Front. Comput. Sci., 2023, 17(1): 171307
https://doi.org/10.1007/s11704-021-0562-y

RESEARCH ARTICLE

SR-AFU: super-resolution network using adaptive frequency component upsampling and multi-resolution features

Ke-Jia CHEN[1,2], Mingyu WU (✉)[3], Yibo ZHANG[3], Zhiwei CHEN[3]

1 School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
2 Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
3 College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract Image super-resolution (SR) is one of the classic computer vision tasks. This paper proposes a super-resolution network based on adaptive frequency component upsampling, named SR-AFU. The network is composed of multiple cascaded dilated convolution residual blocks (CDCRB) to extract multi-resolution features representing image semantics, and multiple multi-size convolutional upsampling blocks (MCUB) to adaptively upsample different frequency components using CDCRB features. The paper also defines a new loss function based on the discrete wavelet transform, making the reconstructed SR images closer to human perception. Experiments on the benchmark datasets show that SR-AFU has higher peak signal to noise ratio (PSNR), significantly faster training speed and more realistic visual effects compared with the existing methods.

Keywords super-resolution, multi-resolution features, adaptive frequency upsampling, wavelet transformation

1 Introduction

Image super-resolution (SR) [1] is a classical problem in computer vision research, which refers to the reconstruction of the corresponding high-resolution (HR) images from the available low-resolution (LR) images. The higher the resolution is, the more details the image can provide. HR image data is of great importance to the image related applications. For example, HR medical images help doctors make the correct diagnosis [2,3]; HR satellite images (HRSI) can help to easily distinguish similar objects [4,5]; With HR images, the performance of pattern recognition can be greatly improved [6]. However, most digital images are currently captured by image sensors such as charge-coupled devices (CCD) or complementary metal oxide semiconductors (CMOS), with the resolution not sufficient to meet the needs for consumer applications and scientific research. Therefore, it is essential to conduct research on image resolution enhancement.

The traditional SR methods are primarily based on interpolation [7], sparse representation (Dictionary Learning) [8], neighbor embedding [9], etc. The quality of HR images obtained by these methods is usually unsatisfactory as they use only the information contained in the LR images, not being able to reconstruct high-frequency details. In recent years, significant breakthrough has been made in SR research with the development of deep convolutional neural networks. Examples include SRCNN [10] proposed by Dong et al., VDSR [11] and DRCN [12] proposed by Kim et al., RDN [13] based on ResNet, and RCAN [14] using attention mechanism.

However, there are still some disadvantages in the methods mentioned above. Firstly, few models take the correlation between distant pixels (i.e., pixel context) into account, which is very valuable for SR reconstruction. For example, if the model can recognize that the LR image depicts small objects and complex textures, it can be inferred that its corresponding HR image contains lots of high-frequency components, so the model can adjust the upsampling strategy to favor high-frequency data. Secondly, most previous methods add only one upsampling layer as the last layer of the network. All the information required for upsampling is only obtained from the feature map of the last layer, while the information of other layers cannot be fully utilized. Thirdly, existing methods often use MAE or MSE to define loss functions. Although the reconstructed SR image could get higher peak signal to noise ratio (PSNR), its visual effect is not always satisfactory. Oppositely, the perceptual loss used in some methods [15] can achieve realistic visual effects despite the relatively low PSNR of the reconstructed images. Finally, methods like RCAN use a second-order module to achieve higher PSNR. However, they often require a larger video memory capacity and have lower prediction speed. These defects are more prominent when the size of predicted images is large. This makes the methods inconvenient for practical use.

To solve the above problems, this paper proposes an adaptive frequency components upsampling (AFU) model

Received November 25, 2020; accepted July 15, 2021

E-mail: wmylxmj@gmail.com

based on the deep parallel dual-network structure for SR image reconstruction. The entire network contains two sub-networks: cascaded dilated convolution residual network (CDCRN) and multi-size convolutional upsampling network (MCUN). The CDCRN is cascaded with 32 cascaded dilated convolution residual blocks (CDCRB) and so is the MCUN that is cascaded with 32 multi-size convolutional upsampling blocks (MCUB). Each CDCRB can perceive the multi-resolution features of the image and extract image semantics based on the pixel context information. Note that in order to reduce the training cost and make the model lighter, we avoid the use of second-order modules, even if doing so may cause a loss of PSNR. The output of each CDCRB is the input of each corresponding MCUB. Each MCUB upsamples an image which has the same size with the final SR image, i.e., (h, w, 3), by sub-pixel convolution, and allocates a coefficient to the image. This image is a component of the SR image, which contains only part of frequency band. The deeper the MCUN layer is, the higher frequency the upsampled image contains. The final SR image is the weighted sum of these components output by each MCUB. Such an upsampling structure can automatically adjust the training rate of each block. Since the low frequency and medium frequency components in the SR image can be quickly learned and reconstructed by shallow blocks, the overall training speed is greatly accelerated. In order to balance PSNR and visual effects, a new loss function based on wavelet transform is defined. Using high-order wavelet decomposition, the function can calculate the errors of different frequency components in the local space of the image, so that the reconstructed image is closer to human perception.

