Residual Shuffle Attention Network for Image Super-Resolution

Zhiwei Li¹, Yaping Zhang¹* and Yuwei Yang²

¹School of Information, Yunnan Normal University, Kunming, Yunnan 650500, China
²Nantong Institute of Technology, Nantong 226000, China
Email: 781303842@qq.com; zhangyp@ynnu.edu.cn; 20180044@ntit.edu.cn

Abstract. In order to improve the accuracy of the super-resolution network and reduce the number of model parameters, this paper improves its RCAB module on the basis of RCAN, and builds a reconstruction network RSAN that can improve the quality and efficiency of image super-resolution reconstruction. By replacing the original channel attention module with a more efficient and lightweight shuffle attention, it is mainly used to reduce the number of parameters, supplemented by improving the accuracy; and replacing part of the ordinary convolution in RCAN with split convolution is mainly used to improve accuracy, supplemented by reducing feature redundancy and parameters. The experimental results show that RSAN in this paper can not only obtain better subjective visual evaluation and objective quantitative evaluation, but also reduce the number of network model parameters and improve the efficiency of the network to a certain extent.

Keyword. Shuffle attention; image super resolution; image processing.

1. Introduction

Image super-resolution (SR) reconstruction is a technology that reconstructs low-resolution (LR) images into high-resolution images (HR). It is widely used in monitoring, remote sensing, military, medical and other fields. It has always been one of the research hotspots in the field of digital image processing. At present, there are mainly two types of methods used for image super-resolution reconstruction. The one is reconstruction-based methods, and the other is the deep learning method that is currently used more popular. Dong et al. [1, 2] are the pioneers of deep learning-based super resolution. They first used convolutional neural networks (CNN) to propose a super-resolution convolutional neural network (SRCNN) and fast super-resolution convolutional neural network (FSRCNN) that can accelerate model training. The SRCNN structure includes three parts: feature extraction, nonlinear mapping and reconstruction, which has a profound impact on the later various deep learning-based network structures. Lai et al. [3] proposed the Laplacian super-resolution network (LapSRN), which is similar to a pyramid in overall structure. The method includes a feature extraction branch, an image reconstruction branch, and then a staged up-sampling to reconstruct the image. At the same time, a new loss function, similar to the loss of HED (Holistically-Nested edge detection) [4] is used to guide model training. Lim et al. [5] proposed an enhanced deep super-resolution network (EDSR), which deletes the BN layer in the ResNet network and partially improves the quality of the generated image. Some scholars [6] used the adversarial generative network (GAN) to achieve super-resolution, and have achieved good results in structural similarity (SSIM). However, as the number of neural network layers increases, the methods based on deep learning make training more and more complex, and it is difficult to achieve satisfactory
results.

Since Vaswani et al. [7] introduced the attention mechanism into CNN, the attention mechanism has achieved good results in a series of CV tasks, such as the squeeze and excitation network (SEnet) proposed by Hu et al. [8], which is a channel attention mechanism. The convolutional block attention module (CBAM) method proposed by Woo [9] includes both spatial attention and channel attention. Since then, a large number of methods based on attention mechanisms have emerged in the field of image super-resolution networks. Zhang et al. [10] proposed a residual channel attention network (RCAN), which uses channel attention inside its residual channel attention block (RCAB). Dai et al. [11] conducted a more in-depth study on the channel attention method. They believed that the SEnet proposed by Hu et al. [8] is only a first-order image feature, ignoring higher-order image features, and limiting the performance of the network. Therefore, a second-order channel attention network (SAN) was proposed.

The attention mechanism in the above methods either only considers the channel attention or both channel and spatial attention, but it is not efficient in the fusion of the channel and spatial attention weights, the amount of parameters is also large, and the accuracy of the model is limited. In order to solve the above problems, this paper introduces the lightweight and efficient shuffle attention mechanism proposed by Yang [12] to replace the original channel attention (CA) module in RCAB to reduce the amount of parameters and improve the accuracy; In order to further reduce the amount of model parameters, feature redundancy, and improve accuracy, this paper partially replaces traditional convolution with split convolution (SP) proposed by Zhang et al. [13]. The experimental results show that the residual shuffle attention network (RSAN) proposed in this paper can not only obtain better subjective visual evaluation and objective quantitative evaluation, but also reduce the number of network model parameters and improve the efficiency of the network to a certain extent.

