Journal of Physics: Conference Series

Deep Residual Fusion Network for Single Image Super-Resolution

Deep Residual Fusion Network for Single Image Super-Resolution

Jia Wang, Chuwen Lan, Zehua Gao*

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

gaozehua@bupt.edu.cn

Abstract. Convolutional neural networks have been applied in the field of single-image super-resolution (SISR) and have achieved a series of outstanding results. However, most of the SISR research still attempt to pursue wider and deeper network structure, without paying enough attention to the correlations between different features. In order to solve these problems, deep residual fusion network (DRFN) is proposed for more powerful feature expression and feature learning. Specifically, we propose a feature fusion group (FFG) structure, which can effectively use the relevant features extracted from the residual attention group (RAG) and fuse them to be more discriminative. Residual attention group (RAG) includes channel attention module (CAM) and spatial attention module (SAM), which uses attention mechanism to refine features. DRFN also makes full use of nested residual connections, skipping redundant low-frequency information to enhance circulation, thereby focusing the calculation on more important high-frequency components. Extensive experimental results have proved the effectiveness of our model. And our model finally achieves excellent performance in terms of both quantitative metrics and visual quality.

1. Introduction

The single-image super-resolution (SISR) problem aims to obtain corresponding high-resolution (HR) images from low-resolution (LR) images, which can be applied to medical imaging [1] and object recognition [2]. The issue has received widespread attention. However, any LR image can correspond to multiple HR images, so the SISR problem is an ill-posed problem. In order to solve this problem, a large number of super-resolution (SR) methods have been proposed to learn the mapping from LR to HR.

Recently, deep convolutional neural networks (CNNs) have achieved unprecedented success in various fields. In the field of SISR, methods based on CNNs have made significant progress compared with traditional methods. A series of CNN-based methods have been proposed, such as SRCNN [3], VDSR [4], EDSR [5]. Using CNN to learn the representation of structured data, the performance of SISR continues to improve. The depth of the network is further deepened, and the connectivity of the architecture is growing. Although the latest results have made great progress, the structure based on CNNs still has some limitations.

On the one hand, most CNN-based SR methods do not make full use of the original information of LR images, and only consider the design of deeper and wider networks. But blindly stacking residual blocks can not achieve better performance. On the other hand, the original LR image has a lot of low-frequency information, which does not need too much calculation. Not paying enough attention to the acquisition of high-frequency information, a lot of high-frequency information is lost in the process of reasoning.

In order to solve these problems, we propose the deep residual fusion network (DRFN) to make full use of the information of the original LR image. We regard the residual attention group (RAG) as the basic module of the network, and improve the resolution of the network through nested residual connections. Inspired by [6], we apply the attention mechanism to the model. The channel attention module (CAM) is added to learn the interdependence between feature channels, and the spatial attention module (SAM) is used for feature refinement, so that our proposed network can focus on more useful positions. Feature fusion group (FFG) fuse multi-scale features to effectively utilize global information, and uses the complementarity between high and low-level features to extract more comprehensive LR features.

Compared with other advanced SR methods, our method obtains better visual quality and restores more image details. In summary, the main contributions of this article are as follows:

(1) A very efficient DRFN network is proposed, which can achieve higher precision on image SR. Compared with other SR methods,

(2) Our DRFN can use the attention mechanism and feature fusion method to effectively combine local and global information, and extract and fuse LR image features more efficiently to obtain better SR performance.

(3) We can skip a large amount of low-frequency information through nested residual connection, and concentrate the computing power on obtaining effective information.

2. Related Work

As a hot topic in the field of computer vision, SISR has been widely studied. The method to solve the problem proposed by the previous researchers is based on interpolation [7], which is simple and fast. However, applying the same method to all images means that the HR image obtained may not be the optimal solution of this LR image. Later, methods based on priors were proposed, such as non-local similarity prior [8] and sparse prior [9]. Compared with interpolation-based methods, this method specifically generates HR images with higher quality, but still has defects in statistics and application of prior information.

