RESIDUAL MAGNIFIER: A DENSE INFORMATION FLOW NETWORK FOR SUPER RESOLUTION

Zhan Shu[1], *Mengcheng Cheng*[1], *Biao Yang*[1], *Zhuo Su*[1,*], *Xiangjian He*[2,3]

[1] School of Data and Computer Science, National Engineering Research Center of Digital Life, Sun Yat-sen University, Guangzhou, China
[2]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China
[3] School of Computing and Communications, University of Technology Sydney
[*] Corresponding Author: suzhuo3@mail.sysu.edu.cn

ABSTRACT

Recently, deep learning methods have been successfully applied to single image super-resolution tasks. However, some networks with extreme depth failed to achieve better performance because of the insufficient utilization of the local residual information extracted at each stage. To solve the above question, we propose a Dense Information Flow Network (DIF-Net), which can fully extract and utilize the local residual information at each stage to accomplish a better reconstruction. Specifically, we present a Two-stage Residual Extraction Block (TREB) to extract the shallow and deep local residual information at each stage. The dense connection mechanism is introduced throughout the model and within TREBs to dramatically increase the information flow. Meanwhile this mechanism prevents the shallow features extracted earlier from being diluted. Finally, we propose a lightweight subnet (residual enhancer) to efficiently recycle the overflow residual information from the backbone net for detail enhancement of the residual image. Experimental results demonstrate that the proposed method performs favorably against the state-of-the-art methods with relatively-less parameters. Code is available at https://github.com/suzhuoi/DIF-Net.

Index Terms— Single image super-resolution, Enhanced residual information, Dense connection

1. INTRODUCTION

Single image super-resolution (SISR) is a classical problem in low-level computer vision, which reconstructs a high-resolution (HR) image from a low-resolution (LR) image without altering the semantics of the image. The same LR image can be obtained by downsampling from an infinite number of HR images. Hence, the SR problem is inherently an ill-posed problem. In order to solve the problem, numerous SISR methods have been presented, including interpolation-based methods and reconstruction-based models. But they normally suffer dramatically degeneration in restoration performance with some large upscaling factors.

We first review CNN-based SR methods and then introduce the main contributions of the proposed approach.

1.1. Related Work

CNN-based SR methods have been witnessed noteworthy progress recently [1, 2, 3]. Dong *et al.* [4] first exploited a three-layer convolutional neural network (SRCNN), to jointly optimize the feature extraction, non-linear mapping and image reconstruction stages into an end-to-end manner. Aimed at the disadvantage of too much computation, a modified method (FSRCNN) is constructed by Dong *et al.* [5], which adopts the original LR image as input without interpolation. These improvements provide FSRCNN [5] better performance but lower computational cost than SRCNN [4]. The work in [6] then presented an efficient sub-pixel convolutional neural network (ESPCN), which replaces the upsampling operation with an efficient sub-pixel convolution.

Nevertheless, constrained by the challenge of training, many deep models cannot achieve ideal results. Kim *et al.* [7] increased the network depth to 20 layers by migrating ResNet into SISR. Residual information is sparse and more accessible to learn, which helps to speed up VDSR's [8] convergence during training. Leding *et al.* [9] presented SRGAN to further increase the depth to 30 layers. Removing some redundant modules from the residual network, Lim *et al.* [10] were able to train their model (EDSR) with 160 layers. Tai *et al.* [11] proposed MemNet with long-term memory. They mimicked the workings of the cerebral cortex and formulated the skip connection mechanism to bridge the long-term dependencies. Haris *et al.* [12] illustrated error feedback mechanism to characterize or constrain the features in early layers. Recently, Zheng *et al.* [13] constructed a simple network (IDN) with the key component DBlock consists of an enhancement unit and a compression unit. Their proposed method achieved real-time speed while still maintaining good reconstruction accuracy.

