Efficient Channel Attention Feature Fusion for Lightweight Single Image Super Resolution

Lingxiu Jiang and Yue Zhou*  
Shanghai Jiao Tong University, School of Electronic Information and Electrical Engineering, Shanghai, China  
*Email: surej0824@sjtu.edu.cn

Abstract. Recent advances in deep learning and convolution neural network have greatly improved the reconstruction performance of SISR compared with the traditional methods. However, complicated models and huge amount of parameters limit the application of those methods in real-world scenes. In our paper, we propose an efficient channel attention feature fusion method on the lightweight super-resolution network (ELSRN) for SISR. We reduce our network parameters through several modules, including binary cascading feature fusion. Besides, we propose to build efficient inverted residual block (EIRB) and stack several EIRBs to capture effective feature information of different scales. Last, we fuse multi-scale features in pairs step by step and finally refine final feature information with different scale features. Several experiments have proved that our EIRBs module and binary cascading method are effective and our network can achieve a great trade-off between reconstruction performance and model size.

1. Introduction
Single image super-resolution (SISR) is a classical low-level task of computer vision and aims to generate a super-resolution (SR) image from the corresponding low-resolution (LR) image. SISR is also an ill-posed problem, since even the identical LR image may result totally different SR images. Recent advances in deep learning and convolution neural network have greatly improved the reconstruction performance of SISR compared with the traditional methods.

Dong [1] was the first to implement SISR with convolution neural network (CNN) to establish an end-to-end model between LR and HR images. Kim [2] thought that the deeper neural network could work better and result higher reconstruction quality of SR images. However, gradient explosion, gradient vanishing and difficulty in training procedure became more and more serious as the neural network depth increased. EDSR [3] modified the residual network to be more suitable for the task of SISR, based on the very deep neural network. Those above large models with huge amount of parameters seriously limit the application in real-world scenes or mobile devices.

In order to reduce the model parameters, DRCN [4] proposed to increase the recursion depth without introducing more parameters. And MemNet [5] employed recursive units and gate units to build long-term memory and choose useful information flowing to the deeper layers. However, these methods reduced the model parameters with the price of computational operations. The burden of computational operations also plays an important role in practical application scenarios. Therefore, during the recent years, more and more lightweight CNNs, like Xception [6] and MobileNetV2 [7], are designed for the mobile devices and resources constrained environment.
Figure 1. Performance comparison between the network and other lightweight models on x2 Set5.

Beside above, channel attention mechanism has also attracted plenty of attention in the recent days. LFFN [8] utilized general channel attention mechanism and proposed a lightweight feature fusion network based on Xception [6].

Inspired from the above problems and works, we propose an efficient channel attention feature fusion based on the lightweight super-resolution network (ELSRN) for SISR. The main contributions of our work consist of three parts:

1) First, we reduce our network parameters through several ways, including efficient inverted residual blocks (EIRBs), efficient channel attention mechanism and binary cascading feature fusion.

2) Second, to extract enough feature information with limited parameters, we propose to stack several EIRBs to capture feature information of different scales.

3) Last, we fuse multi-scale features in pairs step by step and finally refine final feature information with different scale features.

Figure 1 shows our network performs state-of-the-art (SOTA) on Set5 with the scale factor of 2 among several lightweight SR methods. Several experiments have proved that our network can achieve a great trade-off between reconstruction performance and model size.

2. Related Work

2.1. Lightweight CNNs

During the recent years, to reduce model size, lightweight CNNs are stepping on the stage. The main content of MobileNetV1 [9] is to translate the general convolution into depthwise convolution and pointwise convolution for the limited resources of mobile devices and embedded applications. At the same time, ShuffleNet [10] utilized group convolution and channel shuffle to greatly reduce the parameters and computation flops while maintaining the considerable performance on the tasks like image classification and object detection. Depthwise convolution is the special group convolution while setting the number of groups equal to the number of input channels. MobileNetV2 [11] proposed the inverted residual blocks (IRBs) with linear operation in the high dimension instead of non-linear one to further improve the feature representation. MobileNetV2 [11] is one of the most outstanding lightweight CNNs on the tasks of image classification and object detection.

2.2. Channel Attention Mechanism

The aim of channel attention mechanism in CNNs is to consider the interrelationship among channels. Hu [12] proposed SENet to adaptively pay more attention to channels with more informative feature instead of treating every channel equally in the same layer. Zhang [13] proposed to integrate the channel attention with residual learning SR network and achieved the SOTA performance of SISR.
Different from the general channel attention above, ECANet [14] proposed to utilize 1D convolution on the channel dimension, named efficient channel attention (ECA) to replace the general channel attention in SENet [12].

