Two-stage Network For Single Image Super-Resolution

Yuzhuo Han[1]*, Xiaobiao Du[2], Zhi Yang [3]
[1]Dalian University of Technology, China
[2]Zhuhai College of Jilin University, China
[3] Dibaocheng (Shanghai) Medical Imaging Technology Co., Ltd. China
yzhhan5@gmail.com, xbiaodu@163.com, stoneyang159@gmail.com

Abstract

The task of single-image super-resolution (SISR) is a highly inverse problem because it is very challenging to reconstruct rich details from blurred images. Most previous super-resolution (SR) methods based on the convolutional neural networks (CNN) tend to design more complex network structures to directly learn the mapping between low-resolution images and high-resolution images. However, this is not the best choice to blindly increase the network depth, because the performance improvement may not increase, but it will increase the computational cost. To solve this problem, we propose an effective method that learns high-frequency information in high-resolution images to enhance the image reconstruction. In this work, we propose a two-stage network (TSN) to recover clear SR images. The proposed TSN firstly learns the high-frequency information in high-resolution images, then learns how to transform to high-resolution images. A large number of experiments show that our TSN achieves satisfactory performance.

1. Introduction

Due to the limitations of hardware equipment and the real-time requirements of information transmission and processing, the image data that people obtain is often low-resolution (LR) images, but in practical applications, high-resolution (HR) images can provide more information and help professionals make better decisions. Accurate judgment, but also has a better perception effect. Single image super-resolution (SISR) utilizes the inherent relationship between the pixels in the image and the surrounding pixels, learns the implicit redundancy in the natural data, and can recover the missing detail information from an LR image, which can be obtained from the LR image. At present, SISR technology is widely used in many fields, such as social security[1], medical imaging, and military remote sensing. The existing SISR algorithms can be roughly divided into three categories: interpolation-based algorithms[2, 5], reconstruction-based algorithms[4], and learning-based algorithms[11, 13, 15, 18]. The algorithm based on interpolation is simple, but the reconstructed image will introduce artifacts and ringing. Although reconstruction-based algorithms have good reconstruction effects, they have low efficiency and are sensitive to scale scaling factors. The learning-based algorithm solves the problem of sensitivity to scaling factors and has been widely used in the SISR field. Since the SISR problem is inconsistent (there are multiple possible solutions, that is, multiple HR images correspond to the same LR image), certain conditions need to be constructed to constrain the solution space of the reconstructed image. As a batch of the learning algorithm, the deep learning-based SISR algorithm establishes a non-linear end-to-end mapping relationship between input and output through a multilayer convolutional neural network (CNN) (Convolutional Neural Network). The first SISR network model based on deep learning algorithms, SRCNN [7], constructed a simple shallow CNN, and the obtained image reconstruction effect was significantly improved compared to other SR reconstruction algorithms. In recent years, many deep learning super-resolution (SR) reconstruction models have been proposed, such as ESPCN [29], VDSR [25], and RCAN [33], which have further improved the SISR reconstruction effect. Since the SR reconstruction algorithm based on deep learning [28, 27, 26, 9] usually builds an end-to-end network model, the LR image input into the specific network model, optimize the loss function of the network by means of feature mapping and scale enlargement, and then obtain the HR image. We assume that if the low-resolution input with high-frequency information will be more robust to model. In order to achieve this hypothesis, we introduce residual channel attention network (RCAN) to transform HR images with high-frequency information to LR images with high-frequency information. We propose a two-stage network (TSN). The one stage is learning how to transform

*Corresponding authors.

LR images into LR images with high-frequency information. The other stage is learning how to transform LR images with high-frequency information into HR images. To make our network learn two-stage, we also propose a two-stage learning loss, which will guide our network jointly learning how to transform LR images into LR images with high-frequency information and transform LR images with high-frequency information into HR images.

In summary, the main contributions of this paper are listed as follows:

• We introduce residual channel attention groups [33] to transform HR images with high-frequency information to LR images with high-frequency information.

