Parallel Residual Attention Network for Image Super-Resolution

Yongsheng Duan¹,*, Yang Su¹,*, Jiahao Xu², Wei Wu¹

¹School of Information Engineering, Wuhan University of Technology, Wuhan, China
²School of Information Engineering, Chang’an University, Xian, China

*Corresponding author e-mail: 804666940@qq.com, 94520022@qq.com

Abstract. The application of convolution neural networks (CNN) to image super-resolution has achieved excellent results. Inspired by these excellent results, we find that most models only increase the depth of the network in order to obtain deeper features. However, these models often cannot take full advantage of the original feature information from the low-resolution (LR) images. Moreover, the image information is mainly divided into high frequency information and low frequency information, and the existing image super resolution network models are mostly single-branch models, which limits the performance of extracting various aspects of image information. To resolve the problem, we proposed a parallel network model. About the network, we proposed the LR residual block (LRB) and attention block (AB). LRB allows constantly replenishing original LR features in the process of extracting deep features, thus it can effectively extract more information from LR image. The AB contains the attention mechanism, it will adjust their weights based on the importance of the image features to extract more efficient features between low-resolution (LR) images and high-resolution (HR) images. Experiments on benchmark datasets show that our model performs better. We achieve higher accuracy and visual improvements against state-of-the-art methods.

1. Introduction

Image super-resolution generally refers to the single image super-resolution (SISR). SISR aims to get a satisfactory high-resolution (HR) image from a degraded low-resolution (LR) image. In recent years, with the introduction of some typical networks and the improvement of hardware, super-resolution (SR) has achieved many significant breakthroughs. SISR has also been applied to many fields, such as security imaging [2], medical imaging [3], satellite imaging and so on. Due to the outstanding advantages of deep learning in image super-resolution, people pay more attention to the application of deep learning in super-resolution, and there are also many typical network models, such as: SRCNN [1], VDSR [4] and DRCN [5], EDSR [6]etc. While improving the effectiveness of these networks, they are constantly deepening the network. Furthermore, RDN [7] proposed by Zhang et al. In an RDN block, units are connected like DenseNet [8]. Before entering the reconstruction part, features from all previous blocks are fused by dense connection and residual learning. However, these typical models have some drawbacks:
(1) These typical models have all focused on the network depth. The improvement of network depth is necessary for getting the deep feature information, but not the deeper the network, the better. The deeper the network depth, the greater the amount of computation, and the loss of much of the original information from LR image during processing. Although RDN proposed by Zhang improved the shortcoming, it still did not make full use of the original information.

(2) Many recent models treat the features extracted from the LR images without distinction, which makes network unobtrusive to extract features and wastes much time in computing. Everyone knows that a picture is composed of low frequency information and high frequency information. We think a single network is hard to take care of both kinds of information.

Inspired by these drawbacks, we proposed our network to address them. Our network model contains the LR residual block (LRB) and the attention block (AB). LRB allows constantly replenishing original LR features in the process of extracting deep features, it can effectively solve the problem of the loss of the original information from LR image during training. AB consists of attention mechanism. It can make network to pay more attention to certain information, which is high frequency information. Most attention mechanisms are used for classification tasks. It is well known that high frequency is the texture feature of images, and image classification is mainly based on deep high frequency features.

Considering the characteristics of the two modules, we use two modules to build different branch networks. Experiment shows that the features extracted by the branch composed of the LRB and the features obtained by the branch composed of the AB make the generated HR image details better. In summary, our contributions are three-fold:

(1) We proposed the LR residual block (LRB) to supplement more original information for the generated HR image.

(2) We proposed the attention block (AB), AB can get more effective high frequency information for the SR image reconstruction due to its attention mechanism.

(3) We proposed a parallel two-way neural network model. The model contains two branches, one is made up of LRB and the other is made up of AB. The model can effectively obtain the original information and high frequency information from the LR image and fuse them to reconstruct a pleasing high-resolution (HR) image. are identified in italic type, within parentheses, following the example.

2. Related Work

So far, there have been many models based deep learning. And these models have been recognized by everyone. In this section, we only discuss some works on image SR. SRCNN proposed by Dong et al. is known as the first model using deep learning in super resolution, and the model establishes an end-to-end mapping between the degraded LR images and their reconstructed HR images. This has caused a lot of people to pay more attention to super resolution. People began to build their own models based on SRCNN to continuously increase the network depth. VDSR is the first very deep model used in SISR. Like SRCNN, VDSR also takes the LR image from the bicubic interpolation as input. It introduces the residual learning [9] and contains a 20-layer VGG-net. VDSR increased the network depth, but caused a lot of parameters. Kim et al. Proposed DRCN introducing recursive learning for parameter sharing to reduce parameters. Tai et al. introduced residual units in DRRN [5], these residual units are used to form a recursive block. And each block shares the same parameters for reducing parameters.