3. Proposed Method
In this section, we firstly describe the whole architecture of our network. Second, we elaborate two key modules in our method separately.

![Network Architecture of ELSRN](image)

Figure 2. Network architecture of our proposed ELSRN.

3.1. Network Architecture
As shown in Figure 2, the proposed efficient channel attention feature fusion method with lightweight super-resolution network (ELSRN) consists of feature extraction module, efficient feature mapping module stacked with several efficient inverted residual blocks (EIRBs), binary cascading based multi-scale efficient feature fusion module (MEFFM) and reconstruction module.

First, we use a 3x3 convolution layer with 64 filters followed by BN layer and Relu activation layer to extract shallow feature information from LR image. This can be formulated as

\[ x_0 = F_{cat}(I_{lr}) \] (1)

\( I_{lr} \) represents the input LR image, \( F_{cat} \) is the function of shallow feature extraction module and \( x_0 \) is the output of shallow feature extraction module. Efficient feature mapping module consists of several stacked EIRBs.

\[ x_n = F_{eirb}^n(F_{eirb}^{n-1}(...F_{eirb}^1(x_0))) \] (2)

\( x_n \) is the output of the \( n \)th EIRB and \( F_{eirb}^i \) represents the function of \( i \)th EIRB. In order to fuse multi-scale features with few parameters, we design to cascade every two outputs of EIRBs to fuse features two by two. For all cascading outputs, our model continues to cascade two of them until only one output is left. \( F_{cat} \) represents the binary cascading function.

\[ M_j = F_{cat}(x_j, x_{j+1}) \] (3)

\[ M_{j+1} = F_{cat}(x_{j+2}, x_{j+3}) \] (4)

\[ x_n = F_{cat}(M_j, M_{j+1}) \] (5)
After the last binary cascading, the output is $x_n$. Then, the final output of binary cascading is fed into the multi-scale efficient feature fusion module.

$$x = F_{meffm}(x_n)$$ \tag{6}

$F_{meffm}$ denotes the function of the multi-scale efficient feature fusion module and $x$ represents its output. The reconstruction module works after global residual learning.

$$I_{sr} = F_{up}(x_n + x)$$ \tag{7}

$F_{up}$ denotes the operation of the reconstruction module, which consists of 1x1 convolution and sub-pixel convolution layer. The whole architecture of our proposed network is presented in Figure 2.

3.2. **Efficient Inverted Residual Block (EIRB)**

We think, although the depthwise separable convolution used in the IRBs of MobileNetV2 reduces the parameters and computation operations, it ignores the interaction relationship between channels. As this may decrease the reconstruction ability of network. So we think it is necessary to integrate channel attention mechanism to compensate for it. Different from general channel attention used in RCAN [13], we combine efficient channel attention mechanism with inverted residual block, named EIRB, to enhance the correlation between channels without introducing more extra parameters.

![Efficient Inverted Residual Block](image)

**Figure 3.** Right is the composition of EIRB and left is the description of efficient channel attention (ECA) layer.

ECANet [14] and MobinetNetv2 [11] are both designed for those high-level computer vision tasks like image classification, we reduce the number of channels and use local residual learning strategy in our EIRBs for SISR. Depthwise separable convolution and efficient channel attention complement with each other from channel dimension. After several EIRBs, important feature information is refined step by step. Every output of EIRB also represents feature information with different receptive field. The details of EIRB are shown on the Figure 3.

3.3. **Multi-scale Efficient Feature Fusion Module (MEFFM)**

To make full use of effective feature information from EIRBs with fewer parameters, we design the binary cascading method to fuse every two outputs of EIRBs locally. And for all cascading outputs, we continue to cascade every two of them until only the last output is left. Then, the final output of binary cascading is fed into the multi-scale efficient attention feature fusion module. Different from general cascading operation, our binary cascading operations further reduce the parameters and make full use of multi-scale features at the same time. Besides, the binary cascading gradually refines features step by step to extract more important feature information. The details of MEFFM are shown on the Figure 4.
Figure 4. Left describes the multi-scale effective feature fusion module and right represents the detailed composition of multi-scale feature fusion.

4. Experiments and Analysis

4.1. Datasets
During our training procedure, we choose the DIV2K dataset [15] as our training set. DIV2K is one of the largest datasets for super-resolution and image restoration tasks. There are 800 high resolution images in different scenes for training set and 100 images for validation. The corresponding LR images come from bicubic interpolation in Matlab. During testing stage, we evaluate our model on the several standard benchmark datasets: Set5 [16], Set14 [17], BSD100 [18] and Urban100 [19]. We calculate the PSNR and SSIM [20] on the Y channel for better comparing with other methods.