• We propose a two-stage network (TSN). The first stage is learning how to transform LR images into LR images with high-frequency information. The second stage is learning how to transform LR images with high-frequency information into HR images.

• We propose a two-stage learning loss, which will guide our network jointly learning how to transform LR images into LR images with high-frequency information and transform LR images with high-frequency information into HR images.

2. Related Work

2.1. Single Image Super-Resolution

Single image super-resolution is a low-level computer vision task. The popular method in our literature is learning the mapping function from low-resolution images to high-resolution images to reconstruct. Traditional machine learning techniques are widely applied in super-resolution, including kernel method[4], PCA [3], sparse-coding [32], learning embedding [5], etc. There is a powerful method to take full use of the image self-similarity without extra data. In order to obtain a super-resolution image, [13] use the patch redundancy to produce. Freedam et al. [11] further develop a localized searching method. [15] extend this algorithm to guide the patch search by using detected perspective geometry.

Current advances in SISR make full use of the powerful representation capability of convolution neural networks. Dong et al. [7] first proposed SRCNN to recovery high-resolution image. They interpret the architecture in CNN as extraction layer, non-linear mapping layer, and reconstruction layer, corresponding these steps in sparse coding [32]. DRCN [20] further these steps through firstly interpolating the low-resolution image to the desired size so that suffers from the huge computational complexity and some detail loss. Kim el al. [19, 20] adopt the deep residual convolutional neural network to achieve better performance, which use the bicubic interpolation to upsample the low-resolution image to the desired size and then fed in the network to output the super-resolution image. Since then, the deeper CNN-based super-resolution models is a trend to obtain superior performance, such as LapSRN [21], DRRN [30], SR-ResNet [22], EDSR [25] and RCAN [33].

Nevertheless, the depth of the network brings a huge amount of computation and increases the processing time. In order to solve this problem, Dong et al. adopt smaller filter sizes and a deeper network namely FSRCNN [8], which remove the bicubic interpolation layer in SRCNN and embedding the deconvolution layer at the tail of the FSRCNN. To reduce parameters, DRRN [30] proposed the combination of the residual skip connection and the recursive so that compromise the runtime speed. In order to utilize the multi-scale feature, [23] proposed MSRN model to capture the multi-scale feature at different scale sizes.

Although most of the CNN-based super-resolution methods strongly promote progress in this field, most of the advanced models blindly increase the depth and parameters of the network. It is clear that this method increases the running time and does not improve accuracy.

2.2. Attention Model

To human perception, attention generally means human visual systems focus on salient areas [17] and adaptively process visual information. Currently, several studies have proposed embedding attention mechanism processing to improve the performance of CNNs for different tasks, such as image segmentation, image, and video classification [14, 31]. Wang et al. proposed non-local neural network [31] for video classification, which incorporates non-local process to spatially attention long-range feature. Hu et al. proposed SENet [14] to capture channel-wise feature relationships to obtain better performance for image classification. Li et al. [24] proposed an expectation-maximization attention network for semantic segmentation, which borrowed EM algorithm to iteratively optimize parameters and decrease the complexity of the operation in non-local block. Huang et al. proposed criss-cross attention [16] for semantic segmentation, which can efficiently capture contextual patterns from long-range dependencies. Fu et al. proposed a dual attention network (DANet) [12], which mainly consists of the position attention module and the channel attention module. They use the position attention module to learn the spatial interdependencies. The channel attention module is designed to model channel interdependencies. It largely improves the segmentation results through capturing rich contextual dependencies. Zhang et al. are first introduce non-local blocks in single image super-resolution. The proposed residual non-local attention learning [34] to capture more detailed information through preserving more low-level features, being more suitable for super-resolution image reconstruction. The network pursues better network representational ability and achieves high-quality image recon-