In the field of SISR, Dong *et al.* first applied CNN to SR and proposed an SRCNN [3] with three convolutional layers, which achieved the best performance of SR at that time. When He *et al.* proposed ResNet [10], SISR introduced the concept of residual learning. Residual learning bypasses the useless low-frequency information and focuses on the learning of the high-frequency part between the HR image and the LR image. Therefore, it is particularly suitable for SISR. Kim *et al.* designs VDSR [4] and DRCN [11] based on residual learning to deepen the network layer and achieve better results.

Then ledig et al. used the structure of ResNet [10] to build a deeper network SRGAN [12]. Based on their results, Lim et al. proposed EDSR [5] to modify the stacked residual block and delete unnecessary content. With EDSR [5], the peak signal-to-noise ratio (PSNR) has been significantly improved. Network depth is very important for vision tasks, but stacking too many residual blocks does not bring improvement.

In addition to considering depth, other networks such as DBPN [13] use feedback mechanisms to build interdependent up-sampling modules, and DRRN [14] uses recursive blocks. However, most of these methods do not specifically consider the acquisition of high-frequency information, which hinders the improvement of deep network capabilities of feature extraction.

3. Deep Residual Fusion Network (DRFN)

3.1. Network architecture

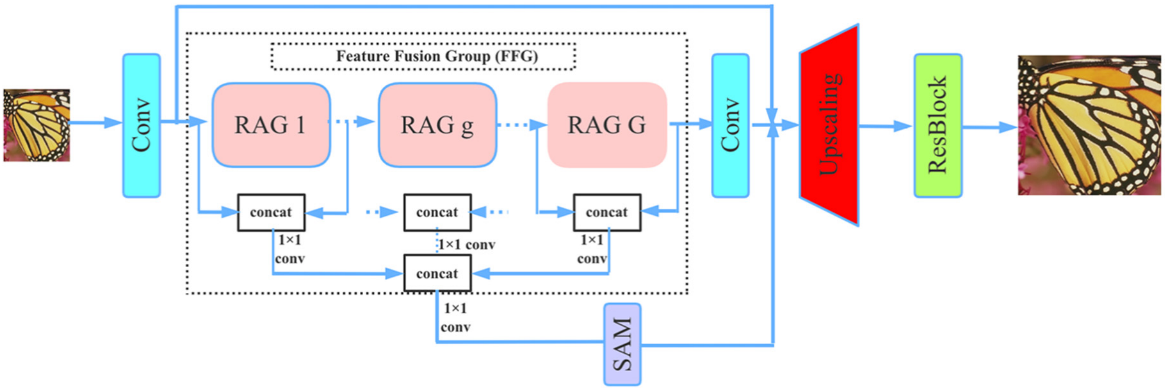

Fig.1 The proposed network architecture DRFN

As shown in Fig.1, our deep residual fusion network (DRFN) is mainly composed of four parts: shallow feature extraction, feature fusion group（FFG）based deep feature extraction, upscale module, reconstruction part.

Taking the low-resolution image I_{LR} as the input of the network and the super-resolution image I_{SR} as the output, we first extract the shallow feature from the LR:

$$F_{SF1} = \Phi_{SF1}(I_{SR}) \tag{1}$$

Φ_{SF1} represents the convolution operation of shallow feature extraction, and F_{SF1} is the feature extracted after this operation. Next, we use F_{SF1} as the input of deep feature extraction to generate deep feature

$$F_{DF} = \Phi_{\mathrm{RA}}(F_{SF1}) + \Phi_{FF}(F_{SF1} + F_{SF2} \ldots + F_{SFG}) + F_{SF1} \tag{2}$$

