This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/icassp39728.2021.9414527, ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

LIGHTWEIGHT NON-LOCAL NETWORK FOR IMAGE SUPER-RESOLUTION

Risheng Wang[1,2], *Tao Lei*[1,2*], *Wenzheng Zhou*[3], *Qi Wang*[4], *Hongying Meng*[5], *Asoke K. Nandi*[5]

[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology
Xi'an 710021, P. R. China
[2]Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology
Xi'an 710021, P. R. China
[3]School of Electrical and Control Engineering, Shaanxi University of Science and Technology
Xi'an 710021, P. R. China
[4]School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China
[5]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge,
Middlesex, UB8 3PH, United Kingdom

ABSTRACT

The popular deep convolutional networks used for image super-resolution (SR) reconstruction often increase the network depth and employ attention mechanism to improve image reconstruction effect. However, these networks suffer from two problems. The first is the deeper network easily causes higher computational cost and more GPU memory usage. The second is traditional attention mechanism often misses the spatial information of images leading the loss of image detail information. To address these issues, we propose a lightweight non-local network (LNLN) for image super resolution in this paper. The proposed network makes two contributions. First, we use non-local module instead of normal attention module to obtain larger receptive field and extract more comprehensive feature information, which is helpful for improving image SR reconstruction results. Secondly, we use the depthwise separable convolution (DSC) instead of the vanilla convolution to reconstruct the residual block, which greatly reduces the number of parameters and computational cost. The proposed LNLN and comparative networks are evaluated on five commonly public datasets, and experiments demonstrate that the proposed LNLN is superior to state-of-the-art networks in terms of reconstruction performance, the number of parameters and storage space.

Index Terms— deep learning, image super-resolution (SR), non-local module, depthwise separable convolution (DSC)

1. INTRODUCTION

Image super-resolution (SR) reconstruction is a technology to restore low-resolution image into high-resolution image and improve image visual quality. It has been widely used in remote sensing [1], medicine [2], criminal investigation [3] and other fields. To achieve image SR, ones can improve image sensors and optical instruments, but a high-precision equipment is usually expensive. Therefore, it is not practical to improve image super-resolution by changing hardware equipment. Due to this reason, researchers begin to pay attention to image super-resolution algorithms. With the continuous development of deep learning, convolutional neural networks (CNN) are increasingly used for image super-resolution reconstruction [4]. By designing a CNN model to learn the nonlinear mapping relationship, it is possible to obtain the reconstructed high-resolution images that correspond to a low-resolution input images. Although existing deep learning algorithms can provide good image super-resolution results than traditional algorithms based on model-driven, they still suffer from some shortcomings.

First of all, image SR belongs to ill posed problem, that is, multiple possible high resolution (HR) images can be obtained from one low resolution (LR) image after reconstruction, and there is no unique solution. Therefore, there are many possibilities to map LR image to HR image, and it is difficult to get the optimal solution. To solve the problem, Kim et al. [5] proposed an accurate image super resolution using very deep convolutional network (VDSR) that increases the network depth to 20 layers. By cascading small filters in the deep network structure, the context information of large image area can be effectively utilized to improve image SR accuracy. The deepening of VDSR network is conducive to enhance the representation ability, but it cannot make full use of the shallow feature information. Therefore, Zhang et al. [6] proposed a residual dense network (RDN) that uses various skip connections and series operations between shallow

*(*Corresponding author: Tao Lei)*

and deep layers to combine dense connection layer and local feature fusion to achieve excellent SR effect. In order to further improve the convergence speed and accuracy of SR networks, Yu et al. [7] proposed a wide activation for efficient and accurate image super-resolution (WDSR) by widening the network width and increasing the characteristics of deep flow. Although these studies mentioned above improve image SR effect by designing deeper or wider networks, they suffer from the problems of a huge mountain of parameters and high computational cost requirement.

Secondly, each feature map channel is equally important in the above mentioned SR networks, but in the actual training, different feature map channels have different importance. Therefore, researchers begin to integrate attention modules to SR networks. Li et al. [8] constructed a multi-scale residual network (MSRN) by introducing convolution kernels of different scales into the network to realize image SR reconstruction. Compared with previous CNNs with single-scale convolution kernel, MSRN can fully extract feature information and improve image SR reconstruction effect. Moreover, Zhang et al. [9] proposed residual channel attention network (RCAN), in which a very deep training network is constructed by residual in residual (RIR) structure. Since both the long skip connection and short skip connection in RIR are helpful for preserving low frequency information of images, RCAN can learn more useful information leading to better image SR reconstruction effect. To improve RCAN, Zhang et al. [10] proposed a densenet with deep residual channel attention (DRCA) based on RCAN. Compared with RCAN, the network requires fewer parameters and less computation. However, DRCA only considers the dependence between feature channels but ignores the spatial correlation of feature maps resulting in the loss of image detail information.