4.2. Implementation Details
We set the batch size as 16 and patch size 48x48 from the LR images as input. We set 64 feature channels for the whole network and totally stacked 16 EIRBs. Our model was trained by ADAM optimizer with L1 loss. The initial learning rate was set to $10^{-4}$ and decreased to half every 50 epochs.

4.3. Ablation Study

4.3.1. Numbers of Stacked EIRBs
To investigate the number of EIRBs in our network, we set {8, 16, 24, 32} for 4 stacked EIRBs groups. We compared them from their parameters, multi-adds and PSNR on the x2 Set5. The compared results are shown in Table 1 and we can find that 8 stacked EIRBs achieve poor performance with minimal parameters. 16 stacked EIRBs perform best among 4 experiment groups. In the last two experiment groups with 24 and 32 EIRBs, parameters and multi-adds significantly increase, but PSNR results hardly improve. Hence, our proposed network stacks 16 EIRBs.

|          | 8   | 16  | 24   | 32   |
|----------|-----|-----|------|------|
| Params   | 769.0K | 1193K | 1616K | 2040.8K |

Table 1. Comparison of PSNR on the scale $\times$ 2 of Set5 among different number of EIRBs.
4.3.2. Binary Cascading
To evaluate the effectiveness of the binary cascading method, we replaced the binary cascading with direct cascading, which directly cascaded all outputs of EIRBs after the last EIRB and fused features directly. We compared the parameters and multi-adds of two models based on different cascading methods. The compared results are shown in Table 2 and we can find that binary cascade method is of few parameters with better reconstruction ability.

Table 2. Comparison of PSNR on the scale $\times$ 2 of Set5 among different cascading methods.  
D represents directly cascading method and B represents ours.

|                  | EIRBs+D | EIRBs+B(Ours) |
|------------------|---------|---------------|
| Params           | 2473K   | 1193K         |
| Multi-adds       | 551.31G | 292.7G        |
| PSNR             | 38.14   | 38.20         |

4.4. Comparisons with the State-of-the-art
For quantitative comparisons, we compared our reconstruction results in PSNR and SSIM against several lightweight CNNs for SISR: VDSR [21], DRCN [4], DRRN [22], MemNet [5], CARN [23], AWSRN [5] in the latest years in Table 3. From Table 3, our proposed network has less multi-adds but better performance compared with CARN and AWSRN. In Figure 5, we presented the visual comparisons between our methods with several others. Obviously, our proposed ELSRN achieved better visual results than others.

Table 3. Comparison with several lightweight SR models.

| Datasets | Scale | DRRN PSNR/SSIM | MemNet PSNR/SSIM | CARN PSNR/SSIM | AWSRN PSNR/SSIM | ELSRN(Ours) PSNR/SSIM |
|----------|-------|----------------|------------------|----------------|-----------------|-----------------------|
| Set5     | x2    | 37.74/0.9591   | 37.78/0.9597     | 37.76/0.9590   | 38.11/0.9608    | 38.20/0.9615       |
|          | x3    | 34.03/0.9244   | 34.09/0.9248     | 34.29/0.9255   | 34.52/0.9281    | 34.63/0.9306       |
|          | x4    | 31.68/0.8888   | 31.74/0.8896     | 32.13/0.8937   | 32.27/0.8960    | 32.40/0.8992       |
| Set14    | x2    | 33.23/0.9136   | 33.28/0.9142     | 33.52/0.9166   | 33.78/0.9189    | 33.90/0.9212       |
|          | x3    | 29.96/0.8349   | 30.00/0.8350     | 30.29/0.8407   | 30.38/0.8426    | 30.51/0.8441       |
|          | x4    | 28.21/0.7720   | 28.26/0.7723     | 28.60/0.7806   | 28.69/0.7843    | 28.91/0.7864       |
| B100     | x2    | 32.05/0.8973   | 32.08/0.8978     | 32.09/0.8978   | 32.26/0.9006    | 32.40/0.9039       |
|          | x3    | 28.95/0.8004   | 28.96/0.8001     | 29.06/0.8034   | 29.16/0.8069    | 29.28/0.8098       |
|          | x4    | 27.38/0.7284   | 27.40/0.7281     | 27.58/0.7349   | 27.64/0.7385    | 27.75/0.7398       |
| Urban100 | x2    | 31.23/0.9188   | 31.31/0.9195     | 31.92/0.9256   | 32.49/0.9316    | 32.80/0.9335       |
|          | x3    | 27.53/0.8378   | 27.56/0.8376     | 28.06/0.8493   | 28.42/0.8580    | 28.38/0.8522       |
|          | x4    | 25.44/0.7638   | 25.50/0.7630     | 26.07/0.7837   | 26.29/0.7930    | 26.56/0.7955       |
| Params   | -     | 297K           | 677K             | 1.592K         | 1.397K          | 1.193K              |
| Multi-Adds | -   | 6796.9G        | 2662.4G          | 222.8G         | 320.5G          | 292.7G              |
5. Conclusions
In this paper, we propose an efficient, effective and lightweight SR network with fewer parameters and multi-adds. We construct the efficient inverted residual blocks with efficient channel attention to extract hierarchical features and cascade those features two by two for further feature fusion. In the end, multi-scale feature fusion provides better feature information of great importance for better reconstruction ability. Several experiments have witnessed that our proposed network can achieve excellent trade-off between performance and model parameters and flops.