Φ_{RA} represents G residual attention groups (RAGs) in the deep feature extraction part. Φ_{FF} represents the feature fusion (FF) operation, F_{SFG} is the feature obtained after passing the $G - th$ RAG. And this part uses residual learning to add the extracted features to F_{SF1}. As far as I know, our method allows the network to achieve a deeper depth, provides a very large receptive field, and bypass the abundant low-frequency information, which is the most important part of the network to extract features. Finally, the upscale module is used to enlarge the image and reconstruct it

$$I_{SR} = \Phi_{\mathrm{REC}}\big(\Phi_{\mathrm{UP}}(F_{DF})\big) = \Phi_{\mathrm{SR}}(I_{LR}) \tag{3}$$

Φ_{UP} represents the upscale module. Φ_{REC} represents the reconstruction module. And Φ_{SR} represents our whole model. At present, there are pre-upscale, transposed convolution [15], and ESPCN [16] methods. After experiments, it is found that the ESPCN [16] method is better than pre-upscale, and can achieve a balance between performance and computational burden.

Finally, DRFN should be optimized through loss function. Some loss functions including L1, L2, perception loss, etc. have been applied to SISR. In order to show the effectiveness of the model, we choose L1 loss. Given N LR, HR images $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$

$$L_C(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathrm{H}_{\mathrm{SR}}\big(I_{SR}^i\big), I_{HR}^i\right\|_1 \tag{4}$$

θ represents the parameters of the network. The loss function adopts the method of adam [17] gradient optimization. More training and testing details will be given in Sec. 4.1.

3.2. Feature fusion group(FFG)

One of our important innovations is the efficient extraction of deep features. Now we will describe this part in detail. The deep feature extraction part is composed of feature fusion group (FFG), which includes G residual attention groups (RAG). As shown in Fig.2, each RAG includes M residual blocks, channel attention module (CAM) and spatial attention module (SAM). This structure can be designed to train a CNN network with more than 400 layers.

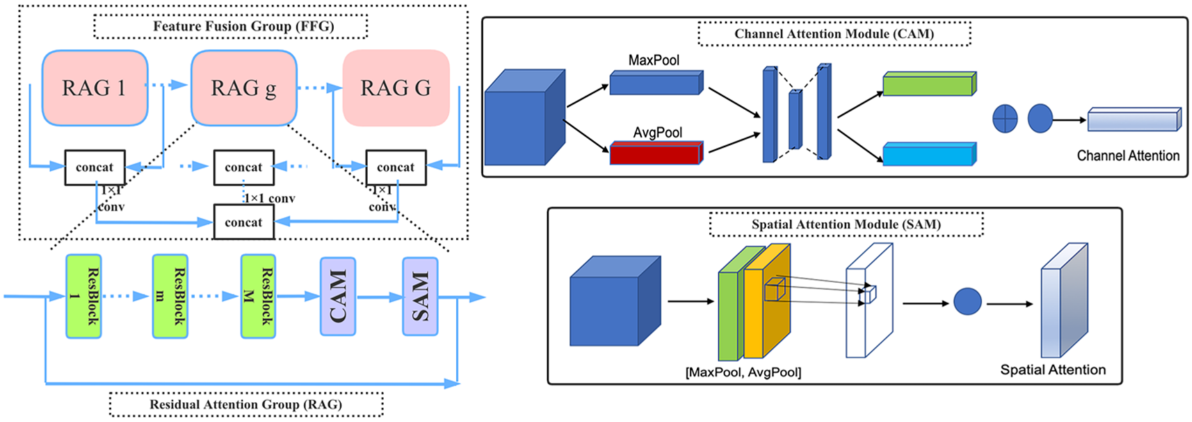

Fig.2 Proposed residual attention group (RAG)

Studies have proved that deepening the layers of CNNs is vital in computer vision tasks, but blindly stacking basic modules in SISR does not improve performance. There are problems such as gradient disappearance and explosion. With this in mind, we thought about designing a hierarchical structure and proposed RAG as our basic module. We stacked$M$ residual blocks to form a basic RAG, and used the attention mechanism to get more discriminative representations.

