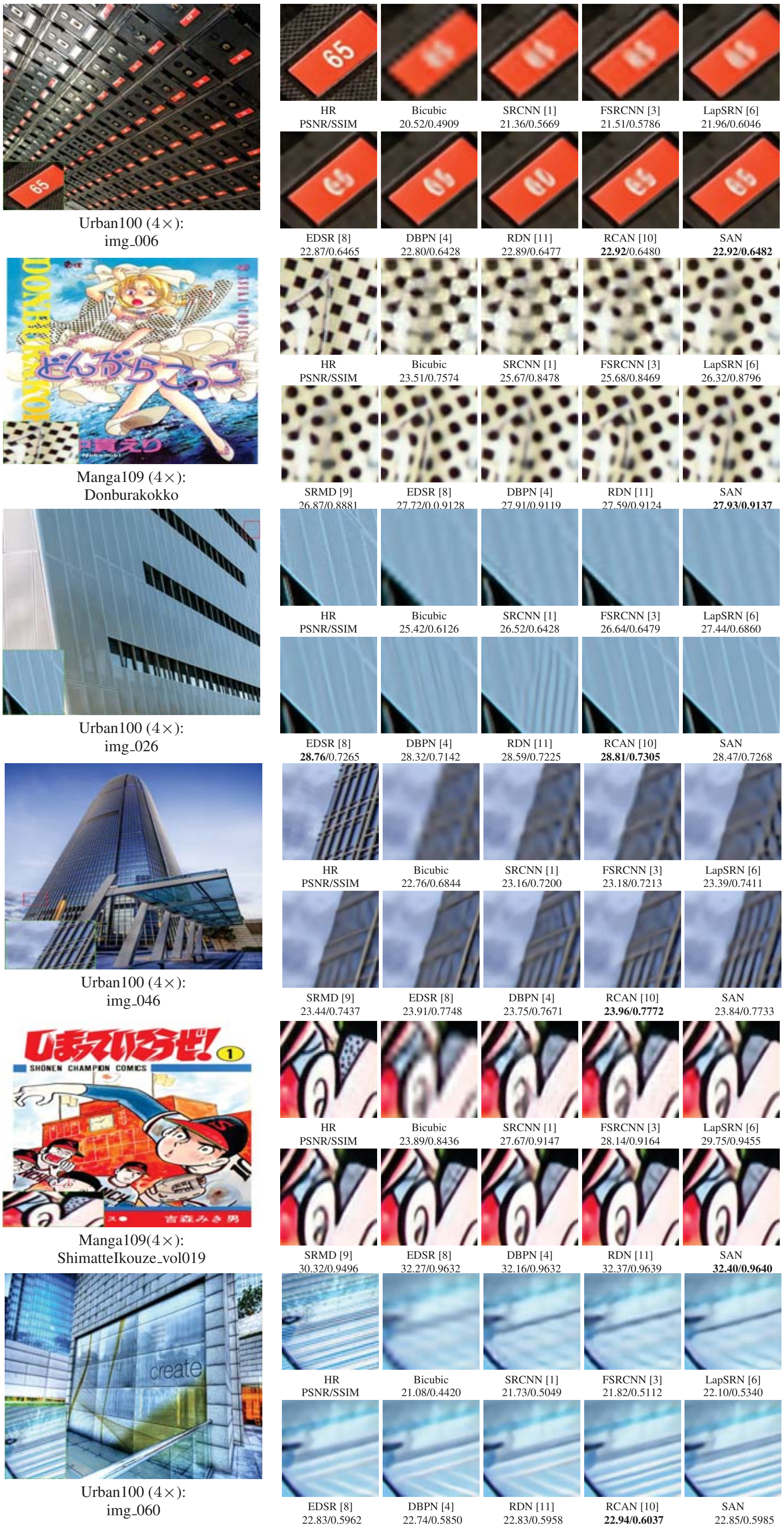

Figure 2. Visual comparison for 4× SR with BI model on Urban100 and Manga109 datasets. The best results are **highlighted**

