Image Super-Resolution via Multi-scale Channel Attention Residual Network

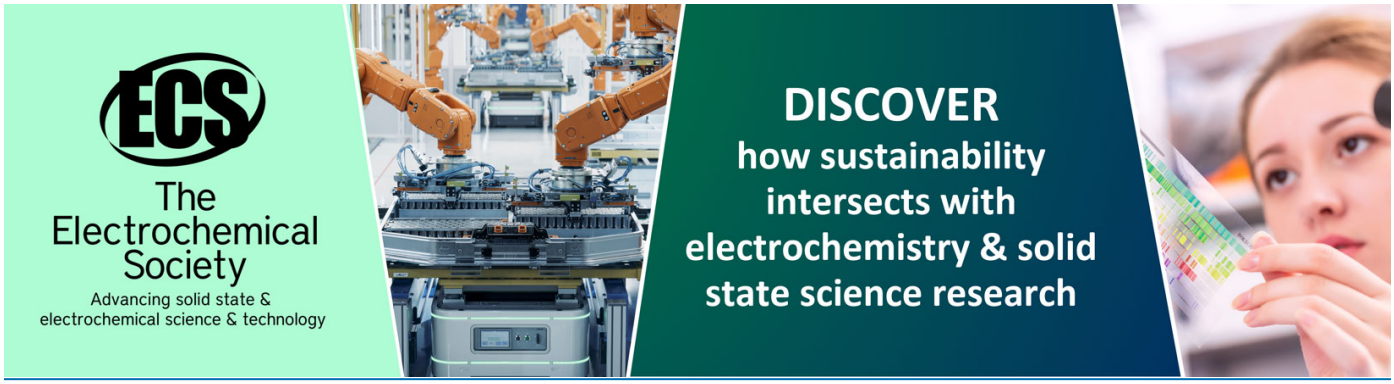


Image Super-Resolution via Multi-scale Channel Attention Residual Network

Caidong Yang[1], Fangwei Sun[1], Chengyang Li[1,2], Heng Zhou[1,3], Ziwei Du[1], Zhongbo Li[1*] and Yongqiang Xie[1*]

[1] Institute of Systems Engineering, Academy of Military Sciences, Beijing 100141, China

[2] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

[3] School of Electronic Engineering, Xidian University, Xi'an 710071, China

[*]Corresponding author's e-mail: lzb05296@163.com yqxie2021@163.com

Abstract. The refinement of existing CNN-based Super-resolution Reconstruction (SR) networks mainly focuses on deeper network architecture, which hinders the transmission of information in the networks. A deeper network is unable to make full use of intermediate correlation features and makes the training of the network difficult. To solve these problems, we propose a multi-scale channel attention residual network (MCAR). Specifically, we propose a multi-scale channel attention fusion module (MCAF) to learn local and global channels feature and capture the long-range dependencies. Furthermore, the multi-scale block is adopted to get the different scale feature representations. The experimental results on four benchmark datasets demonstrate that our models can effectively improve the visual effect of images, and outperform most of the advanced SISR methods in PSNR and SSIM.

1. Introduction

Low-resolution (LR) images are blurred and detailed information is lost, so the SR technology is to complete and perfect this part of the information to obtain high-resolution (HR) images. The generated images can also provide services for downstream computer vision-related tasks[1][2], which can enhance the task effect and improve recognition accuracy.

The traditional SR methods have achieved great success because of their interpretability and ease to accomplish. However, these methods need more and more artificially defined prior knowledge with the zoom scale increasing and no longer meeting the needs. It's difficult to achieve the purpose of high-quality reconstruction.

The CNNs have power feature representation ability to learn the mapping relationship. The earliest CNN-based SISR is the SRCNN[3] proposed by Dong et al.in 2014, obtaining a higher PSNR/SSIM index than the traditional method. In VDSR[4] which has 20 layers, Kim et al. used residual blocks to connect networks to increase the depth. In ResNet[5] , Kaiming He had proved that increasing the depth of the network can fully release the potential of the network and enhance its learning ability.