2. Method

2.1. Network Structure

The overall network structure of this article is shown in figure 1, which is divided into three parts.

2.1.1. Shallow Feature Extraction. The shallow feature extraction part contains only a common convolution layer with the kernel size of 3*3, 3 input channels and 64 output channels.

2.1.2. Deep Feature Extraction. The extraction of deep features is completed by the new residual block (residual in residual, RIR) in this paper. The RIR network is composed of 10 external residual blocks and a long skip connection. Replace the last ordinary convolution layer of the fifth external residual block with SP. Each external residual block contains 20 residual shuffle attention block (RSAB). Each RSAB is composed of 2 convolution layers with the same number of input and output channels, a 3*3 convolution kernel, a ReLU activation function, a shuffle attention module, and a short skip connection.

2.1.3. Upsampling and Reconstruction. In the image reconstruction part, the scaling factor S is 4 in this paper. The input channel is C, the output channel is C*4, and the convolution layer with the kernel size of 3*3 and the Pixelshuffle layer are connected in series. The Pixelshuffle layer proposed by Shi et al. [14] completes the actual upsampling process. See equations (1) and (2) for specific calculations.

\[ M = f_{conv}(t_{res}) \]  
\[ M' = \text{Pixelshuffle}(M) \]

where \( t_{res} \) represents the tensor output by the previous residual block set, and \( f_{conv} \) represents a two-dimensional convolution operation with C input channels and C*4 output channels. The dimension of M is \([N, C*4, W, H]\).

The dimension of \( M' \) in equation (2) is \([N, C,W \times \sqrt{4}, H \times \sqrt{4}]\).
So far, the above process has completed the operation of magnifying the feature map by 2 times. Since the scale factor in this article is 4, the above process has to go through again to get $M''$ with the dimension of $[N, C, W\times4, H\times4]$. Finally, we need to pass $M''$ through a convolution layer with an input channel number of $C$, an output channel number of 3, and a convolution kernel size of 3*3 to obtain the final high-resolution image.

![Figure 1. Overall structure diagram of RSAN network.](image)

2.2. Shuffle Attention Module

The attention mechanism has shown its advantages in many computer vision tasks. Yang et al. [12] proposed a lightweight shuffle attention mechanism. Compared with the previous attention mechanism, the biggest difference is that the feature maps are grouped, learned and finally shuffled in the channel dimension, so that the feature information can be communicated with each other between the channels.

![Figure 2. Shuffle attention mechanism [12].](image)

As shown in figure 2, the entire attention mechanism consists of 4 parts:

1. Feature grouping: given an input feature map $X \in \mathbb{R}^{C\times W\times H}$, where $C$, $W$, $H$ represent the channel number, width and height of the feature map respectively. Split the feature map $X$ into $G$ groups for the first time along the channel dimension, namely $X = [X_1, \cdots, X_G], X_i \in \mathbb{R}^{(C/G)\times W\times H}$, then each group is split into two branches along the channel direction $X_{i1}, X_{i2} \in \mathbb{R}^{((C/G)/2)\times W\times H}$. One branch uses the interrelationship between channels to generate a channel attention map, while the other branch uses the spatial relationship between features to generate a spatial attention map. It is found through experiments that $G$ set to 16 is more appropriate.

2. Channel attention: The SEnet proposed by Hu [8] et al. will lead to a large number of parameters, and it is difficult to make a trade-off between speed and accuracy. Wang [15] and others proposed efficient channel attention (ECA), but when the size of the convolution kernel is large, ECA is not applicable. In order to make improvements, Yang [12] and others adopted a lighter solution, which is realized with the combination of global average pooling + scaling + activation function. The specific calculation formulas are as follows:
\[ s = f_{gb}(X_{i1}) = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} X_{i1}(x, y) \]  
\[ X_{i1}' = \sigma(f_c(s)) \cdot X_{i1} = \sigma(W_1s + b_1) \cdot X_{i1} \]

where \( W_1, b_1 \in R^{((C/G)/2)^{1+1}} \) are two parameters that can be trained continuously through the network. \( \sigma \) represents the Sigmoid activation function.