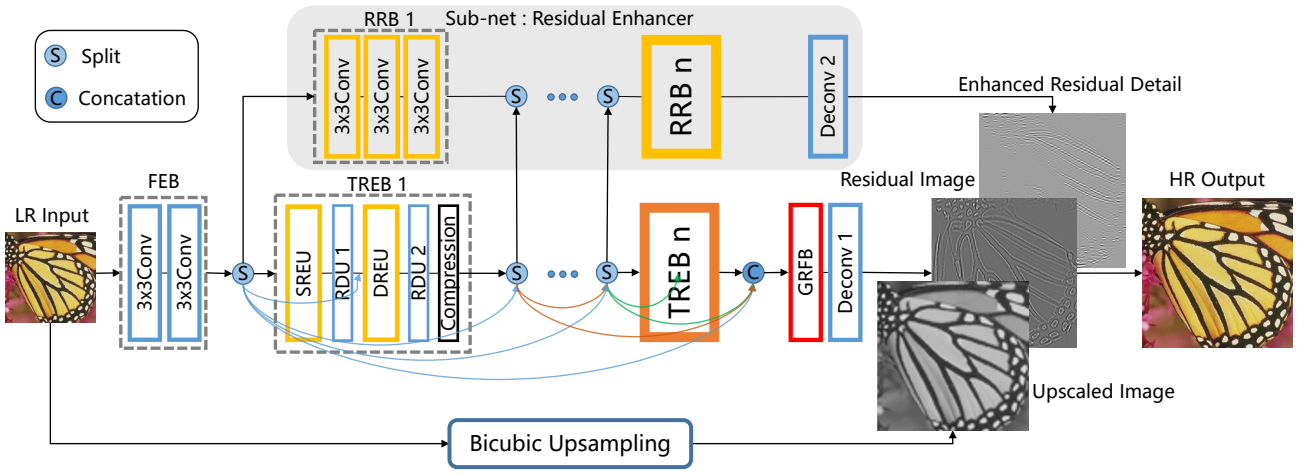

Fig. 1. Architecture of the proposed network. Backbone net contains two 3×3 convolutional layers for extracting residual information from LR images. The following TREBs extract information progressively. All information will flow to GRFB for a global fusion. The light gray component is the residual enhancer for recycling the overflow information.

Although achieving prominent performance, most of the deep networks still have some drawbacks. Most networks lack sufficient information flow, which may lead to a lower information utilization rate[1]. Taking [13, 8] as examples, information extracted by each block only flow to the next block, and there are no direct connections between nonconsecutive blocks. The straight-forward structure dilutes previously extracted information as continuous convolution operations. Though using skip connection mechanism to enhance the contacts between different blocks, the structure of MemNet [11] is still sparse. And these models only pay attention to the depth of networks rather than the compactness, which results in the insufficient excavation of hierarchical features and restriction of information utilization rate.

1.2. Contributions

To address these drawbacks, we propose a novel model DIF-NET (Fig. 1) with dense connections both locally and globally to increase the information flow and extract hierarchical residual information in each phase as much as possible. The main contributions of the proposed method are three-fold.

- **Two-stage Residual Extraction Block**. TREB is the core module for DIF-NET. We divide the procedure of extracting local residual information into two stages: firstly shallow information and then the deep one. Different from IDN [13], we utilize dense connection mechanism to ensure sufficient excavation of local residual information. Two Information Denoising Units are placed behind shallow/deep residual extraction unit separately. Compression unit fuses the extracted shallow and deep information as well as reduces the dimension of feature maps. The ingenious two-stage structure prevents the dilution of extracted shallow information while still obtaining deep information.

- **Global Residual Fusion Mechanism**. Different from skip connection mechanism presented in MemNet [11], we utilize dense connections to reinforce the associations between TREBs. As illustrated in Fig. 2, the compact connections between blocks increase the information flow remarkably, leading to the growth of the information utilization rate.

- **Residual Enhancer**. Dense connections do not guarantee the full utilization of information, a huge amount of information flow will lead to the overflow of effective information. We innovatively design a residual enhancer consisting of Residual Recycle Blocks (RRBs) to recycle the residual information spilled from each TREB. A residual image with clearer texture will be obtained after enhancement of subnet.

2. PROPOSED METHODS

In this section, we first describe the proposed model architecture and then suggest the MREB, the residual enhancer and global residual fusion mechanism.

2.1. Network Structure

As shown in Fig. 1, our model consists of a backbone net and a subnet (residual enhancer). In the backbone net, information passes through Feature Extraction Block (FEB), several TREBs, Global Residual Fusion Block (GRFB) and finally a deconvolution block successively. The split operations are placed between TREBs to implement the dense connection

[1] The utilization rate can be defined as the ratio of data to parameters, and less parameters means higher utilization rate once data is fixed.