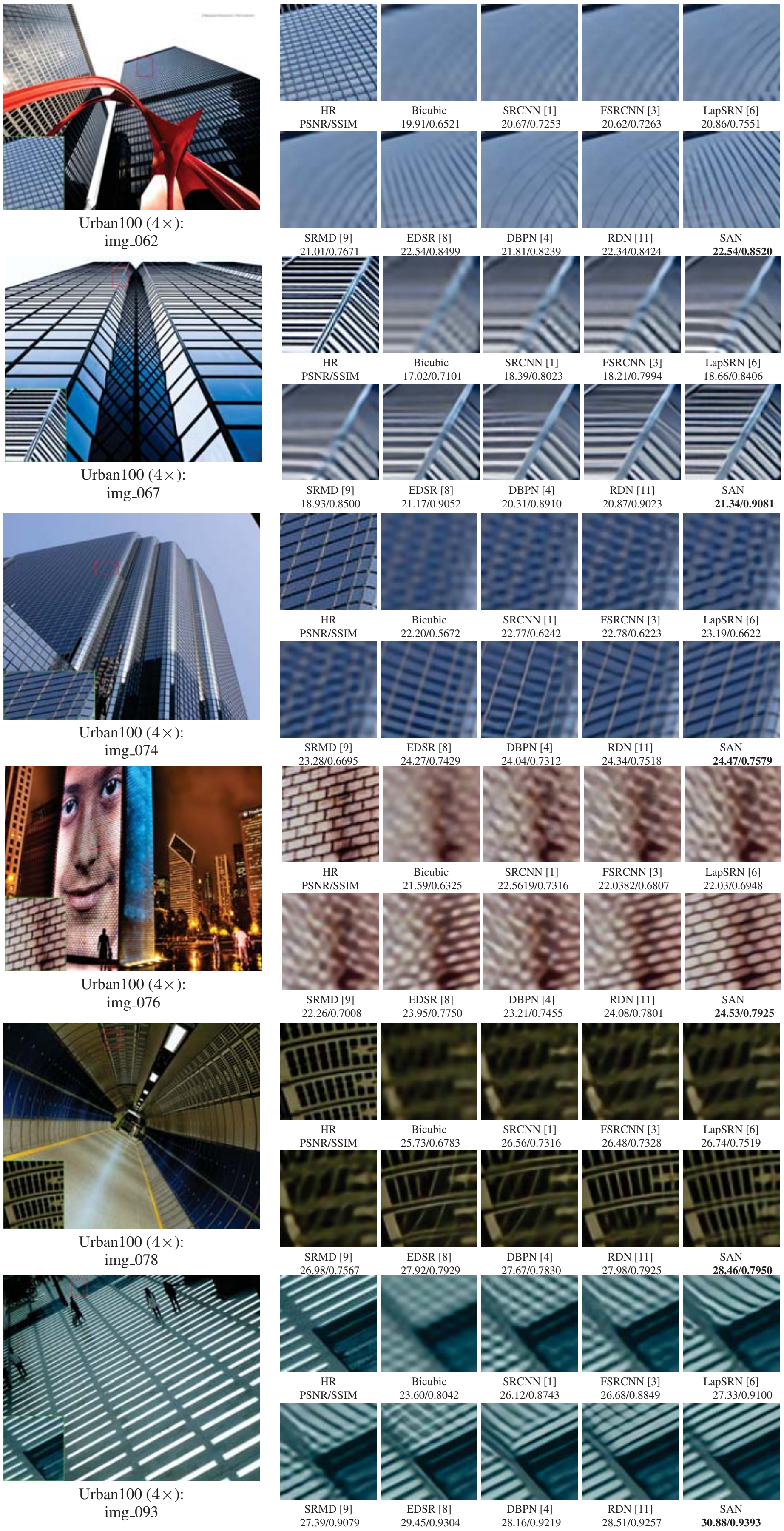

Figure 3. Visual comparison for 4× SR with BI model on Urban100 dataset. The best results are **highlighted**