(3) Spatial attention: Spatial attention can be regarded as a supplement to channel attention. Here, the group normalization (GN) operation proposed by Wu et al. [16] is used. The specific formula is as follows:

\[ X_{i2}' = \sigma(W_2 \cdot GN(X_{i2}) + b_2) \cdot X_{i2} \]

where \( W_2, b_2 \in R^{((C/G)/2)^{1+1}} \) are also two parameters that can be trained continuously through the network. \( \sigma \) represents the Sigmoid activation function.

(4) Aggregation: After completing the two kinds of attention learning and re-calibrating the features, the two branches need to be spliced and aggregated, \( X'_i = [X'_{i1}, X'_{i2}] \in R^{((C/G) \times W \times H)} \). Then all the sub-features are aggregated, and finally the channel shuffle operation is performed.

3. Experimental Result

3.1. Basic Experiment

Experimental data set: The DIV2K is selected as training data set, which contains 800 HR training images and 100 verification images. The scaling factor selected in this paper is 4, and the evaluation indicators are peak signal-to-noise ratio (PSNR) and structural similarity measure (SSIM). Train for 30 epochs.

Experimental results: We verify the performance of the algorithm proposed in this article on the DIV2K validation set, set5, set14, and BSD100. Bicubic, EDSR [5], SRGAN [6], RCAN [10], RDN [17] these 5 networks are chosen to compare with RSAN. Among them, Bicubic is just a mathematical interpolation. SRGAN is trained on TensorFlow, RCAN, RDN, EDSR, and RSAN in this article are all trained on PyTorch. The comparison results of PSNR and SSIM are shown in table 1. Finally, as shown in figures 3 and 4, we give some SR images to visually show the reconstruction effect.

It can be seen from table 1 that the method in this paper performs poorly on BSD100 under the 4× scale factor S. The possible reason is that the generalization ability of the model is insufficient. The effect on Set14 and Set5 is not bad. In general, it can be seen that the RSAN proposed in this article can obtain satisfactory reconstructed images.

| Dataset   | S | Bicubic | SRGAN | RCAN | RDN | EDSR | OUR |
|-----------|---|---------|-------|------|-----|------|-----|
|           | P | S       | P     | S    | P   | S    | P   |
| Set5      | 4 | 28.42   | 0.81  | 24.58| 0.77| 28.79| 0.85| 28.44| 0.84| 27.43| 0.80| 29.23| 0.86|
| Set14     | 4 | 26.00   | 0.70  | 22.43| 0.68| 25.74| 0.75| 25.57| 0.74| 24.86| 0.72| 26.12| 0.75|
| BSD100    | 4 | 25.96   | 0.66  | 23.34| 0.67| 25.66| 0.71| 25.56| 0.71| 25.08| 0.69| 25.86| 0.71|
| DIV2K     | 4 | 26.66   | 0.85  | 23.80| 0.75| 28.07| 0.81| 27.87| 0.80| 27.20| 0.78| 28.35| 0.81|

3.2. Ablation Studies

In order to better explore the influence of SP convolution and shuffle attention mechanism on the model, we use Set14 as the test data set to perform ablation experiment. As shown in table 2, the first line is original RCAN, the second line indicates that only shuffle attention is used to replace the original CA module (RCAN+SA), the third line indicates that only SP convolution is used to replace the original traditional convolution (RCAN+SP), and the last line contains the above two replacement (RCAN+SA+SP). For experimental comparison, Params is the number of parameters of the entire model.
network, Tmadds is the total number of multiplication and addition operations, and GFlops means giga floating point operations per second. The statistical work of the above parameters is implemented by torchstat.

Figure 3. Comparison of 4x reconstruction of “female” in Set 14 data set.

Figure 4. Comparison of 4x reconstruction of “butterfly” in Set5 data set.