However, Most existing deeper and deeper CNN-based SR methods have the following problems:(1) simply stacking the residual block to increase the depth will increase the training difficulty and hardly

obtain the better improvements; (2) ignoring the relevant information between the middle layers; (3) gradually weakening the transmission of information flow.

To address these problems, we proposed a MCAR model to exploit the different scale feature from the input image and obtain a better feature representation ability. The MCAF module is proposed for the MCAR model. Firstly, MCAF extracts different scale features by different convolution kernels. Secondly, at different scales, we utilize local and global attention branches respectively to learn channel-wise correlational features. Lastly, we fuse the two branches which have different scale features to reconstruct high-resolution images via a sub-pixel convolution upscaling module.

2. Related work

2.1. CNNs for SR

The main improvement direction of the CNN-based algorithm model is the depth of the convolutional layer. In theory, the deeper network can capture more higher-lever features and provide better expression capabilities. In 2015, Kim et al. [4] used residual blocks to increase the network depth, which solved the problem of slow convergence speed, and improved network performance. In 2018, Zhang Y et al boiled a deep network using residual mapping and achieved good results in high-frequency information processing of images.

2.2. Attention mechanism

Processing the entire image needs too many resources, and increases inference time. The human visual system often does not pay attention to every detail when quickly observing the outside world, but selectively to the areas that are interested in and considered important and ignores some unimportant areas. Based on this idea, J. Hu et al. [6] proposed a SENet model with different degrees of attention to different channel features by learning different weights using the obtained weights value to enhance useful features and suppress useless features.

2.3. Residual network

The essence of the SISR is to generate the texture details from the low-resolution image information. So, researchers believe that the learning ability of the networks should be focused on the residual part between the LR and HR images. Then the network can avoid the complete transformation. The global residual reduces the amount of model calculation. The local residual is ResNet[5] proposed by K. He et al., which is mainly used to solve the model degradation caused by the deep network. The local residual achieves by adding skip connections inside the network.

3. Proposed method

3.1. Network architecture

We will describe and formulate the network we proposed. As shown in Figure 1, the MCAR has three parts, including LR image feature extraction module, MCAF for nonlinear feature mapping learning, and the reconstruction part.

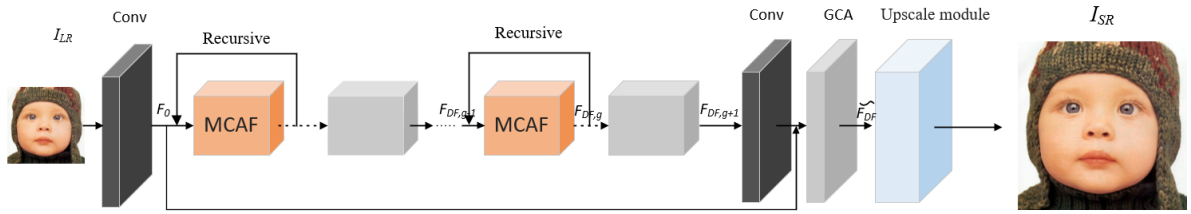


Figure 1. The network architecture of MCAR.

The input is I_{LR} and output is I_{SR}, our network uses the 9 × 9 convolutional kernel to extract shallow features F_0:

$$F_0 = H_{EF}(I_{LR}) \tag{1}$$

Where $H_{EF}(\cdot)$ denotes the feature extraction function, and F_0 is the extracted shallow feature. Then F_0 is used for the MCAF module to learn deep features:

$$F_{DF} = H_{MCAF}(F_0) \tag{2}$$

where $H_{MCAF}(\cdot)$ denotes the functional function of our proposed MCAF structure, and F_{DF} is the deep feature. MCAF can learn deep features well through global channel attention and local channel attention. Then learn the F_{DF} channel attention through the globe channel attention module (GCA):