Although a lot of studies on image SR have been reported, they still face two challenges. First, the channel attention mechanism cannot capture the spatial relationship of images. Secondly, the number of parameters and computational cost of the network becomes higher to improve feature representation. In this paper, we propose a lightweight non-local network for image super resolution (LNLN), our main contributions include:

(1) We integrate both spatial attention and channel attention to the proposed LNLN, which achieves better SR reconstruction results than normal channel attention networks.

(2) We employ depthwise separable convolution (DSC) to achieve a lightweight network to reduce the number of parameters. The proposed LNLN is superior to popular SR networks since it only requires 8.72MB memory.

2. THE PROPOSED NETWORK

For existing depth network models, the DRCA is one of the most popular models for image SR reconstruction. DRCA first uses a convolutional layer to extract the shallow features of the input image. Secondly, the shallow features, dense connection residual group, and the last projection layer (1×1 convolutional layer) are connected by long skip connection to obtain a larger receptive field. Then the upscale module composed of convolutional layer and pixel shuffle layer (PSL) is used for upsampling. Finally, a convolution layer is used to map the upscaled features to the SR image to obtain the reconstructed high-resolution image. DRCA is mainly composed of five residual blocks, and the vanilla convolution layers in the residual blocks bring many parameters, which leads to the requirement of large storage space and high computational cost.

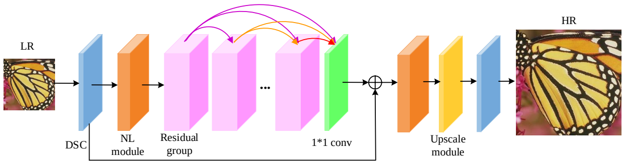


Fig. 1. The framework of LNLN.

In view of the problems existing in DRCA, we propose a lightweight non-local network for image SR. As shown in Fig. 1, the structure of LNLN is composed of four stages. In the first stage, we use a 3×3 depthwise separable convolution layer to extract shallow features, and then use a non-local module to obtain the feature map with spatial correlation. In the second stage, deep feature information is further extracted through densely connected blocks of residuals to obtain more details regions. In the third stage, the fusion feature maps of spatial correlation and channel correlation are obtained through non-local module. In the last stage, the reconstructed image is obtained by upsampling and convolution operation. The LNLN shows two advantages than popular DRCA: (1) the LNLN provide better image SR reconstruction results due to the use of non-local module. (2) the LNLN requires fewer parameters and memory usage due to the employment of DSC.

2.1. Global Context Capture Module (GCCM)

In neural network architectures, the larger receptive field is often obtained by stacking more convolutional layers. However, a vanilla convolutional layer only provides local spatial information while missing global spatial information. To solve the problem of integrating both local and global spatial information to SR networks, we use non-local [11] operations to directly calculate the relationship between two locations, so as to quickly capture the long-distance correlation and get more global feature information.

Non-local operations can be defined as:

$$y_i = \frac{1}{c\left(x\right)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right) \qquad (1)$$

where

$$f\left(x_i, x_j\right)=\exp\left(\theta\left(x_i\right)^T \emptyset\left(x_j\right)\right) \tag{2}$$

$$c\left(x\right)=\sum_{\forall j} f\left(x_i, x_j\right) \tag{3}$$

the x is the input feature map, i is the output feature position index and j is the index of all possible positions. x and y are inputs and outputs of non-local operations. The $f\left(x_i, x_j\right)$ calculates the similarity between x_i and x_j. the function $g(x_j)$ is the representation of feature map at position j, y is obtained by standardizing the response factor $c\left(x\right)$.

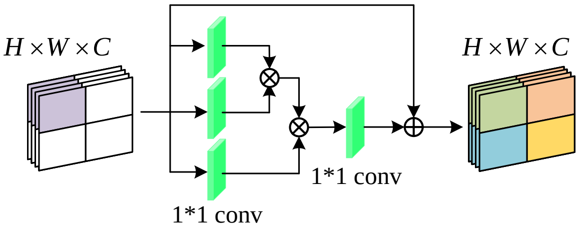


Fig. 2. Non-local module.