Let F_{SFg} be the feature obtained after $g-$th RAG, the$g-$th RAG can be expressed as

$$F_{SFg} = \Phi_{\mathrm{SA}}\big(\Phi_{\mathrm{CA}}(X_{g,M})X_{g,M}\big)\Phi_{\mathrm{CA}}(X_{g,M})X_{g,M} + F_{g,\mathrm{M}-1} \tag{5}$$

Φ_{SA} and Φ_{CA} represent spatial attention and channel attention operations, respectively. F_g is the feature obtained after g RAG modules, and $X_{g,m}$ is the feature obtained after residual blocks. Feature fusion (FF) operation splices and convolves the adjacent features extracted by RAG to fuse the features. After that, SAM operation is performed to refine the features. Finally, all the feature maps are merged together

$$F_{FF} = \Phi_{FF}\big(F_{SF1} + \cdots + F_{SFg} \ldots + F_{SFG}\big) \tag{6}$$

Φ_{FF} is the FF operation, F_{FF} is the feature after FF.

3.3. *Implementation details*

We now describe the details of the proposed DRFN algorithm. First, we applied a convolutional layer with a kernel size of 3×3 for shallow feature extraction. After the convolution, the kernel size of the channel-downscaling, channel-upscaling, and feature fusion convolution layers of CAM is 1×1. The convolution kernel of SAM is 7×7. In addition to these, all others use a convolution kernel of 3×3 to keep the size of the feature map fixed.

In the part of deep feature extraction, we set the number of RAG to 16. There are 8 residual blocks in each RAG. In CA module, we use 1×1 convolution filter with reduction ratio 16, thus channel-downscaling has 4 filters. The other RAG convolution filters are all 64. For upscale module, we choose to use ESPCN [16]. The final convolutional layer outputs 3 channels to produce RGB images.

4. Experiments

4.1. *Setting details*

(1) datasets

We choose the DIV2K [18] data set as the training set, which includes 800 high-definition training pictures and 100 test pictures. For testing, we use 5 standard benchmark datasets: Set5 [19], Set14 [20], B100 [21], Urban100 [22], and Manga109 [23]. The final SR image is evaluated by peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) indicators on Y channel of transformed YCbCr space.

(2) training settings

In each mini-batch, we randomly select 16 LR patches as the network input. The model uses Adam [17] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 10e\text{-}8$. The initial value of the learning rate is set to 10e-4, and the halving operation is performed every 1000 epochs. The algorithm uses the PyTorch [24] framework on an Nvidia 1080Ti GPU. each epochs requires approximately 80s.

In order to show the effectiveness of the model, the experiment was performed on Bicubic (BI) degradation models [25], and the scaling factor was selected as 4.

4.2. Ablation study

The main modules of DRFN are FFG and RAG. We use the Set14 [20](×4) data set as the test set to compare the PSNR of model variants and measure the effects of FF and connection of RAG in the network. The results are shown in Tab.1.

(1) Feature fusion (FF)

To prove the effect of our proposed FF, we removed this operation from the very deep network. When FF is deleted, the PSNR value on Set14 [20] is relatively low, regardless of other parts. Compared with the original network after adding FF, the performance has been improved. These comparisons prove that our proposed FF structure is effective for deep networks.

(2) long and short connection

First of all, we found that the model with long and short connection is much better than the baseline, indicating that the module is useful. After adding other modules, it can still improve PSNR. It is very difficult to improve the PSNR when the current network depth is very deep, and RAG can improve the algorithm performance because the module can make the low frequency information flow and extract high frequency information more.

Table 1. Investigations of Feature fusion (FF) and long and short connection

Feature fusion (FF)		√		√
long and short connection			√	√
PSNR on Set14 (4×)	27.56	27.76	27.70	27.79

4.3. Compare with the state-of-the art methods

Compare our proposed DRFN with the latest algorithms: SRCNN [3], FSRCNN [26], VDSR [4], LapSRN[27], MemNet [28], EDSR [5], SRMDNF [25], RDN[29], DCSR [30] and MSRN [31]. Tab.2 shows the quantitative evaluation results of our ×4 scale factor.