$$\widetilde{F_{DF}} = H_{scale}(F_{DF}) \tag{3}$$

Where $H_{scale}(\cdot)$ denotes the globe channel attention module (GCA), and $\widetilde{F_{DF}}$ is the learned deep feature. Then input $\widetilde{F_{DF}}$ to the upsampling module for size enlargement:

$$F_{\uparrow} = H_{\uparrow}(\widetilde{F_{DF}}) \tag{4}$$

Where $H_{\uparrow}(\cdot)$ is the up-sampling function, and $F_{\uparrow}$ is larger size feature. Finally, input $F_{\uparrow}$ into the reconstruction module to complete the final image reconstruction:

$$I_{SR} = H_{rec}(F_{\uparrow}) = H_{MCAR}(I_{LR}) \tag{5}$$

where $H_{rec}(\cdot)$ denotes the reconstruction module, $H_{MARFN}(\cdot)$ denotes our proposed MCAR network, and I_{SR} is the final generated image.

MSE can improve PSNR and SSIM very well, but it will cause the generated image to be smooth and blurred visual effects. Experiments have proved that the L1 loss function is better than that of MSE, and the reconstruction effect is more realistic. For simplicity, we choose L1 loss to optimize the proposed model and then use MSE for the loss fine-tuning.

$$L_{MSE}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left\|H_{MCAR}(I_{LR}^{i}) - I_{HR}^{i}\right\|^2 \tag{6}$$

$$L_{L1}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left\|H_{MCAR}(I_{LR}^{i}) - I_{HR}^{i}\right\|_1 \tag{7}$$

3.2. MCAF module

The convolutional neural network extracts information features by fusing information[7] that satisfies a certain spatial and channel distribution in the local receptive field. We utilize convolutional learning to focus on the relationships between different channels. As shown in Figure 2, the model can select the channel features that are effective and contribute more to the reconstruction and suppress the other channel features.

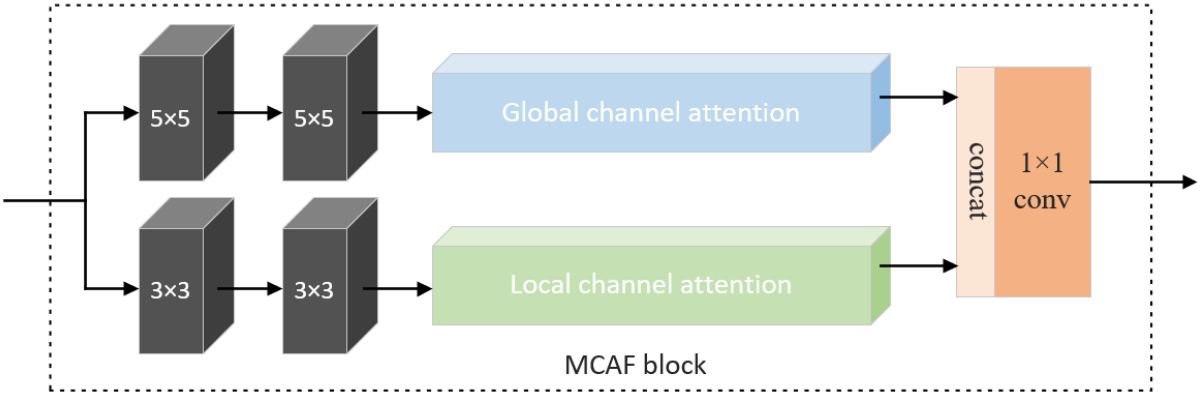


Figure 2. Multi-scale channel attention fusion module.

3.2.1. Global channel attention branch. We choose a 5×5 convolution to extract image information. As shown at the top of Figure 3, this branch first compresses the features of the entire spatial dimension into (C,1,1) through the global average pooling operation and then uses two one-dimensional

convolutions to complete the learning of the weights of each channel. Last, multiply the weight value with each original feature channel.