EDSR proposed by Lee et al. is the state-of-the-art breakthrough. EDSR removes the usage of batch normalization (BN), which was designed for classification. For classification, representations are very abstract and the correlation of these representations is poor due to the shift brought by BN. But for SISR, correlation between the input and output is very important. Poor correlation may harm the final performance.

Although the improvement of network depth has improved the super-resolution network, it also increases the amount of calculation and weakens the correlation of features, which brings difficulties to HR image reconstruction. To solve the problem, Dong et al. used the transposed convolution layer for upsampling to reconstruct the fine resolution. Shi et al. introduced the sub-pixel convolution layer to upscale the final LR feature maps into HR image in ESPCN. Rather than increasing resolution by
explicitly enlarging feature maps as the deconvolution layer does, ESPCN expands the channels of the output features for storing the extra points to increase resolution, and then rearranges these points to get the HR output through a certain mapping criterion. As the expansion is carried out in the channel dimension. The sub-pixel convolution works well, and it is commonly used.

Recently, RDN proposed by Zhang et al. RDN uses the residual dense block (RDB) as the basic units, which are densely connected like DenseNet, and at the end of an RDN block, a bottleneck layer is used following with the residual learning across the whole block. Before entering the reconstruction part, features from all previous blocks are fused by dense connection and residual learning. Although RDN uses a lot of residual learning to compensate for the loss of information from the original LR image in the convolution process, it only uses the input and output of the previous layer as the input of the next layer. As the network deepens, it still fails to make full use of the original information. Inspired by RDN, we proposed the LR residual block (LRB) and the attention block (AB). We will detail our model in next section.

3. Network

![Figure 1. The Network Structure of PRAN](image)

![Figure 2. (a) Residual block in MDSR [6]. (b) Dense block in SRDenseNet [10]. (c) residual dense block in RDN [7]. (d) our LR residual block](image)
Figure 3. (a) Channel Attention Block. (b) Our Attention block

3.1. Network Structure

As shown in Figure 1, our network structure is divided into four parts: original feature extraction of low-resolution images (LRFE), deep feature extraction network (DFE), feature fusion (FF), and up-sampling network (UPNET). We use ILR and IHR as the input and output of network. Inspired by RDN, we also use two Conv layers to extract original features from LR. Unlike the RDN, we use the output \( F_0 \) of the first Conv layer as the global residual learning for all blocks, as depicted in Figs. 2 and 3.

In the low-resolution image original information extraction part, we use a convolution layer to obtain the initial information \( F_0 \), and then use it as the original information of the image for use by all subsequent modules. As shown below:

\[
F_0 = O_{\text{LRFE}}(I_{LR})
\]

Where \( O_{\text{LRFE}}(*) \) denotes convolution operation of the original feature extraction of low-resolution images (LRFE).

Deep feature extraction, as shown in Figure 1, is composed of two branches, and the upper branch is composed of L LRBS, which is used to extract deep information of the image, which contains most of the low frequency information and some high frequency information. The other branch is composed of A blocks of AB, and the function of the branch is to extract deep details of the image, mainly high frequency information. As shown below:

\[
F_{\text{LRBS}} = O_{\text{LRBS}}(F_1)
\]

\[
F_{\text{ABS}} = O_{\text{ABS}}(F_1)
\]

Where \( O_{\text{LRBS}}(*) \) denotes the operation of the upper branch of deep feature extraction and \( O_{\text{ABS}}(*) \) denotes the operation of the lower branch of deep feature extraction. The is obtained by convolution operation, and are the obtained features by operation. In the feature fusion part, we add the deep information obtained by the two branches in the deep feature extraction part, and then use the two convolution layers to simply extract the fused information for use in the up-sampling part. As shown below:
\[ F_{-FF} = O_{-FF}(F_{-LRBS}, F_{-ABS}) \]  

(4)

Where \( O_{-FF}(\ast) \) denotes the operation of the feature fusion. \( F_{-FF} \) is the result of the feature fusion.

Inspired by ESPCNN [30], after obtaining the required feature information, we use sub-pixel convolution for up-sampling to obtain the HR image. As shown below:

\[ I_{HR} = O_{-UPNET}(F_{-FF}) \]  

(5)

Where \( O_{-UPNET}(\ast) \) denotes the operation of the image reconstruction.

3.2. LR Residual Block

Now, we explain our LRB module in detail. As shown in Figure 2, our LRB block consists of three parts, local residual, global residual learning, and feature combination (FC). For local residual learning, we use the residuals to connect two adjacent convolution layers, so that in the process of extracting features, we can continuously supplement local information and prevent over-fitting, thus appropriately increasing network depth. And our LRB block is different from the Dense.