The results clearly show that our method is superior to most current methods. After self-ensemble operation, even the highest PSNR was achieved in all models. Although RDN [29] shows better performance, we can see from the Fig.5 that it has approximately 22M parameters. Compared with the 5.2M parameters of our model, we have achieved similar performance with huge differences in parameters. Our DRFN can use FF to integrate global information without excessively increasing the depth and width of the network to extract more resolving features. The network we proposed achieves a balance between computing power and accuracy.

Visual quality

Moreover, we have given qualitative results in the Fig.4 and Fig.5, and the results of other methods are all from RDN [29]. Observed by the image "img_004", the images restored by most methods are relatively blurry. Only our images can get a more accurate lattice structure and restore more details. In the "Yumeiro-Cooking" image, we can also see more details restored by our proposed DRFN model, which is different from the fuzzy lines produced by other methods.

These results all show that our model can extract deep features in the image. The recovery of high frequency information is more difficult. But our DRFN can get more image details.

Table 2. Quantitative Comparisons of state-of-the-art methods for BI degradation model.

Method(×4)	Set5[19]		Set14[20]		B100[21]		Urban100[22]		Manga109[23]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRCNN [3]	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
FSRCNN[26]	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280	27.90	0.8610
VDSR [4]	31.35	0.8830	28.02	0.7680	27.29	0.0726	25.18	0.7540	28.83	0.8870
LapSRN [27]	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet [28]	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
EDSR[5]	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	-/-	-/-
SRMDNF[25]	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
RDN [29]	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
DCSR [30]	31.58	0.8870	28.21	0.7715	27.32	0.7264	27.24	0.8308	-/-	-/-
MSRN [31]	32.07	0.8903	28.60	0.7750	27.52	0.7273	26.04	0.7896	30.17	0.9034
DRFN	32.16	0.8936	28.62	0.7797	27.55	0.7345	25.92	0.7795	30.26	0.9027

4.4. Model complexity analysis

We hope to get better results and improve the accuracy of the results with limited computing resources. Therefore, we compare the performance of the models. In Fig.5, we can find that our model achieves similar performance with fewer parameters than RDN [29] and EDSR [5]. These results all show that our model performs better under the same computing power.

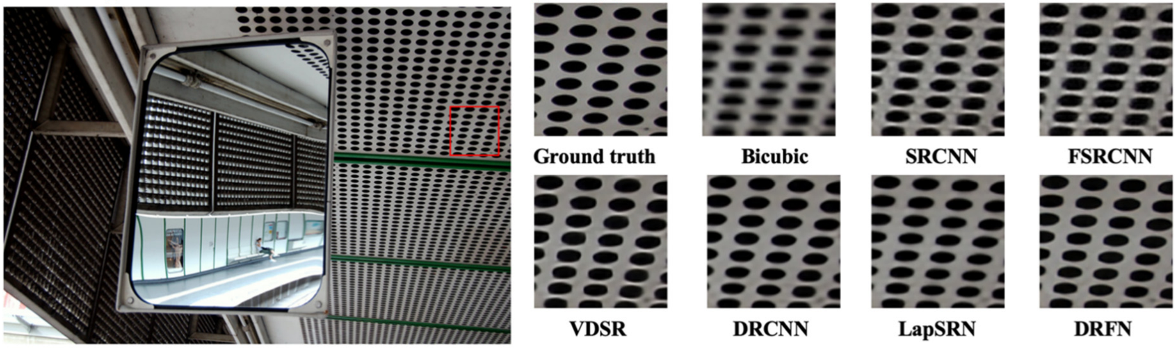

Fig.3 Qualitative results for 4× SR with BI model on Urban100 [22]