References
[1] Chao D., Chen C.L., He K., Tang X. 2014 Learning a deep convolutional network for image super-resolution European Conference on Computer Vision
[2] Lim B., Son S., Kim H., Nah S., Lee K.M. 2017 Enhanced deep residual networks for single image super-resolution in: CVPRW
[3] Martin D., Fowlkes C.C., Tal D., Malik J 2001 A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics Tech. Rep. UCB/CSD-01-1133 (EECS Department, University of California, Berkeley)
[4] Kim J, Kwon Lee J, and Mu Lee K 2016 Deeply-recursive convolutional network for image super-resolution Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1637-45
[5] Wang C, Li Z, Shi J 2019 Lightweight image super-resolution with adaptive weighted learning network. arXiv: Computer Vision and Pattern Recognition
[6] Howard A., Zhu M., Chen B., Kalenichenko D., Wang, W., Weyand, T., Andreetto, M., Adam, H. 2017 Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: Computer Vision and Pattern Recognition
[7] Hu, J., Shen, L., Sun, G. 2018 Squeeze-and-excitation networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 7132-41
[8] Zeyde, R., Elad, M., Protter, M. 2010 On single image scale-up using sparse representations International Conference on Curves and Surfaces pp 711-730
[9] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. 2017 Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: Computer Vision and Pattern Recognition
[10] Zhang, X., Zhou, X., Lin, M., Sun, J. 2017 Shufflenet: An extremely efficient convolutional neural network for mobile devices. arXiv: Computer Vision and Pattern Recognition
[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. 2018 Mobilenetv2: Inverted residuals and linear bottlenecks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4510-20
[12] Hu, J., Shen, L., Sun, G. 2018 Squeeze-and-excitation networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 7132-41

[13] Zhang Y, Li K, Li K, Wang L, Zhong B, and Fu Y 2018 Image super-resolution using very deep residual channel attention networks In Proceedings of the 15th European Conference, (Munich, Germany, 8–14 September 2018) pp. 294–310

[14] Wang Q L, Wu B G, et al 2019 Eca-net: Efficient channel attention for deep convolutional neural networks. arXiv:1910.03151

[15] Agustsson E, Ntire T R 2017 Challenge on Single Image Super-Resolution: Dataset and Study Computer Vision and Pattern Recognition 2017 1122-31

[16] Bevilacqua M, Roumy A, Guillemot C, and Alberimcore M L 2012 Low-complexity single-image super-resolution based on nonnegative neighbor embedding British Machine Vision Conference pp 1-10

[17] Zeyde, R., Elad, M., Protter, M. 2010 On single image scale-up using sparserepresentations (Curves & Surfaces, Avignon-France, June 24-30, 2010) pp 711-730

[18] Martin D, Fowlkes C C, Tal D, Malik J 2001 A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics Tech. Rep. UCB/CSD-01-1133 (EECS Department, University of California, Berkeley)

[19] Huang J, Singh A, Ahuja N 2015 Single image super-resolution from transformed self-exemplars IEEE Conference on Computer Vision and Pattern Recognition pp 5197-206

[20] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity IEEE Transactions on Image Processing 13(4), 600-12

[21] Kim J, Lee J K, Lee K M 2016 Accurate image super-resolution using very deep convolutional networks CVPR 2016 (Seoul National University) Oral Paper

[22] Tai Y, Yang J, Liu X 2017 Image super-resolution via deep recursive residual network Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition pp 3147-3155

[23] Ahn N, Kang B, and Sohn K 2018 Fast, accurate, and lightweight super-resolution with cascading residual network European Conference on Computer Vision (Springer, Cham) pp 256-72