Dense block uses a lot of residual learning, and it allows any two networks to connect directly, in the same block. This can prevent the loss of information and the over-fitting of the network, thus deepening the network. But it has caused a lot of reuse of the same information, increasing the amount of calculation. LRB reduces the reuse of the same information on the network without losing information. This is also a highlight of the LRB block.

Global residual learning is the biggest bright spot of the LRB block. As shown in Figure 2, in the LRB block, we not only link the adjacent two-layer convolution network, but also compensate for the loss of the original information caused by network deepening by introducing original information (F0). The original information (F0) is combined with the output of the LRB block as the input to the next layer, which we call global residual learning. The introduction of this global residual learning is quite important. On the one hand, it makes full use of the original information from the LR image and provides more original information for the later reconstruction. On the other hand, the original information contains a lot of low-frequency information, the low-frequency information is the profile of the image, which is the basis of an image. The global residual learning makes the convolution network retain this foundation in the process of extracting deep feature information.

Feature combination is a feature information storage mechanism, as shown in the concat1 block in Figure 2, which is equivalent to a memory block. This mechanism is capable of storing the output of each LRB block as follows:

\[ F(\text{concat1}) = FC(LRB_1, LRB_2, ..., LRB_L) \]  

(6)

The storage mechanism combines the feature information of different levels to provide more rich information for the reconstruction of the later image.

3.3. Attention Block

As the name implies, the mechanism of attention is to focus on the aspects of interest and to assign different energies to different things. Here, we mean to give different weights to the different features extracted by the convolution layer, so that the useful information weight becomes larger, and the useless information makes its weight smaller. In this way, the network focuses on more useful features, making useful information more prominent. In the past, the characteristics of the different layers of the network to be extracted from the convolution layer were the same, which could not highlight the deep features of the extraction. Moreover, attention mechanisms are mostly used in natural language
processing and classification tasks. Image classification requires a high degree of detail in the image, which is the high frequency information of the image. As we said above, an image consists of low-frequency information and high-frequency information. The low frequency information represents the outline of the image, and the high-frequency information reflects the details of the image. The classification is more focused on the details of the thing, that is, high-frequency information. Therefore, according to our analysis, we boldly use the attention mechanism to extract the detailed feature information of the image (high frequency information), as shown in (b) of Figure3.

We are inspired by CBAM [11] and use only the channel attention block according to our needs, as shown in (a) of Figure3. The channel attention block passes the input feature map through global max pooling and global average pooling based on width and height, respectively, and then passes through the MLP. The feature of the MLP output is subjected to an elementwise-based addition operation, and then a sigmoid activation operation is performed to generate a final

Channel attention feature map. The channel attention feature map and the input feature map are elementwise multiplied to generate output features.

The attention block we proposed is shown in (b) of Figure3. The AB block consists of three parts, a CAB block, local residual learning and feature combination (FC). The function of the normal convolution layer is to extract feature information and use it as input to the CAB block. The function of the channel attention block is to nonlinearly process the input multi-layer feature information, adjust the weights of different features of different channels, and let the network pay more attention to the detailed features (high-frequency information), thereby obtaining more detail features. The feature join is similar to the LRB. The feature combination makes the feature information of different levels merge together, which can prevent the loss of information and provide different levels of features for image reconstruction, making the information more comprehensive.

4. Experiment

4.1. Datasets and metrics
At present, super-resolution generally adopts the DIV2K dataset [12] because of the diversity of data and the sufficient quantity. The DIV2K dataset is an open source dataset published by Timofte et al. DIV2K has a total of 1000 images and is divided into 800 training images, 100 validation images, and 100 test images. We use 800 training images from DIV2K to train our network model. Then we select 6 validation images in the training process and use 100 validation images as test sets. In addition, for testing, we also use three standard benchmark datasets: Set5, Set14, Urban100. We use Bicubic Interpolation (BI) to degrade the training set. Then we use degraded training set as the input for our network. We use PSNR and SSIM as our evaluation criteria to evaluate our training results.

4.2. Experiment Details
Training Details. In our network, we set all convolution kernel size as $3 \times 3$ (the kernel size is usually an odd, and taken the increasing of receptive field into account) except that in feature combination, whose kernel size is $1 \times 1$. For Convolution layer with kernel size $3 \times 3$, we pad zeros to each side of the input to keep size fixed. And we set the values of L (the number of LRB) and A (the number of AB) to 16. Original feature extraction layers, feature combination layers have 64 filters. Other layers also have 64 filters and are followed by ReLU. The final Convolution layer has 3 output channels, as we output color HR images. We use the L1 loss function according to [14], and our model is trained by ADAM optimizer[13] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The value of the initial rate is $10^{-4}$. The learning rate is halved at every $2 \times 10^{-5}$ iterations. We set the value of epoch to 100. We use tensorflow to implement our network models with a P6000 GPU.
4.3. Experiments Results

We present the experimental results of PRAN networks consisting of our proposed LRB blocks and AB blocks and a comparison with SRCNN, EDSR, RDN. These comparisons are based on test sets: set5, set14 and DIV2K. Here we only compare the task of image bicubic ×3 super resolution on the data sets. In the experiment we set the value of L to 5, and the other values remain unchanged. They were evaluated using PSNR and SSIM.