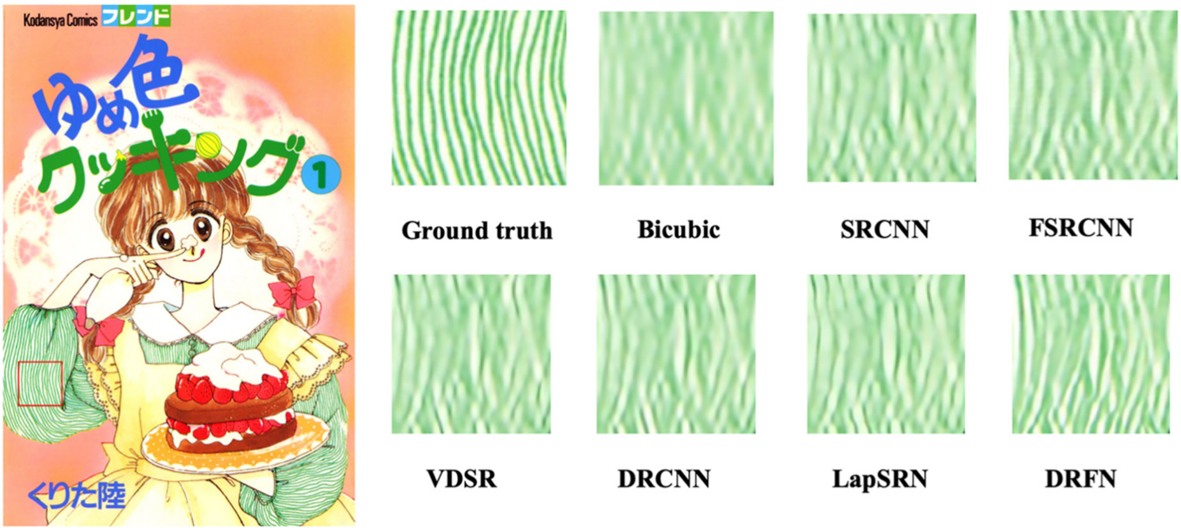

Fig.4 Qualitative results for 4× SR with BI model on Manga109 [23]

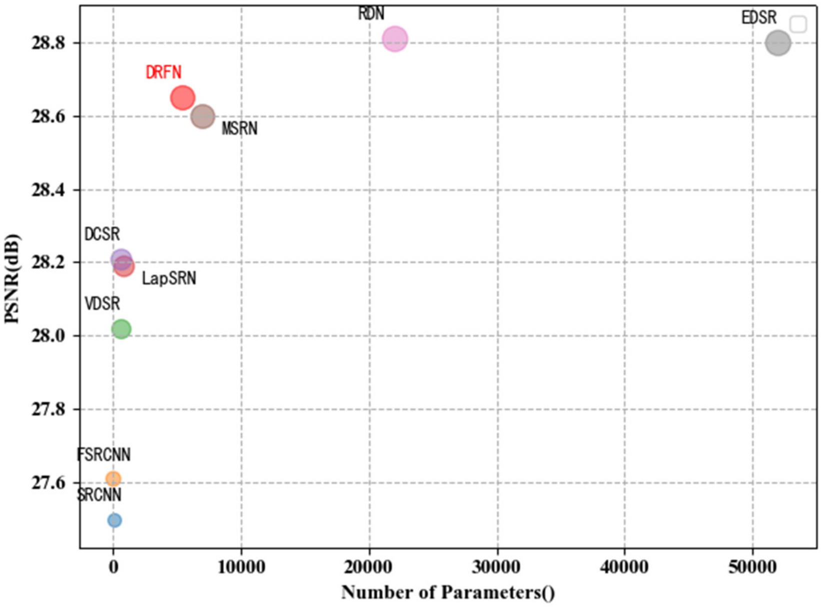

Fig.5 Comparison of memory and performance. Results are evaluated on Set14 [20](×4).