Although the non-local module can obtain more comprehensive feature information, it will cause a substantial increase in computing cost. Therefore, we only add non-local blocks in low-level and high-level feature spaces, which is a good way to supply global information to the network for image restoration while slightly increasing the computation cost. In addition, we divide the feature map into four regional grids as shown in Fig. 2. After each grid performs a non-local operation, the area grids are reconnected to the feature map and processed by subsequent layers, which further effectively reduces the computational cost of the non-local module [12]. In fact, the non-local module not only ensures the training efficiency of the network, but also helps the network learn more complex and useful features to obtain higher reconstruction performance.

2.2. Model compression

Generally, with the deepening of network models, the feature representation will be improved and the number of parameters will be multiplied. A very deep network often requires a lot of storage resources and easily causes over fitting risk. Therefore, we use DSC [14] instead of vanilla convolution to construct residual groups, which can effectively compress the network model while ensuring the network training accuracy. The vanilla convolution needs to consider both spatial information and channel correlation. In DSC, the channel convolution is performed first, and then the pointwise convolution is performed. DSC not only reduces the computational complexity of convolution operation and reduces the number of parameters, but also increases the network width and extracts richer feature information, it thus does not cause too much loss to the accuracy of reconstruction results. Suppose that the size of the input feature map in the vanilla convolution is $W \times H \times C_1$, the output feature map size is $W \times H \times C_2$, where W and H are the width and height of the feature map, respectively. C_1 and C_2 represent the numbers of channels, the size of convolution kernel is $K \times K$, then the computational complexity of vanilla convolution is $K^2 \times W \times H \times C_1 \times C_2$ while the computational complexity of DSC is $W \times H \times (K^2 \times C_1 + C_1 \times C_2)$. Compared with the vanilla convolution, the computational complexity of DSC can be reduced to $(1/C_2 + 1/K^2)$ of the vanilla convolution. In LNLN, we replace the vanilla convolution in the residual group with DSC and construct a lightweight residual group (LRG) module. The parameter number of the LRG is 0.38M, which is only 14.12% of the corresponding module in DRCA network. The whole network model is compressed from 55.4MB to 8.72MB.

3. EXPERIMENTS

Experiments are performed on a workstation with Intel Core i7 8700X @ 3.2GHz, 64GB RAM, NVIDIA GeForce RTX 2080Ti GPU, Windows 10 Pro, and PyTorch 0.4.

We used 800 high-resolution images from DIV2K dataset [15] as training set, and used Set5 [16], Set14 [17], BSD100 [18], Urban100 [19], Manga109 [20] as test sets. Firstly, LR image is obtained by bicubic downsampling of HR image. Then, we increase the number of samples by randomly rotating 90°, 180°, 270° and horizontally flipping.

3.1. Training

For the LNLN training, the hyper-parameter values are set as follows. The initial learning rate is 10^{-4}, and then reduced to half every 200 epochs. We use L1 loss as training loss and Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 10^{-8}$. In addition, we use the ReLU as the activation function. After the training, we use the optimal model parameters to reconstruct the test set. We transform the SR image into YCbCr space and evaluate the reconstructed image effectively on Y channel.

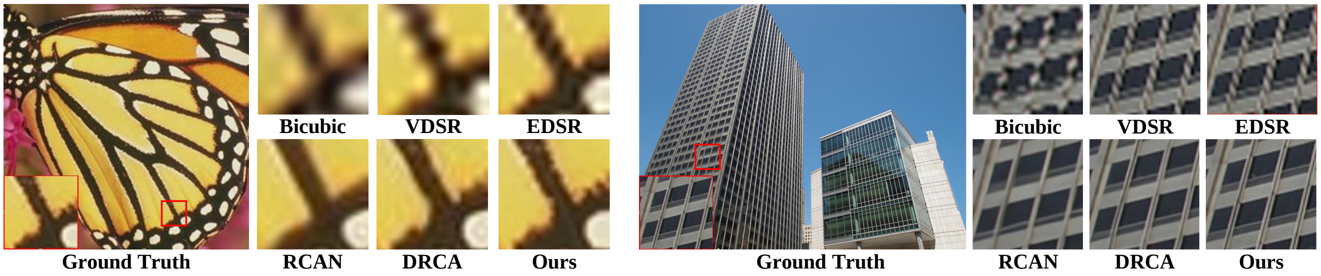
3.2. Evaluation and results

We selected peak signal to noise ratio (PSNR) and structural similarity (SSIM) as important indicators to evaluate the quality of image reconstruction. The higher values of PSNR and SSIM correspond to the better SR reconstruction effect.