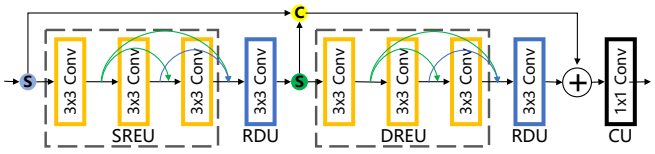

Fig. 2. Architecture of the proposed TREB.

mechanism. Subnet contains several RRBs and a deconvolution block. I_L and I_H represent the input and output of our model separately. FEB, which are firstly utilized to extract the features from the LR image, can be formulated as:

$$R_0 = F(I_L) = F_2(F_1(I_L)), \tag{1}$$

where R_0 denotes the output of FEB, $F_1(\cdot)$ and $F_2(\cdot)$ indicate the first two convolution operations, $F(\cdot)$ represents FEB function. It is worthy of note that R_0 only contains high frequency information that related to details in LR image, i.e. residual information. Theoretically FEB can be regarded as a high-pass filter which selectively suppresses background-related low frequency information and extracts high frequency information. R_0 is then utilized to extract hierarchical information by series of TREBs. There exists n blocks of TREBs, the k-th output R_k can be expressed as:

$$R_k = M_k(S_{k,out}), \tag{2}$$

and $M_k(\cdot)$ denotes the k-th TREB function. The input of a split operation is the same as all its outputs, which can be described as:

$$S_{k,in} = S_{k,out} = R_{k-1} \oplus S_{k-1,out} \oplus ... \oplus S_{1,out}. \tag{3}$$

R_{k-1} indicates the output of the (k-1)-th TREB, $S_{k,in}$ and $S_{k,out}$ represent the input of the k-th split operation as well as all its outputs respectively. $\oplus$ denotes the sum operation. Hierarchical information will be extracted sufficiently at each stage through TREBs. After extracting features deep enough with a set of TREBs, we further conduct GRFB, which fuses all the hierarchical information from every stage. The output R_G of GRFB can be obtained by:

$$R_G = G(C([R_n; S_{n,out}; S_{n-1,out}; ...; S_{1,out}])), \tag{4}$$

where $G(\cdot)$ denotes the GRFB function. Since we have obtained sufficient features that stock residual information from each stage, a deconvolution block is placed at the end of the backbone net to map the features to HR space:

$$I_H^r = D_1(R_G). \tag{5}$$

Here, I_H^r denotes the output of the backbone net (i.e. the residual image), $D_1(\cdot)$ indicates deconvolution function. Dense connections are used to increase the information flow of our model, meanwhile residual enhancer recycle the spilled information from the backbone net. The output B_k of the k-th RRB can be expressed as:

$$B_k = R^k(C([B_{k-1}; S_{k,out}])), \tag{6}$$

where B_{k-1} demonstrates the k-th RRB function. Finally a deconvolution block is utilized to map the output of the last RRB to HR space, this procedure can be formulated as:

$$I_H^f = D_2(B_n), \tag{7}$$

specifically I_H^f denotes the output of the residual enhancer and we regard it as enhanced residual detail. Residual enhancer will be specifically described in Section 2.3. Hence the output of the proposed model can be summarized as:

$$I_H = \mathcal{F}(I_L) = I_B + I_H^r + I_H^f, \tag{8}$$

and I_B represents the image after bicubic upsampling processing, $\mathcal{F}(\cdot)$ is the construction function of our model.

2.2. Two-stage Residual Extraction Block

As illustrated in Fig. 2, TREB can be roughly divided into five units: shallow residual extraction unit (SREU) and deep residual extraction unit (DREU) followed by a residual denoising unit (RDU) respectively, a 1x1 convolution layer named compression unit (CU) is utilized to computation reduction. The output U_1^k of k-th SREU can be expressed as:

$$U_1^k = H_s^k(S_{k-1,out}) = H_{s3}^k(H_{s2}^k(H_{s1}^k(S_{k-1,out}))), \tag{9}$$

where $H_s^k(\cdot)$ represents the k-th SREU function. $H_{s1\sim3}^k(\cdot)$ indicates the first three convolution operation of the k-th TREB. Dense connection is also implemented to allow information flows across layers. We further have U_2^k:

$$U_2^k = H_{r1}^k(U_1^k), \tag{10}$$

and $H_{r1}^k(\cdot)$ is the RDU function. RDUs stabilize the training by removing noise from residual information. U_2^k is then divided into two parts by slice operation:

$$U_{2,1}^k = \zeta_{1-q}(U_2^k),\ U_{2,2}^k = \zeta_q(U_2^k). \tag{11}$$

Specifically we know that the dimension of $U_{2,2}^k$ is q ($q \in (0, 1)$) times that of U_2^k. $U_{2,2}^k$ is then concatenated with $S_{k,out}$ in channel dimension, we have:

$$R_c^k = C[U_{2,2}^k, S_{k-1,out}], \tag{12}$$

letting C denotes the concatenation operation. Therefore partially local shallow residual information is reserved to prevent from being diluted. The rest information is further used to extract local deep residual information:

$$\begin{aligned} U_3^k &= H_d^k(U_{2,1}^k) = H_{d3}^k(H_{d2}^k(H_{d1}^k(U_{2,1}^k))), \\ U_4^k &= H_{r2}^k(U_3^k), \end{aligned} \tag{13}$$

where $H_d^k(\cdot)$ denotes the DREU function and $H_{r2}^k(\cdot)$ denotes the RDU function. U_3^k and U_4^k represents the outputs separately. Finally the CU fuses the extracted local shallow residual information and deep residual information:

$$R_k = H_c^k((U_4^k \oplus R_c^k)), \tag{14}$$

and $H_c^k(\cdot)$ represents the k-th CU function.

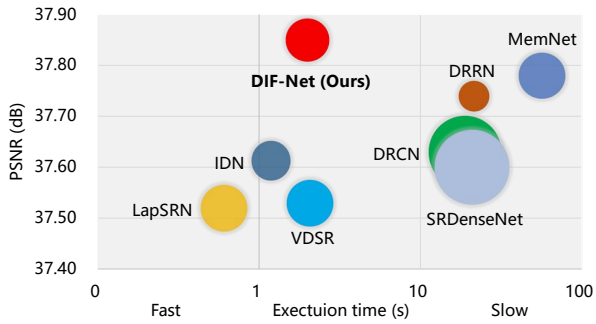

Fig. 3. Trade-off between average performance vs. speed and parameters on the Set5 dataset with 2 times. The size of the circle represents the number of parameters.

2.3. Residual Enhancer

Most methods improve reconstruction performance by deepening the depth of the networks, which may lead to a decline in information utilization. Constrained by the number of paraments, effective information will inevitably spill when flowing through blocks. The proposed lightweight residual enhancer is utilized to recycle the spilled residual information to extract more subtle details further. These enhanced details are finally mapped to the HR space at the end of the subnet.

As illustrated in Fig. 1, subnet is composed of a set of RRBs and a deconvolution block. Each RRB consists of three layers of 3×3 convolution layers followed by rectified linear units (ReLU) respectively. Specifically, for the output B_k of the k-th RRB, we have:

$$B_k = R^k(C[B_{k-1}, S_{k,out}]), \tag{15}$$

where B_{k-1} is the output of the previous RRB. $R^k(\cdot)$ indicates the k-th RRB function. The output I_H^f of the subnet is called enhanced residual detail, as illustrated in Eq. 7.

2.4. Global Residual Fusion Mechanism

As described in Section 2.2, the straight-forward information flow mode dilutes the shallow residual features extracted formerly as continuous convolution operations. Moreover, the deficiency of information flow in a way restricts the expressive ability of the model, which likewise leads to a low information utilization rate. We introduce dense connections, named as global residual fusion mechanism, to connect the whole network. Residual information extracted by each TREB will flow into all the following TREBs and subnet, which increase network flow remarkably. Meanwhile, all residual information will gather at the end of the backbone net through concatenation operation. Eventually we obtain the final residual information using a 3×3 GRFB.

3. EXPERIMENTAL RESULTS

In this section, we first introduce training details, then analysis our models, and compre it with state-of-the-arts.