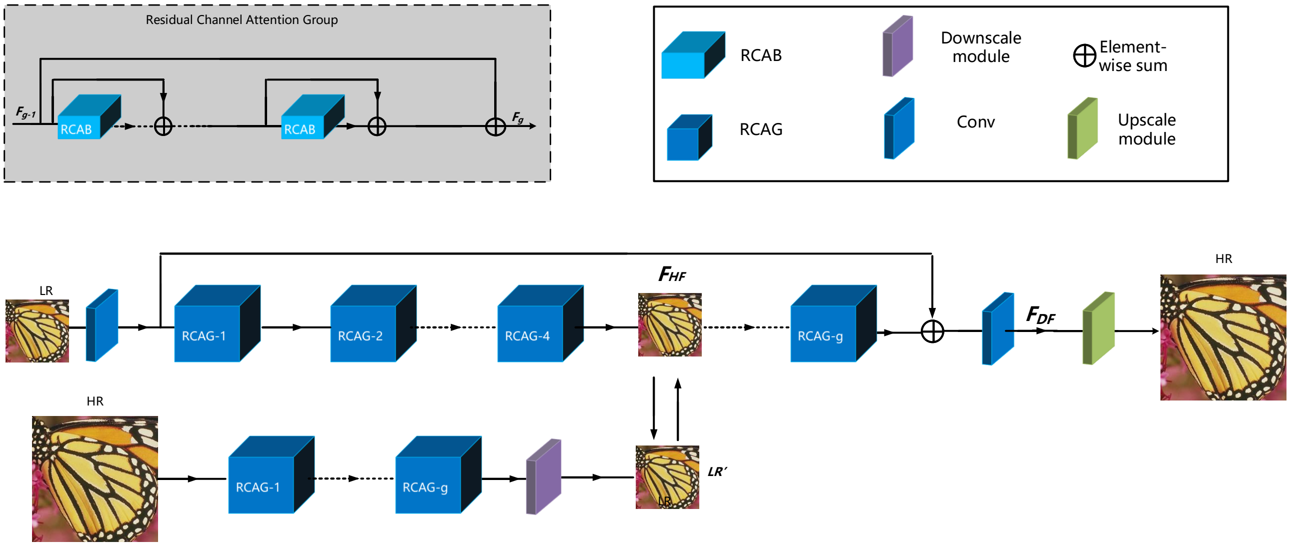

Figure 1. Framework of the proposed two-stage network (TSN). RCAG consists of several residual channel attention blocks (RCAB) [33]. The first stage is learning how to transform LR images into LR images with high-frequency information. The second stage is learning how to transform LR images with high-frequency information into HR images.

struction results. Dai et al. proposed non-locally enhanced residual group (NLRG) [6] to capture spatial contextual information so that hugely improves the performance of the model.

3. Two-Stage Network

3.1. The First Stage

The purpose of the first stage is learning how to transform LR images into LR images with high-frequency information. As shown in Figure 1, our two-stage network (TSN) mainly consists of three parts: shallow feature extractor, residual channel attention group (RCAG) based deep feature extraction, and reconstruction layer. Give I_{LR} and I_{SR} as the input and output of our TSN. We denote I_{HR} as ground truth. We firstly use several residual channel attention groups (RCAGs)[33] and a downscale module as the high-frequency extracting network to extract high-frequency information in I_{HR}. The I_{HR} fed in this downscale network to generate low-resolution images $I_{LR'}$ with high-frequency information.

$$I_{LR'} = H_{down}(I_{HR}) \tag{1}$$

where H_{down} denotes the network consisted of several residual channel attention groups (RCAGs)[33] and a downscale module. To the downscale module, we use a ”max pooling” layer to operate it. Following the [25, 23], we apply one convolution layer to capture the shallow feature F_0 from the LR input

$$F_0 = H_{SF}(I_{LR}) \tag{2}$$

where H_{SF} represents the convolution operation. Then the shallow feature F_0 fed in RCAG , which thus obtains the deep feature as

$$F_{HF} = H_{RCAG4}(F_0) \tag{3}$$

where H_{RCAG4} stands for the first four RCAGs, which consists of several residual channel attention block[33]. It should be noted that F_{HF} with high-frequency information for further transforming high-resolution images.