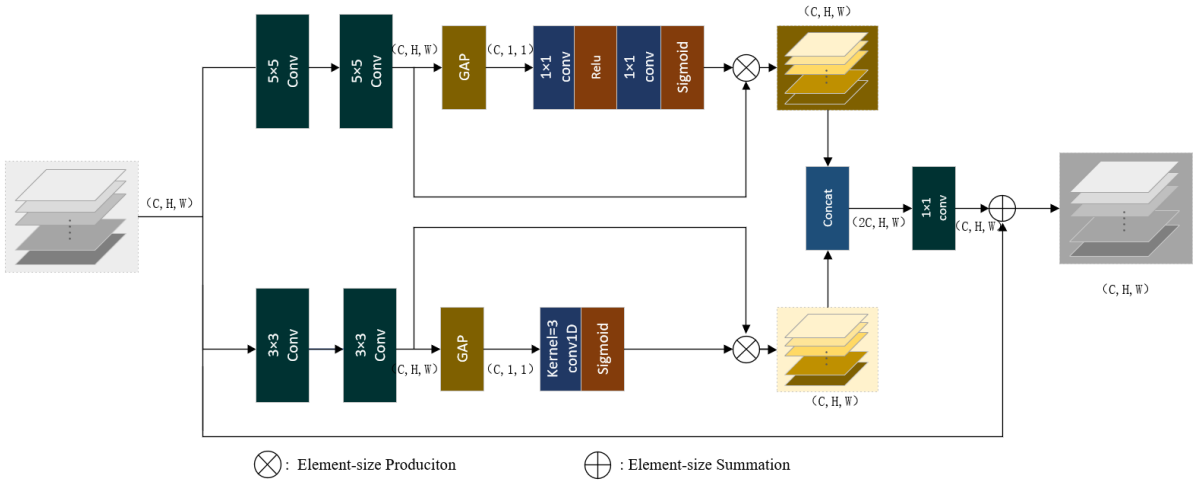


Figure 3. Global and local channel attention feature learning module.

3.2.2. Local channel attention branch. The model obtains a larger receptive field on a shallower network by the globally operating and then captures long-range dependencies in information. But for SISR, more information is more beneficial to image reconstruction. Some information that is not important from a global perspective can also be important on the reconstruction. Therefore, to compensate for the problem brought by global channel attention, we used the local channel attention[7] to enhance the effect of local channel features on image reconstruction. Because only a small number of parameters is added, no dimensionality reduction operation is required, which avoids the problem of information loss during the dimensionality reduction process.

As shown at the bottom of Figure 3, the GAP is first performed to compress the features of the entire spatial dimension into 1×1×C size, and then one-dimensional convolution which the kernel is 3 is used to complete the local cross-channel interaction, and extract the local inter-channel interaction dependencies. Finally, we obtain the weights between the local channels by the sigmoid function.

3.3. Multi-scale model

Multi-scale features can contain richer image features. We adopt a parallel structure and use convolution blocks with kernel sizes of 3×3 and 5×5 to extract features respectively in the same layer. The larger the convolution kernel can extract more features. Therefore, we use a 5×5 convolution kernel in the global channel attention branch to extract information and a 3×3 convolution kernel in the local channel attention branch.

3.4. Feature fusion block and reconstruction module

After obtaining the global and local channel attention features, we perform a concat operation on the two pieces of information, and finally, use 1×1 convolution to fuse the two features. In addition to the 1×1 convolution, there are also methods such as pixel value-weighted average, pixel value average[8], etc. It is proved by comparative experiments that 1×1 convolution works best.

The reconstruction module uses the feature learned from the previous layers to reconstruct different scale images. Compared with the deconvolution layer and nearest neighbour, sub-pixel convolution[9] has superior performance and a better reconstruction effect. So, we choose sub-pixel convolution as our reconstruction mothed.