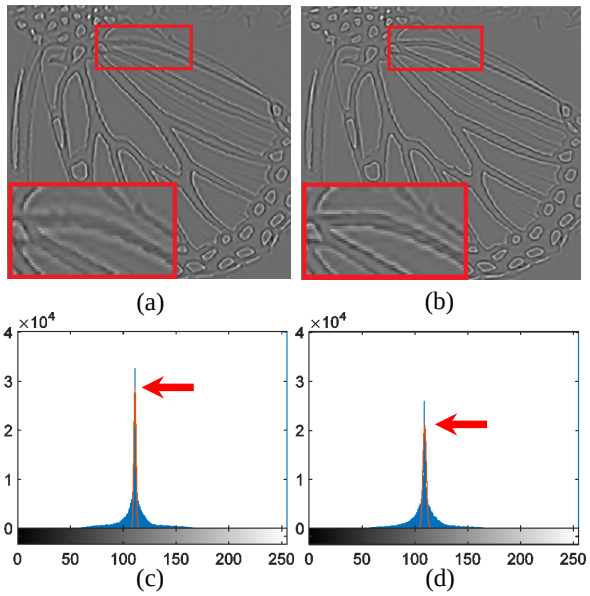

Fig. 4. The residual images comparison and corresponding data distribution histogram analysis using the "butterfly" image from Set5 dataset.

3.1. Implementation and training details

We use a high-quality (2K resolution) dataset DIV2K dataset as the training data. DIV2K consists of 800 training images, 100 validation images, and 100 test images. We train all of our models with 800 training images and use 5 validation images in the training process. For testing, we use four standard benchmark datasets: Set5, Set14, BSD100 and Urban100. The SR results are evaluated with PSNR and SSIM on Y channel (i.e., luminance) of transformed YCbCr space.

We set the parameters of mini-batch size and weight decay to 64 and 1e-4 respectively. Training a DIF-Net roughly takes a day with a NVIDIA GeForce GTX 1080 GPU for 2x model.

3.2. Model Analysis

As illustrated in Fig. 3, due to the concise structure of proposed DIF-Net and parameters sharing strategy within and between modules, it is faster than several CNN-based SR methods and maintains better reconstruction accuracy. Here, our DIF-Net model outperforms all state-of-the-art models, and has less parameters than other models.

As shown in Fig. 4, (a) and (b) represent the residual images without and with enhancement by the residual enhancer respectively. A clearer texture in (b) can be told from the magnified details in (a) and (b). The blurred and jagged edge in (a) becomes observably clearer and smoother as shown in (b) with reinforcement of residual enhancer. The corresponding data distribution histograms are shown in (c) and (d) separately, which both subject to the Gaussian distribution. The larger variance value (indicated by red arrows) in (d) repre-