3.2. The Second Stage

The sake of the second stage is learning how to transform LR images with high-frequency information into HR images. The high-frequency feature F_{HF} produced by the first stage is further via

$$F_{DF} = H_{RCAG8}(F_{HF}) \tag{4}$$

where H_{RCAG8} stands for the last four RCAGs, which consists of several residual channel attention block[33]. Then the extracted deep feature F_{HF} is upsampled through the upscale module via

$$F_{\uparrow} = H_{\uparrow}(F_{DF}) \tag{5}$$

where $H_{\uparrow}$ and $F_{\uparrow}$ are a upsample layer and upsampled feature respectively. In the previous works, there are several choices to perform an upscale part, such as transposed convolution [8], ESPCN [29]. Embedding upscaling feature in the last few layers achieve a good trade-off between performance and computational burden, thus is preferable in recent SR models [8, 6, 25]. Then upscaled feature is through one convolution layer

$$I_{SR} = H_R(F_{\uparrow}) = H_{TSN}(I_{LR}) \tag{6}$$

where H_R, $H_{\uparrow}$, and H_{TSN} are the reconstruction layer, upsample layer, and the function of TSN, respectively.

3.3. Two-stage Learning Loss

Then TSN will be optimized with a loss function. Some loss functions have been widely used, such as L2, L1, perceptual losses. In order to verify the effectiveness of our TSN, we adopt the L1 loss functions followed by previous works. Our design principle is to make the low-resolution images transform to images with high-frequency information, then transform to high-resolution images. Therefore, we transform F_{HF} to RGB images I_{HF} to compute loss for high-frequency feature learning. Given a training set with N low-resolution images and high-resolution images denoted by $\{I_{LR}, I_{HR}\}^N$, the purpose of the TSN is to optimize the loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \lambda||I_{HR} - I_{SR}||_1 + ||I_{HF} - I_{LR'}||_1 \tag{7}$$

where θ represents the parameter set of STN. We choose Adam algorithm to optimize the loss function.

3.4. Implementations

We set the RCAG number as $G = 8$ and λ as 5. In each RCAG, we set 15 residual channel attention blocks. In addition to the shallow extract layer and upscale layer, we set the number of the filter as *C* = 64. For the upscale layer, we follow the works in [35, 33] and apply sub-pixel convolution [29] to upscale and reconstruct the deep feature, followed by a 1 x 1 convolution with three filters to output RGB images.

4. Experiments

4.1. Setup

Following [25, 35], we trained 800 training images in the DIV2K data set. In order to verify the effectiveness of our network, we choose 5 benchmark data sets: Set5, Set14, BSD100, Urban100, and Manga109. For the degraded model, we use the Matlab resizing function with the bicubic operation. For metrics, we use PSNR and SSIM to evaluate SR results.

For training, low-resolution images are enhanced by horizontally flipping and TSNdomly rotating $90°, 180°, 270°$. For each min-batch, we set 16 low-resolution image blocks with a resolution of 48×48 as input. We use the ADAM algorithm to optimize the model with $beta_1 = 0.9$, $beta_2 = 0.99$ and $epsilon = 10^{-8}$, and initialize the learning rate to 10^{-4}, and then reduced by half every 200 cycles. We use the Pytorch framework to train the proposed TSN on Nvidia 1080Ti GPU.

4.2. Ablation Study

To show the effectiveness of the two-stage method, we train and test TSN with its variants on the Set5 dataset. Table 1 shows the specific performance.