4. Experiments

In this section, we present the test results of our proposed model on four benchmark datasets.

4.1. Settings

Datasets. We use the DIV2K dataset as training set. The details of the image in DIV2K dataset are clear and thus very suitable for use in super-resolution training. Meanwhile, we use four benchmark datasets as domain Validation set: Set5, Set14, BSD100, and Urban100, which widely used for model performance evaluation in SR.

Evaluation metrics. We used the PSNR and SSIM as the evaluation metrics. PSNR is generally defined by Mean Square Error (MSE). The MSE is defined as in equation (8):

$$MSE = \frac{1}{WH}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}[X(i,j) - Y(i,j)]^2 \tag{8}$$

Where W and H is the wide and high of the image, the X is the generated image and Y is the original image. So PSNR is based on the error between corresponding pixels.

$$PSNR = 10\,lg\left(\frac{X_{MAX}^2}{MSE}\right) \tag{9}$$

SSIM is defined by bright-ness $l(x,y)$, contrast $c(x,y)$, and structure $s(x,y)$:

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1} \\ c(x,y) &= \frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2} \\ s(x,y) &= \frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \end{aligned} \tag{10}$$

Here, x and y is the generated and the original image successively, c is the constant, μ is the mean and the σ is the variance, σ_{xy} is the Covariance. We set $\alpha = \beta = \gamma = 1$, $c_3 = c_2/2$, and get the simplified SSIM:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y+c_1)\{(\sigma)_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{11}$$

Implementation details. During the training process, we expand the DIV2K dataset by randomly cropping, rotating and flipping the images horizontally and vertically to increase training data samples. One sub-image is randomly cropped from the original image, and then down-sampled to 48×48, 24×24 and 16×16 according to different scale alignments. 64 images are selected as a batch for each training. Feature extraction is performed on LR images using a 9×9 convolution kernel, followed by multi-scale channel features learning with 16 MCAF modules. The convolution of the global branch is 5 and the local branch is 3. The model is trained with the ADAM optimizer with β_1=0.9, β_2=0.99, ε=10^{-8}. The learning rate is initialized to 10^{-4} and reduced to half when training to epoch=400. We implement on NVIDIA GeForce RTX 2080Ti GPU using the Pytorch framework.

4.2. Effects of local and global channel attention branch

To demonstrate the effect of the local branch and global branch in the MCAF module, we successively removed. Else, to verify the impact of large-scale feature extraction on global, we adopted convolution kernel sizes of 3 and 5. We set the number of MCAFs to 8 and trained to 300 epochs. Experiments show that when only the local channel attention branch is used, the PSNR index of Set5(2×) is 37.26 dB, which is relatively the lowest. Using only global channel attention, when the convolution kernel is 3 and 5 respectively, it reaches 37.35 dB and 37.71dB, which are relatively increased, and the effect of the convolution kernel is 5 is better. It can be seen that large-scale convolution is efficient for global channel attention. Finally, the experiment of combining local and global is carried out, and the indicators are the best, 37.78 dB. These comparisons illustrate the effectiveness of local and global channel attention in our proposed MCAF for image super-score reconstruction.

Table 1. Investigations of MCAF module on Set5(2×).

MCAF	Local branch (3×3)	✓	×	×	✓	✓
	Global branch (3×3)	×	✓	×	✓	×
	Global branch (5×5)	×	×	✓	×	✓
PSNR on Set5（2×）		37.26	37.35	37.71	37.75	37.78