In Table 1, it can be seen that the values of PSNR and SSIM provided by the proposed method on Set5, Set14, B100, Urban100 and Manga109 datasets are basically higher than comparative models. Although the results of Manga109 are slightly lower than those on RCAN when the scaling factor

Table 1. Quantitative comparisons with different networks for scale factor of 2 and 4.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.8	0.9339
VDSR [5]		37.53	0.959	33.05	0.913	31.9	0.896	30.77	0.914	37.22	0.975
EDSR [13]		38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.1	0.9773
RCAN [9]		38.27	0.9614	34.12	0.9216	32.41	0.9027	33.26	**0.9384**	**39.44**	**0.9786**
DRCA [10]		38.28	0.9615	34.13	0.923	32.39	0.9023	33.25	0.9382	39.4	0.9779
Ours		**38.32**	**0.9618**	**34.15**	**0.9232**	**32.42**	**0.9028**	**33.27**	**0.9384**	39.43	0.9784
Bicubic	4	28.42	0.8104	26	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
VDSR [5]		31.35	0.883	28.02	0.768	27.29	0.7251	25.18	0.754	28.83	0.887
EDSR [13]		32.46	0.8968	28.8	0.7876	27.71	0.742	26.64	0.8033	31.02	0.9148
RCAN [9]		32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
DRCA [10]		32.68	0.9009	28.91	0.7898	27.8	0.7444	26.94	0.8111	31.37	0.9182
Ours		**32.72**	**0.9012**	**28.92**	**0.7899**	**27.82**	**0.7446**	**26.96**	**0.8112**	**31.38**	**0.9183**



(a) Visual comparison for scaling factor s=2

(b) Visual comparison for scaling factor s=4

Fig. 3. Visual comparison of different models when s=2 and s=4.

is 2, the results on other datasets are better than those on RCAN. As we all know, with the increase of scaling factor, the image reconstruction will become more difficult. When the scaling factor is 4, the proposed LNLN obtains higher values of PSNR and SSIM than other models as shown in Table 1. In addition, the LNLN provides a clearer image texture with fuller edges and sharper results that are closer to the original image in Fig. 3.

Table 2 shows the comparison of the number of training parameters and storage usage of networks. Compared with EDSR, RCAN and DRCA, the LNLN achieves a significant compression on the model size. Although the number of trainable parameters and model size are slightly higher than VDSR, the LNLN achieves better reconstruction results due to the use of the non-local module. In summary, our proposed LNLN achieves the better balance between the reconstruction effect and the model size.

Table 2. Comparison of the efficiency of different networks.

Method	Parameters (M)	Storage usage (MB)
VDSR	0.66	2.6
EDSR	41.3	10.8
RCAN	15.8	34.9
DRCA	14.2	55.4
ours	**2.04**	**8.72**

4. CONCLUSION

In this paper, in order to ensure the accuracy of image reconstruction and reduce the memory requirement of CNNs, we propose a lightweight non-local network for image super resolution. On the one hand, we use non-local module to obtain the long-distance dependence of images and global information leading to better image SR results. On the other hand, we use DSC to reduce the number of model parameters. Finally, the propose LNLN achieves a good balance between reconstruction effect and model size. Experiments demonstrate that the proposed LNLN shows the advantages of higher reconstruction accuracy, fewer parameters and lighter model compared with the popular networks in five commonly used datasets.

In the future, we will study the application of semi supervised learning and few-shot learning in image super-resolution reconstruction.

5. REFERENCES

[1] Wen Ma, Zongxu Pan, Jiayi Guo, and Bin Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3512–3527, 2019.

[2] Daniel McDuff, "Deep super resolution for recovering physiological information from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1367–1374.

[3] Yunfeng Zhang, Qinglan Fan, Fangxun Bao, Yifang Liu, and Caiming Zhang, "Single-image super-resolution based on rational fractal interpolation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3782–3797, 2018.

[4] Rushi Lan, Long Sun, Zhenbing Liu, Huimin Lu, Zhixun Su, Cheng Pang, and Xiaonan Luo, "Cascading and enhanced residual networks for accurate single-image super-resolution," *IEEE transactions on cybernetics*, 2020.

[5] 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*, 2016, pp. 1646–1654.

[6] 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*, 2018, pp. 2472–2481.

[7] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[8] 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)*, 2018, pp. 517–532.

[9] 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)*, 2018, pp. 286–301.

[10] Dong-Won Jang and Rae-Hong Park, "Densenet with deep residual channel-attention blocks for single image super resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[11] 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*, 2018, pp. 7794–7803.

[12] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11065–11074.

[13] 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*, 2017, pp. 136–144.

[14] 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*, 2017, pp. 3147–3155.

[15] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 114–125.

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

[17] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[18] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. IEEE, 2001, vol. 2, pp. 416–423.

[19] 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*, 2015, pp. 5197–5206.

[20] 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*, vol. 76, no. 20, pp. 21811–21838, 2017.