R_a denotes the single forward network without a two-stage strategy, which only obtains 32.23 PSNR. With the two-stage strategy, we can find that R_b obtain 32.45 PSNR, which indicates our proposed two-stage method can improve the performance and demonstrates our hypothesis. To the R_b, we only use one convolution layer and downscale module to transform HR images to LR images with high-frequency information. But when we introduce RCAG to consist the deeper network, it extensively improves the performance, which suggests RCAG make an import rule to learn how to transform HR into LR space. To R_c, R_d, and R_e, we argue what is the value of λ should properly be. It should be found that R_d achieves the best result between R_c and R_e. When λ is larger, our TSN will tend to learn the upscaling procedure. But when λ is too lager, the procedure obtained LR with high-frequency information will be hard to learn. Therefore, we use $\lambda = 5$ as our best parameter.

4.3. Comparisons With the State-of-the-Art Methods

As illustrated in Table 2, TSN was compared with more than 7 SR methods, including Bicubic, SRCNN[7], FSRCNN [8], VDSR [19], LapSRN[21], MSRN[23], and SeaNet[10]. The above methods contain conventional models and CNN-based methods. The methods based on CNN are further divided into methods with images prior and without image prior.

Table 2 demonstrates the quantitative comparison between TSN and other super-resolution methods. Obviously, our TSN is superior among these methods. As shown in Table 2, we demonstrate performance comparisons with several advanced CNN-based super-resolution models. The structure of these CNN-based models is carefully designed and achieves satisfactory results at the time. Nevertheless, they ignore how to extract more high-frequency information. As a result, the reconstructed super-resolution images exist some blurry texture. Different from these models, we designed TSN for image super-resolution reconstruction to learn high-frequency information in two-stage. Using the two-stage method, TSN obtained the best results in all test data sets.

In Figure 2, we demonstrate the visual comparisons on ×4 scale factor, respectively. It is clear that most SR methods cannot reconstruct texture. Nevertheless, with the image categorical-prior assistance, our TSN can reconstruct clear super-resolution images.

Table 1. Effects of different modules. We report the PSNR on Set5 datasets in the 200 epoch.

Method	R_a	R_b	R_c	R_d	R_e
Two-stage		✓	✓	✓	✓
RCAG			✓	✓	✓
λ=1		✓	✓		
λ=5				✓	
λ=10					✓
PSNR	32.23	32.45	33.16	33.23	33.21

Table 2. Quantitative comparisons of the state-of-the-art methods.

Method	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	2	33.66/.9299	30.24/.8688	29.57/.8434	26.88/.8403	30.80/.9339
SRCNN [7]	2	36.66/.9542	32.45/.9067	31.36/.8879	29.50/.8946	35.60/.9663
FSRCNN [8]	2	37.05/.9560	32.66/.9090	31.53/.8920	29.88/.9020	36.67/.9710
VDSR [19]	2	37.53/.9590	33.05/.9130	31.90/.8960	30.77/.9140	37.22/.9750
LapSRN [21]	2	37.52/.9591	33.08/.9130	31.08/.8950	30.41/.9101	37.27/.9740
MSRN [23]	2	38.07/.9608	33.68/.9184	33.68/.9184	32.32/.9304	38.64/.9771
SeaNet [10]	2	38.08/.9609	33.75/.9190	32.27/.9008	32.50/.9318	38.76/.9774
RAN	2	**38.19/.9611**	**33.82/.9191**	**32.27/.9008**	**34.52/.9357**	**38.89/.9821**
Bicubic	3	39.32/.9792	27.55/.7742	27.21/.7385	24.46/.7349	26.95/.8556
SRCNN [7]	3	32.75/.9090	29.30/.8215	28.41/.7863	26.24/.7989	30.48/.9117
FSRCNN [8]	3	33.18/.9140	29.37/.8240	28.53/.7910	26.43/.8080	31.10/.9210
VDSR [19]	3	33.67/.9210	29.78/.8320	28.83/.7990	27.14/.8290	32.01/.9340
LapSRN [21]	3	33.82/.9227	29.87/.8320	28.82/.7980	27.07/.8280	32.21/.9350
MSRN [23]	3	34.48/.9276	34.48/.9276	29.13/.8061	29.13/.8061	33.56/0.9451
SeaNet [10]	3	34.55/.9282	30.42/.8444	29.17/.8071	28.50/.8594	33.73/.9463
RAN	3	**34.61/.9392**	**30.58/.8542**	**29.21/.8174**	**28.52/.8601**	**33.96/.9521**
Bicubic	4	28.42/.8104	26.00/.7027	25.96/.6675	23.14/.6577	24.89/.7866
SRCNN [7]	4	30.48/.8628	27.50/.7513	26.90/.7101	24.52/.7221	27.58/.8555
FSRCNN [8]	4	30.72/.8660	27.61/.7550	26.98/.7150	24.62/.7280	27.90/.8610
VDSR [19]	4	31.35/.8830	28.02/.7680	27.29/.0726	25.18/.7540	28.83/.8870
LapSRN [21]	4	31.54/.8850	28.19/.7720	27.32/.7270	25.21/.7560	29.09/.8900
MSRN [23]	4	32.25/.8958	28.63/.7833	27.61/.7377	27.61/.7377	27.61/.7377
SeaNet [10]	4	32.33/.8970	28.72/.7855	27.65/.7388	26.32/.7942	30.74/.9129
RAN	4	**32.42/.8984**	**28.75/.7862**	**27.67/.7410**	**26.37/.7961**	**31.05/.9241**