4.3. Benchmark results

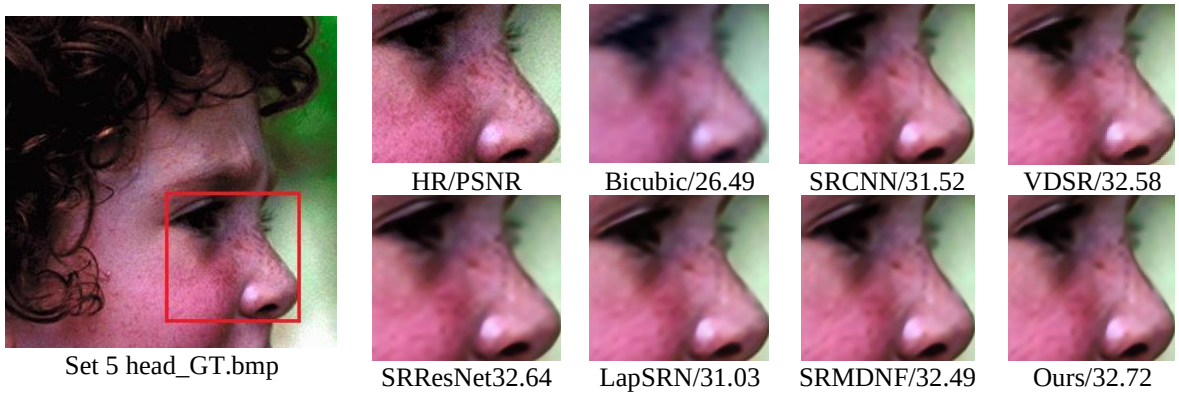
In table 2, we provide the results of quantitative evaluation of our model on benchmark datasets in the super-resolution domain. Our model is compared with the SOAT model, including SRCNN, VDSR, LapSRN, SRResNet and SRMDNF, etc. For the existing models, we use the authors published results. For comparison, we transform the generated image to YCbCr space and then compute PSNR and SSIM metrics. Although the PSNR on the BSD100 dataset in the ×4 case is slightly lower than SRMDNF by 0.06dB, the other datasets achieve the best results. The SSIM of our method on the four benchmark datasets mostly exceeds other SR models.

Table 2. Average PSNR/SSIMs of MCAR for Scale ×2 and ×4.

Dataset	scale	bicubic	SRCNN	VDSR	LapSRN	SRResNet	SRMDNF	Ours
Set5	×2	33.66/0.929	36.66/0.954	37.53/0.958	37.52/0.959	-	37.79/0 .960	**37.79/0.964**
	×4	28.42/0.810	30.48/0.863	31.35/0.873	31.54/0.885	32.05/0.891	31.96/0.893	**32.21/0.904**
Set14	×2	30.24/0.869	32.45/0.907	33.05/0.911	33.08/0.913	-	33.32/0 .916	**33.46/0.924**
	×4	26.00/0.703	27.50/0.751	28.02/0.763	28.19/0.772	28.53/0.780	28.35/0.777	**28.61/0.798**
BSD100	×2	29.56/ 0.843	31.36/0.888	31.90/0.896	31.80/0.895	-	**32.05**/0 .899	31.99/**0.907**
	×4	25.96/0.668	26.90/0.710	27.29/0.713	27.32/0.728	27.57/0.735	27.49/0.734	**27.63/0.754**
Urban100	×2	26.88/0.840	29.50/0.895	30.77/0.914	30.41/0.910	-	31.33/ 0.920	**31.59/0.928**
	×4	23.14/0.658	24.52/0.722	25.18/0.754	25.21/0.756	26.07/0.784	25.68/0.773	**26.15/0.800**

4.4. visualization results

We also present partial visualization results in the case of scale 4 in Figure 5. Through the comparison, our model can reconstruct more image detail information, and the contours are more obvious, which shows that our model is effective the effectiveness.