As shown in Table 1, It clearly shows our PRAN network performance is much higher than the other models. Compared with the RDN model and EDSR model, our network has shown superior performance in both high-definition test sets and non-HD test sets. PRAN contains two branches consisting of blocks and AB blocks, one branch is used to enhance the extraction of the low frequency information of the original input image, and the other is used to extract more high frequency information from the original input image.

Figure 4 shows our partial visualization comparison on scale. In the visual comparison, we only select RDN and PRAN for comparison. Meanwhile the original image is larger, we intercepted the part of the image here as a display.

For the image “img 0900”, we can see that our proposed PRAN restored more details of the letters, which allows everyone to recognize the original meaning of the letters. However, in the RDN model, we can see that the letter “O” is displayed as the letter “C” and the letter “H” at the beginning of the third line is restored to the fuzzy “E”. For the image “img 091”, we can see from the image that the PRAN model retains the original structure of the road surface, while the RDN loses the smoothness of the road surface, making the road rough and fuzzy.

5. Conclusion

In this paper, we proposed a very efficient two-way parallel residual attention network (PRAN) for super-resolution tasks. The network consists of two branches, one of which is composed of our proposed the LR residual block (LRB), and the other branch is composed of the attention block (AB) we proposed. The LRB block uses global residual learning and local residual learning to effectively improve the extraction and utilization of the original information. The AB block uses the attention mechanism and uses residual learning similar to the LRB block, which can extract the high-frequency information of the image more effectively based on the extraction of low-frequency information, making the image details richer and more realistic. In the whole network, we also use the feature connection storage mechanism and the corresponding feature fusion, so that the different levels of features are fully utilized. Experiments show that our network makes full use of the original high and low frequency information of the image and is superior to other single channel networks.

Table 1. Benchmark results. average psnr/ssim values for scaling factor ×3

| Dataset | SRCNN [1] | EDSR [6] | RDN [7] | PRAN (OUR) |
|---------|-----------|----------|---------|------------|
| SET5    | 32.75/0.9090 | 34.65/0.9282 | 34.71/0.9296 | 34.95/0.9430 |
| SET14   | 29.79/0.8320 | 30.52/0.8462 | 30.57/0.8468 | 31.19/0.8579 |
| DIV2K   | / | 31.26/0.9340 | 32.42/0.9221 | 32.56/0.9531 |
Figure 4. Visual comparison for ×3 SR.

Acknowledgments
Here, I am very grateful to Professor Yang Su and Professor Wei Wu for guiding me and providing me with the corresponding experimental environment. I also thank them for their support and help in life. Thanks to the classmate Han Wang, Xin Li and Jiahao Xu.

References
[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, volume 8692, pages 184–199. Springer International Publishing, Cham, 2014.
[2] Wilman W W Zou and Pong C Yuen. Very low resolution face recognition problem. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 21(1):327–40, 2012.
[3] Shi Wenzhe, Caballero Jose, Ledig Christian, Zhuang Xiahai, Bai Wenjia, Bhatia Kanwal, Dawes Tim, O’Regan Declan, and Rueckert Daniel. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. Med Image Comput Comput Assist Interv, 16(Pt 3):9–16, 2013.
[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. arXiv:1511.04587 [cs], November 2015.
arXiv:1511.04587.

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-Recursive Convolutional Network for Image Super-Resolution. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1637–1645, Las Vegas, NV, USA, June 2016. IEEE.

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1132–1140, Honolulu, HI, USA, July 2017. IEEE.

[7] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual Dense Network for Image Super-Resolution. arXiv:1802.08797 [cs], February 2018. arXiv: 1802.08797.

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, Honolulu, HI, July 2017. IEEE.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs], December 2015. arXiv: 1512.03385.

[10] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In IEEE International Conference on Computer Vision, 2017.

[11] Sanghyun Woo, Jongchan Park, Joon Young Lee, and In So Kweon. Cbam: Convolutional block attention module. 2018.

[12] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super resolution: Dataset and study. In IEEE Conference on Computer Vision & Pattern Recognition Workshops, 2017.

[13] D. P. Kingma and L. J. Ba. Adam: A Method for Stochastic Optimization. 2015.

[14] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging, 3 (1):47–57, 2017.