Table 3. Computational and parameter comparison (2X) Set5.

	EDSR	MemNet	MSRN	SeaNet	TSN
Para.	43M	677k	6M	8M	9.6M
PSNR	38.11	37.78	38.07	38.08	38.19

4.4. Model Size Analyses

Table 3 shows the model size and performance of the current CNN SR model. Among these methods, MemNet and NLRG contain much fewer parameters, which reduces performance. TSN not only has fewer parameters than EDSR, MSRN, and SeaNet, but also has better performance, which means that TSN can make a big performance trade-off between model complexity and performance.

5. Conclusion

In this work, we propose a two-stage network (TSN). The one stage is learning how to transform LR images into LR images with high-frequency information. The other stage is learning how to transform LR images with high-frequency information into HR images. Meanwhile, we also propose a two-stage learning loss, which will guide our network jointly learning how to transform LR images into LR images with high-frequency information and transform LR images with high-frequency information into HR images. Extensive experiments show our TSN can reconstruct the clear super-resolution images with fewer parameters.

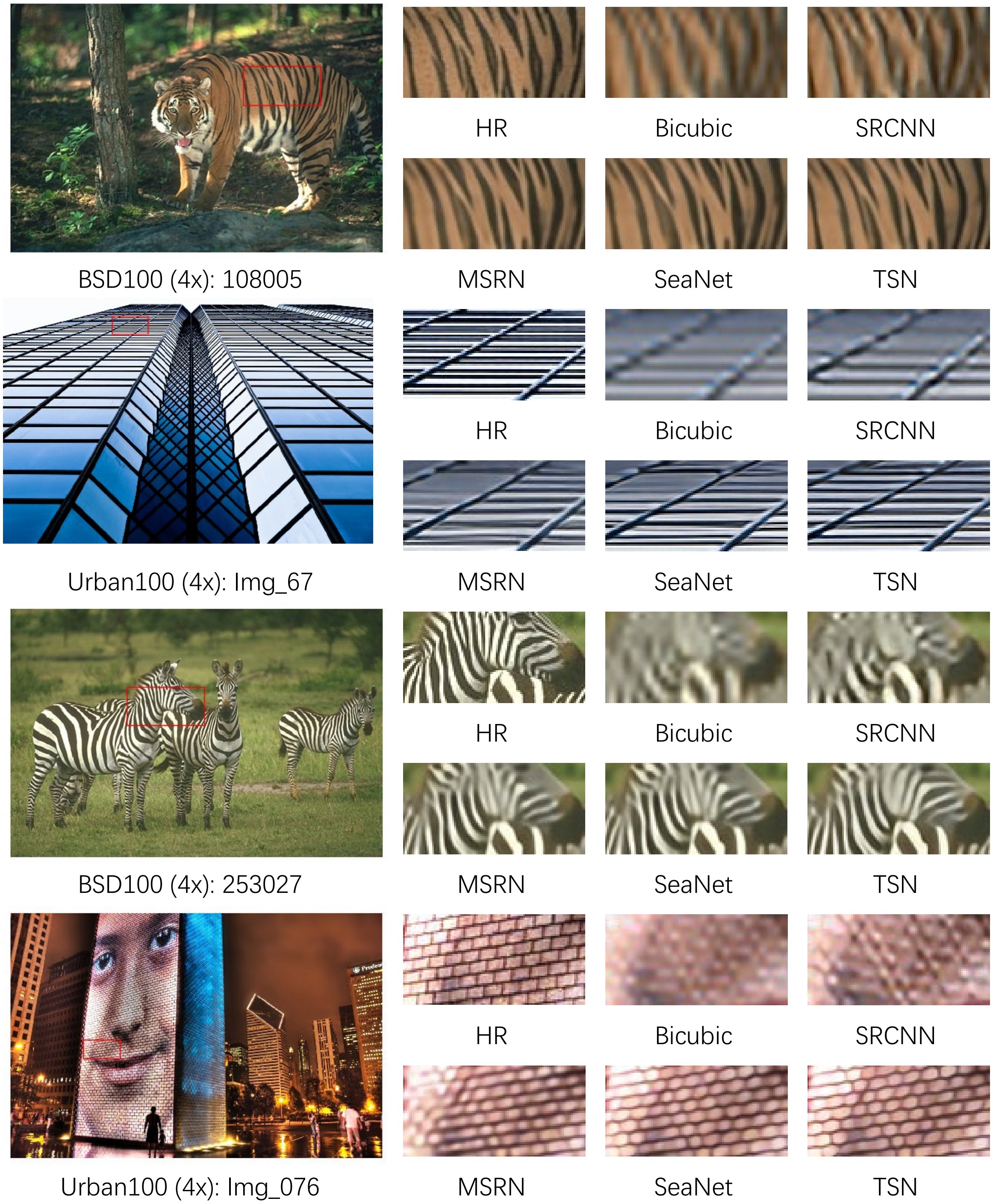

Figure 2. Visual comparison for 4x SR with BI model.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018. 1

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 1

[3] David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. In *Proceedings*

of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 2, pages II–II. IEEE, 2001. 2

[4] Ayan Chakrabarti, AN Rajagopalan, and Rama Chellappa. Super-resolution of face images using kernel pca-based prior. *IEEE Transactions on Multimedia*, 9(4):888–892, 2007. 1, 2

[5] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 1, 2

[6] 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. 3

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2, 4, 5

[8] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 2, 3, 4, 5

[9] X. Du, S. Jiang, Y. Si, L. Xu, and C. Liu. Mixed high-order non-local attention network for single image super-resolution. *IEEE Access*, 9:49514–49521, 2021. 1

[10] Faming Fang, Juncheng Li, and Tieyong Zeng. Soft-edge assisted network for single image super-resolution. *IEEE Transactions on Image Processing*, 29:4656–4668, 2020. 4, 5

[11] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 1, 2

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2

[13] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009. 1, 2

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 1, 2

[16] 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 International Conference on Computer Vision*, pages 603–612, 2019. 2

[17] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 2

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1

[19] 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. 2, 4, 5

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[21] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 2, 4, 5

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[23] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018. 2, 3, 4, 5

[24] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9167–9176, 2019. 2

[25] 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. 1, 2, 3, 4

[26] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 1

[27] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 1

[28] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018. 1

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan

Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 1, 3, 4

[30] 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. 2

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[32] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 2

[33] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 1, 2, 3, 4

[34] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2

[35] 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. 4