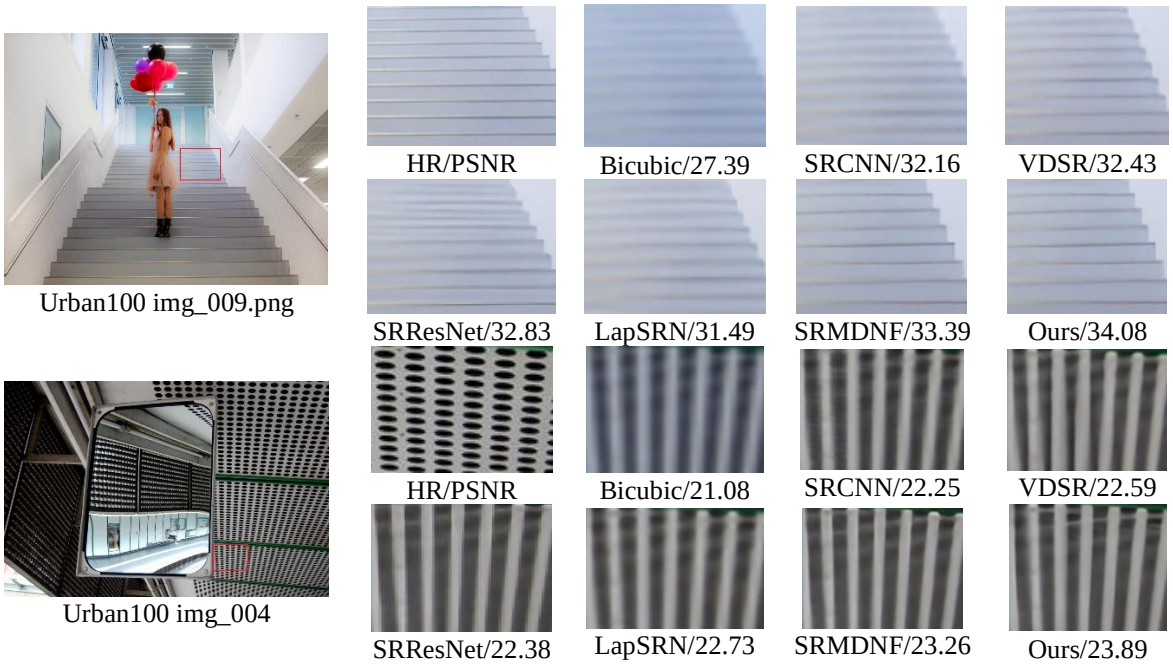


Figure 4. Visual comparison for 4× SR with our model on Set5 and Urban100 datasets.

5. Conclusions

We propose MCAR networks for high-precision image SR. Specifically, based on the basic residual structure, we utilize the global and local channel attention structures to improve the network learning efficiency. We utilize convolution of different sizes to make full use of the original image information. The channel attention learning is carried out Based inside the MCAF module, and the residual connection is used to allow low-frequency information direct skipping outside the MCAF module, making the network more focused on the learning of high-frequency information. Finally, the fusion of different feature maps is achieved with few parameters through 1×1 convolution. The future work is to focus on optimizing the network structure and improving the reconstruction effect of the model on texture details. In addition, the reconstruction speed and light-weight of the model can also be studied.

References

[1] Sajjadi M, Scholkopf B, Hirsch M, et al. Enhancenet: Single Image Super-Resolution Through Automated Texture Synthesis[C].2017 IEEE International Conference on Computer Vision (ICCV), 2017, 4501-4510.

[2] Dai D, Wang Y, Chen Y, et al. Is image super-resolution helpful for other vision tasks? [C]. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE,2016:1-9.

[3] Chao D, Chen C, He K, et al. Learning a Deep Convolutional Network for Image Super-resolution[C]. European conference on computer vision. Springer, Cham, 2014: 184-199.

[4] Kim J, Lee J K, Lee K M. 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] 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.

[6] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[7] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C].2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).2020:11531-11539
[8] Qin J, Huang Y, Wen W. Multi-scale feature fusion residual network for single image super-resolution[J]. Neurocomputing, 2020, 379: 334-342.
[9] 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.