5. Conclusion

We proposed deep residual fusion network (DRFN) to obtain high-quality super-resolution (SR) images. Specifically, feature fusion group (FFG) uses the attention mechanism to learn the correlation between features. Feature fusion utilizes the complementarity between high and low level features to extract more comprehensive features. At the same time, the network uses residual learning to circulate redundant information and concentrate the calculations on high-frequency components. A large number of experiments on SR using the bicubic (BI) degradation model proved the effectiveness of DRFN.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities NO. 2020RC02; National Natural Science Foundation of China (61905021, 61672108, 61976025); State Key Laboratory of Millimeter Waves (K202008); Shenzhen Technical Plan Project (JCYJ20180305164708625).

References

[1] Shi W, Caballero J, Ledig C, et al. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2013: 9-16.

[2] Sajjadi M S M, Scholkopf B, Hirsch M. Enhancenet: Single image super-resolution through automated texture synthesis[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 4491-4500.

[3] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[4] Kim J, Kwon Lee J, Mu Lee K. Accurate image super-resolution using very deep convolutional networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1646-1654.

[5] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.

[6] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[7] Zhang L, Wu X. An edge-guided image interpolation algorithm via directional filtering and data fusion[J]. IEEE transactions on Image Processing, 2006, 15(8): 2226-2238.

[8] Zhang K, Gao X, Tao D, et al. Single image super-resolution with non-local means and steering kernel regression[J]. IEEE Transactions on Image Processing, 2012, 21(11): 4544-4556.

[9] Dong W, Zhang L, Shi G, et al. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization[J]. IEEE Transactions on image processing, 2011, 20(7): 1838-1857.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Kim J, Kwon Lee J, Mu Lee K. Deeply-recursive convolutional network for image super-resolution[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1637-1645.

[12] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[13] Haris M, Shakhnarovich G, Ukita N. Deep back-projection networks for super-resolution[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1664-1673.

[14] Tai Y, Yang J, Liu X. Image super-resolution via deep recursive residual network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3147-3155.

[15] Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style[J]. arXiv preprint arXiv:1610.07629, 2016.

[16] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[18] Timofte R, Agustsson E, Van Gool L, et al. Ntire 2017 challenge on single image super-resolution: Methods and results[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 114-125.

[19] Bevilacqua M, Roumy A, Guillemot C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[J]. 2012.

[20] Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations[C]. International conference on curves and surfaces. Springer, Berlin, Heidelberg, 2010: 711-730.

[21] Martin D, Fowlkes C, Tal D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]. Vancouver: Iccv, 2001.

[22] Huang J B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5197-5206.

[23] Matsui Y, Ito K, Aramaki Y, et al. Sketch-based manga retrieval using manga109 dataset[J]. Multimedia Tools and Applications, 2017, 76(20): 21811-21838.

[24] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch[J]. 2017.

[25] Zhang K, Zuo W, Zhang L. Learning a single convolutional super-resolution network for multiple degradations[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3262-3271.

[26] Dong C, Loy C C, Tang X. Accelerating the super-resolution convolutional neural network[C]. European conference on computer vision. Springer, Cham, 2016: 391-407.

[27] Lai W S, Huang J B, Ahuja N, et al. Fast and accurate image super-resolution with deep laplacian pyramid networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2599-2613.

[28] Tai Y, Yang J, Liu X, et al. Memnet: A persistent memory network for image restoration[C]. Proceedings of the IEEE international conference on computer vision. 2017: 4539-4547.

[29] Zhang Y, Tian Y, Kong Y, et al. Residual dense network for image super-resolution[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.

[30] Zhang Z, Wang X, Jung C. DCSR: Dilated convolutions for single image super-resolution[J]. IEEE Transactions on Image Processing, 2018, 28(4): 1625-1635.

[31] Li J, Fang F, Mei K, et al. Multi-scale residual network for image super-resolution[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 517-532.