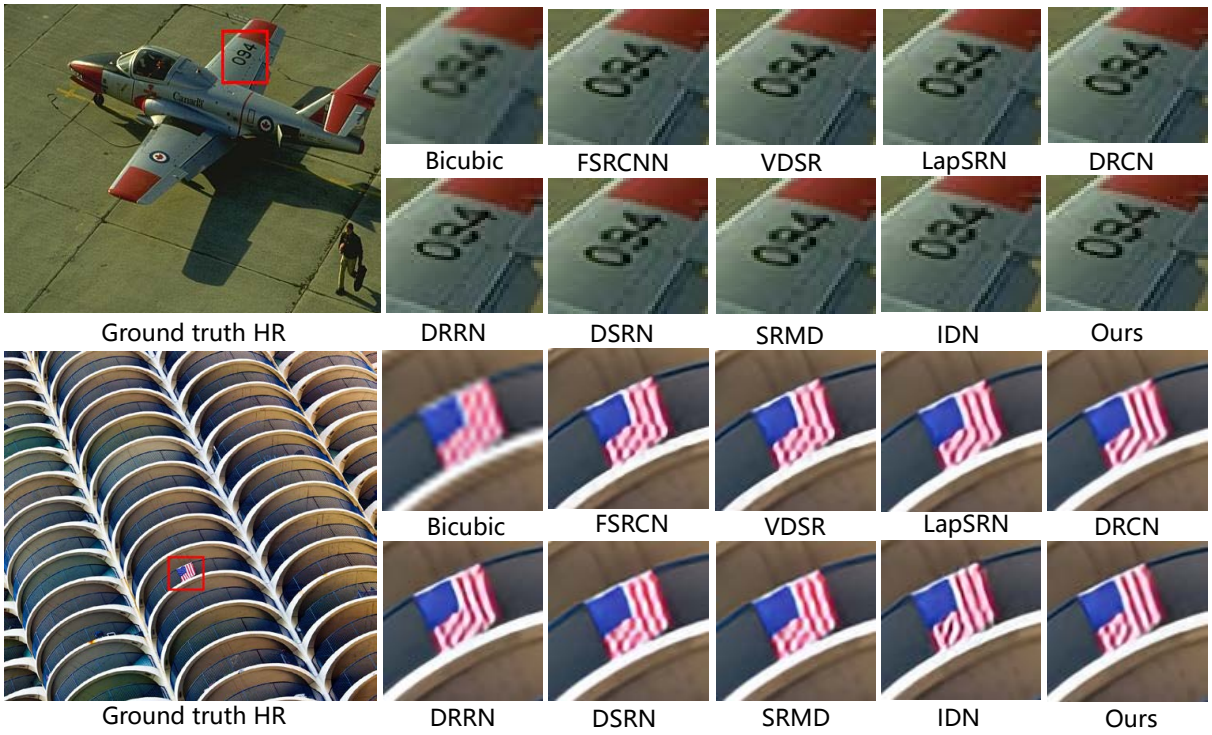

Fig. 6. Visual qualitative comparison of our models with other works. The first sample is 2x on BSD100 dataset and the second is 4x on Urban100 dataset.

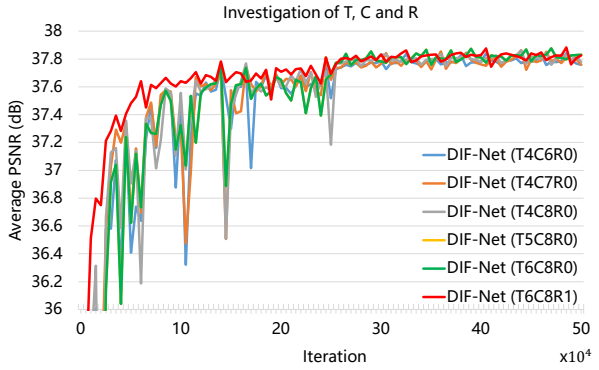

Fig. 5. Convergence analysis of DIF-Net with different values of T, C, and R. The curves for each combination are based on the PSNR on Set5 with scaling factor x2.

sents a higher image contrast in (b), conforming to the visual cognition that (b) has a clearer texture.

As shown in Fig. 5, we investigate the basic network parameters: the number of TREB (T), the number of Conv layers per TREB (C), and whether use residual enhancer (R), larger T or C and R=1 would lead to higher performance.

3.3. Comparison with State-of-the-art Methods

We compare the proposed method with other SR methods, including bicubic, FSRCNN [5], VDSR [8], DRCN [7], LapSRN [14], DRRN [15], DSRN [16], SRMD [17] and IDN [13]. Table 1 and Fig. 6 shows PSNR and SSIM values on four benchmark datasets. The proposed performs favorably against state-of-the-art results on most datasets.

4. CONCLUSION

This paper has presented a novel network for single image super-resolution, which can fully use the hierarchical information. We have utilized the dense connections to increase the information flow as well as prevent the dilution of shallow features. The proposed residual enhancer with relatively-less parameters can efficiently recycle the spilled residual information. Experimental results have demonstrated that the proposed method performs favorably against the state-of-the-art methods on four benchmark datasets, especially in terms of PSNR, SSIM and time performance.

5. ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation of China (61872394), Guangxi Innovation Driven Development Special Fund Project (AA18118039), and Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University, MJUKF201701). Zhuo Su is the corresponding author (suzhuo3@mail.sysu.edu.cn).

Table 1. Quantitative evaluation of state-of-the-art SR algorithms: average PSNR/SSIM for scale 2x, 3x and 4x. Red color indicates the best and blue color indicates the second best performance.