Table 2. Comparison of ablation experiments.

| Method      | PSNR (↑) | SSIM (↑) | Params (↓) | TMAAdd (↓) | GFLOPs (↓) |
|-------------|----------|----------|------------|------------|------------|
| RCAN        | 25.74    | 0.75     | 15,592,379 | 1.9        | 950.1      |
| RCAN+SA     | 25.77    | 0.75     | 15,477,179 | 1.9        | 950.1      |
| RCAN+SP     | 26.04    | 0.75     | 15,569,083 | 1.89       | 948.72     |
| RCAN+SA+SP  | 26.12    | 0.75     | 15,453,883 | 1.89       | 948.72     |

3.3. Supplementary Experiment
Training for 30 epochs may not be enough. Therefore, the better networks RDN and RCAN are selected to further compare with RSAN for 100 epochs. The test data set is still Set14. The result is shown in Table 3.

Table 3. Complementary experiment comparison.

| Method | PSNR (↑) | SSIM (↑) | LPIPS (↓) |
|--------|----------|----------|-----------|
| RCAN   | 26.45    | 0.76     | 0.290     |
| RDN    | 26.26    | 0.76     | 0.299     |
| OUR    | 26.51    | 0.76     | 0.284     |

4. Conclusion
This paper proposes a super-resolution reconstruction networks RSAN that can not only obtain better subjective visual evaluation and objective quantitative evaluation, but also reduce the number of network parameters and improve the network efficiency to a certain extent. However, through many training
experiments, it is found that the networks proposed in this paper still performs too smoothly and does not perform well for subtle image details, such as hair. In order to solve this problem, The next stage of work is to introduce constraints from HR image to LR image to solve multiple solution spaces from LR image to SR image.

Acknowledgments
This work is supported by the Yunnan Ten-thousand Talents Program and Project of Nantong Science and Technology Bureau under grant JC2019108.

References
[1] Dong C, Loy C C, He K, et al. 2014 Learning a deep convolutional network for image super-resolution European Conference on Computer Vision pp 184-199.
[2] Dong C, Loy C C and Tang X 2016 Accelerating the super-resolution convolutional neural network European Conference on Computer Vision pp 391-407.
[3] Lai W S, Huang J B, Ahuja N, et al. 2017 Deep Laplacian pyramid networks for fast and accurate super-resolution Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 624-632.
[4] Xie S and Tu Z 2015 Holistically-nested edge detection Proceedings of the IEEE International Conference on Computer Vision pp 1395-1403.
[5] Lim B, Son S, Kim H, et al. 2017 Enhanced deep residual networks for single image super-resolution Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 136-144.
[6] Ledig C, Theis L, Huszár F, et al. 2017 Photo-realistic single image super-resolution using a generative adversarial network Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4681-4690.
[7] Vaswani A, Shazeer N, Parmar N, et al. 2017 Attention is all you need Advances in Neural Information Processing Systems 5998-6008.
[8] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 7132-7141.
[9] Woo S, Park J, Lee J Y, et al. 2018 CBAM: Convolutional block attention module Proceedings of the European Conference on Computer Vision (ECCV) pp 3-19.
[10] Zhang Y, Li K, Li K, et al. 2018 Image super-resolution using very deep residual channel attention networks Proceedings of the European Conference on Computer Vision (ECCV) pp 286-301.
[11] Dai T, Cai J, Zhang Y, et al. 2019 Second-order attention network for single image super-resolution Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 11065-11074.
[12] Zhang Q L and Yang Y B 2021 Sa-net: Shuffle attention for deep convolutional neural networks ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE) pp 2235-2239.
[13] Zhang Q, Jiang Z, Lu Q, et al. 2020 Split to be slim: An overlooked redundancy in vanilla convolution arXiv preprint arXiv:2006.12085.
[14] Shi W, Caballero J, Huszár F, et al. 2016 Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1874-1883.
[15] Wang Q, Wu B, Zhu P, et al. 2020 ECA-Net: Efficient channel attention for deep convolutional neural networks Conference on Computer Vision and Pattern Recognition (CVPR).
[16] Wu Y and He K 2018 Group normalization Proceedings of the European Conference on Computer Vision (ECCV) pp 3-19.
[17] Zhang Y, Tian Y, Kong Y, et al. 2018 Residual dense network for image super-resolution Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 2472-2481.