Overall, this paper has three main contributions:

- A cascaded dilated convolution residual block (CDCRB) is proposed to provide additional receptive fields, which is connected to the upsampling block to form a parallel dual-network structure;
- A multi-size frequency component upsampling block (MCUB) is proposed to make full use of the features from each CDCRB and the training rate of each MCUB can be automatically adjusted to accelerate the entire training speed;
- A new loss function based on high-order wavelet decomposition is defined, which makes the reconstructed SR images closer to human perception.

The rest of the paper is organized as follows: Section 2 introduces the related work in the SR community. Section 3 details the SR-AFU method proposed in this paper. Experiments in Section 4 not only verify the effectiveness of the AFU module and the wavelet decomposition-based loss function but also compare the performance of SR-AFU and other related models including several state-of-art methods using benchmark datasets. Section 5 concludes the paper.

2 Related work

In the study of single image super-resolution (SISR), traditional interpolation-based or reconstruction-based methods [16] have been gradually replaced by CNN-based single image super-resolution (SISR) methods [17] due to their excellent performance. Dong et al. [10] are the first to introduce CNN to image SR task and proposed a super-resolution convolutional neural network (SRCNN). The method uses bicubic interpolation to enlarge the low-resolution image to the target size, and uses a three-layer convolutional network to fit a non-linear map to achieve good results. Later, Dong et al. [18] further propose a fast super-resolution convolutional neural network (FSRCNN), adding a deconvolution layer at the end of the CNN for upsampling, so that the original low-resolution images can be directly input to the network.

Note that ResNet [19] can alleviate the training difficulties while improving learning performance when the network is deep. Kim et al. proposed a very deep convolutional network (VDSR) [11] and a deeply-recursive convolutional network (DRCN) [12] to further improve SRCNN. The two residual network-based methods not only accelerate the convergence speed, but also avoid the problem of gradient disappearance or explosion. Lai et al. [20] proposed Laplacian pyramid super-resolution network (LapSRN), which takes coarse-resolution feature maps as input to predict high-frequency residuals and then uses transposed convolution to upsample them to a finer level.

Although the above methods achieve a high PSNR, the reconstructed SR images tend to be smooth due to the use of single pixel loss without processing high-frequency image details. Ledig et al. [21] proposed to use generative adversarial network (GAN) in image super-resolution task (i.e., SRGAN) to reconstruct realistic textures from a large number of down-sampled images. Subsequently, Wang et al. [22] proposed an enhanced super-resolution generative adversarial network (ESRGAN) to address the issue of hallucinated details. The reconstructed images have more realistic natural textures. However, the PSNR of both SRGAN and ESRGAN are not satisfactory. To further improve the PSNR, Lim et al. [23] proposed an enhanced deep super-resolution network (EDSR), which removes the redundant modules of SRResNet to increase the model depth and achieves better PSNR. Yu et al. [24] used wide activation for efficient and accurate image super-resolution (WDSR), which further improves the structure of EDSR. Specifically, WDSR removes a lot of redundant convolutional layers to reduce training parameters and enlarges the feature map before the ReLU activation function in the residual module, so it can improve the accuracy of super-resolution while reducing the training time.

However, the above proposed methods mainly focus on improving the network structure, ignoring the existence of redundant low-frequency information in the extracted image features and thus treating high-and low-frequency information equally. To solve this problem, Zhang et al. [14] introduced the attention mechanism to the SR task and proposed a channel attention (CA) mechanism that takes into account the correlation between feature channels to adaptively rescale features. The attention mechanism allows the model to reconstruct more details and textures. Based on RCAN, Dai et al. [25] further proposed a second-order attention network (SAN) for more powerful feature expression and feature correlation learning. A trainable second-order channel attention (SOCA) module is developed to adaptively rescale the channel-wise

features by using second-order feature statistics for more discriminative representations.

After reviewing the existing SR methods, we believe that the following important issues remain unsolved. Firstly, there exists a lot of unnecessary training in previous models, which leads to slow training convergence. They reconstructed all frequency components of SR images simultaneously through the same upsampling layer. Since low-frequency components can be easily reconstructed without deep structures, a better method is to reconstruct different frequency details of the image through different upsampling layers. Secondly, most SR methods do not learn the multi-resolution features of the image, so they cannot focus on the global and local areas of the image at the same time. Thirdly, many SR methods with high PSNR use L1 or L2 loss in the spatial domain, making it difficult to reconstruct high-frequency details of the image and thereby making the reconstructed SR image too smooth. Fourthly, although the RCAN or SAN methods can obtain higher PSNR, they use a second-order module, which has the problems of large number of model parameters, slow prediction speed, and high memory capacity, making them difficult to adapt to real-world application scenarios.