DataSet	Scale	Bicubic	FSRCNN [4]	VDSR [8]	LapSRN [14]	DRCN [7]	DRRN [15]	DSRN [16]	SRMD [17]	IDN [13]	DIF-Net (Ours)
Set5	x2	33.66/0.930	36.99/0.955	37.53/0.958	37.52/0.959	37.63/0.959	37.74/0.959	37.66/0.959	37.53/0.959	37.75/0.959	37.84/0.960
	x3	30.39/0.868	33.15/0.913	33.66/0.921	33.81/0.922	33.82/0.922	33.93/0.923	33.88/0.922	33.86/0.923	33.92/0.923	34.02/0.924
	x4	28.42/0.810	30.71/0.865	31.35/0.882	31.54/0.885	31.53/0.884	31.68/0.888	31.40/0.883	31.59/0.887	31.44/0.884	31.64/0.887
Set14	x2	30.24/0.869	32.73/0.909	32.97/0.913	33.08/0.913	32.98/0.913	33.23/0.913	33.15/0.913	33.12/0.914	33.10/0.913	33.35/0.915
	x3	27.55/0.774	29.53/0.826	29.77/0.831	29.79/0.833	29.76/0.833	29.94/0.834	30.26/0.837	29.84/0.833	29.87/0.833	29.94/0.834
	x4	26.00/0.703	27.70/0.756	28.03/0.770	28.19/0.772	28.04/0.770	28.21/0.772	28.07/0.770	28.15/0.772	28.06/0.769	28.15/0.772
BSD100	x2	29.56/0.843	31.51/0.891	31.90/0.896	31.80/0.895	31.85/0.894	32.05/0.897	32.10/0.897	31.90/0.896	32.02/0.898	32.08/0.899
	x3	27.21/0.739	28.52/0.790	28.82/0.798	28.82/0.798	28.80/0.797	28.91/0.799	28.81/0.797	28.87/0.799	28.86/0.798	28.91/0.800
	x4	25.96/0.668	26.97/0.714	27.29/0.726	27.32/0.728	27.24/0.724	27.38/0.728	27.25/0.724	27.34/0.728	27.27/0.725	27.34/0.728
Urban100	x2	26.88/0.840	29.87/0.901	30.77/0.914	30.41/0.910	30.76/0.913	31.23/0.919	30.97/0.916	30.89/0.916	31.13/0.918	31.33/0.919
	x3	24.46/0.735	26.42/0.807	27.14/0.828	27.07/0.828	27.15/0.828	27.38/0.833	27.16/0.828	27.27/0.833	27.16/0.830	27.43/0.835
	x4	23.14/0.658	24.61/0.727	25.18/0.753	25.21/0.756	25.14/0.752	25.44/0.764	25.08/0.747	25.34/0.761	25.09/0.752	25.39/0.761

6. REFERENCES

[1] R. Liu, X.Wang, X.Fan, H. Li, and Z. Luo, "Deep hybrid residual learning with statistic priors for single image super-resolution," *In ICME*, pp. 1111–1116, 2017.

[2] Z. Su, L. Li, J. Li, and X. Luo, "Maximised self-similarity upsampler," *IET Image Processing*, vol. 11, no. 12, pp. 1229–1237, 2017.

[3] L. Li, Z. Su, X. Shi, E. Huang, and X.Luo, "Mutual-details convolution model for image super-resolution reconstruction," *Journal of Image and Graphics*, vol. 23, no. 4, pp. 572–582, 2018.

[4] C. Dong, C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," *In ECCV*, pp. 184–199, 2014.

[5] C. Dong, C. Loy, K. He, and X. Tang, "Accelerating the super-resolution convolutional neural network," *In ECCV*, pp. 391–407, 2016.

[6] W. Shi, J. Caballero, and F. Huszar, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *In CVPR*, pp. 1874–1883, 2016.

[7] J. Kim, J. Lee, and K. Lee, "Deeply-recursive convolutional network for image super-resolution," *In CVPR*, pp. 1637–1645, 2016.

[8] J. Kim, J. Lee, and K. Lee, "Accurate image super-resolution using very deep convolutional networks," *In CVPR*, pp. 1646–1654, 2016.

[9] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, and A. Tejani, "Photo-realistic single image super-resolution using a generative adversarial network," *In ICCV*, pp. 4681–4690, 2017.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. Lee, "Enhanced deep residual networks for single image super-resolution," *In ICCV*, pp. 1132–1140, 2017.

[11] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," *In ICCV*, pp. 4539–4547, 2017.

[12] G. Shakhnarovich M. Haris and N. Ukita, "Deep back-projection networks for super-resolution," *In CVPR*, pp. 1664–1673, 2018.

[13] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," *In CVPR*, pp. 723–731, 2018.

[14] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *In CVPR*, pp. 5835–5843, 2017.

[15] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," *In CVPR*, pp. 2790–2798, 2017.

[16] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. Huang, "Image super-resolution via dual-state recurrent networks," *In CVPR*, pp. 1654–1663, 2018.

[17] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," *In CVPR*, pp. 3262–3271, 2018.