3 The proposed method

An image can be seen as the result of superimposing with different frequency components [25]. The low-frequency information can be effectively reconstructed only with the information of surrounding pixels, but the reconstruction of high-frequency information usually requires long-distance correlation between pixels. Therefore, a larger receptive field and a deeper convolutional network are required to extract multi-resolution features, and a more differentiated upsampling method is needed to reconstruct different frequency components.

This paper proposes a Super-Resolution network based on Adaptive Frequency component Upsampling (SR-AFU). The method uses cascaded dilated convolutional residual blocks (CDCRB) to extract multi-resolution features and multi-size convolutional upsampling blocks (MCUB) for SR image reconstruction.

The entire framework (shown in Fig. 1) is mainly composed of two networks, named as cascaded dilated convolutional residual network (CDCRN) and multi-size convolutional upsampling network (MCUN), respectively. CDCRN is cascaded with multiple CDCRBs, which can exponentially increase the receptive field for understanding the image semantics and extracting potential features. MCUN is cascaded with MCUBs, which can upsample different frequency components to reconstruct the SR image. Note that the output of each CDCRB is also part of the input of the corresponding MCUB.

3.1 Cascaded dilated convolution residual block

After down-sampling, it is difficult even for human eyes to recognize the content of the image area far away from the shooting point. Therefore, it is necessary to use the surrounding pixels to assist in determining the image semantics. Pooling is a commonly used technique, but it reduces the resolution and makes it difficult to reconstruct small objects. Instead, dilated convolution [7] supports exponential expansion of the receptive field without reducing resolution or coverage. Therefore, dilated convolution is used in SR-AFU to establish the correlation between a wide range of pixels. It is also the first time that dilated convolution is introduced into the SR problem.

Specifically, we use cascaded convolutions with dilation rates of 1, 2, 4, and 8 in each CDCRB and concatenates feature maps under different receptive fields (3×3, 7×7, 15×15, and 31×31) (see Fig. 2). Thus, each CDCRB has a maximum receptive field of 31×31 and can obtain multi-resolution features of images.

Here, the receptive field is defined as:

$$f_l = (f_{l-1} - 1) + (d * (k - 1) + 1), \tag{1}$$

where f_l is the size of the receptive field of the lth layer, d is the dilation rate, and k is the size of the convolution kernel.

Correspondingly, the resolution is defined as:

$$R_i = R_{i-1} + 2p - k - (k - 1) * (d - 1) + 1, \tag{2}$$

where R_i is the resolution of the ith feature map and p is the padding size.

To sum up, a large receptive field can be quickly obtained to extract the contextual information with the help of cascaded dilated convolution and the model can obtain more diverse features by generating feature maps with different resolutions.

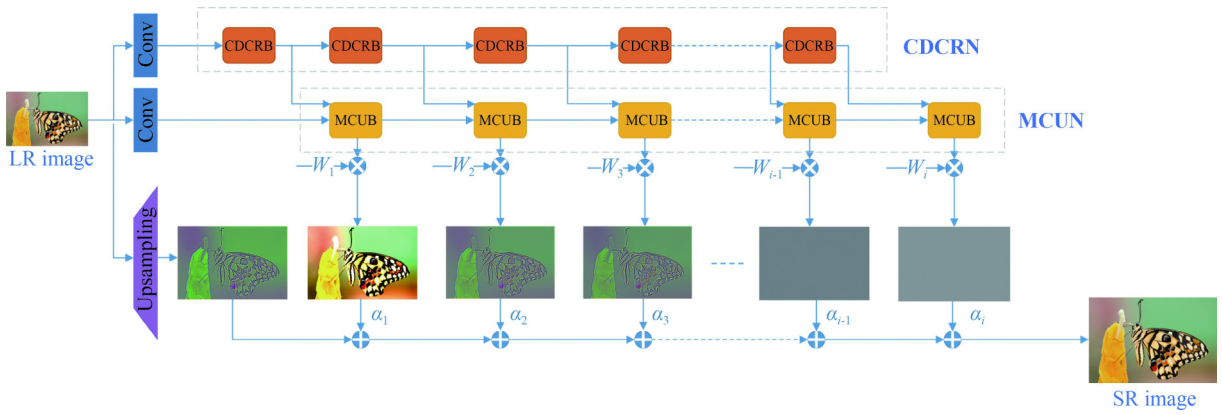

Fig. 1 The framework of SR-AFU

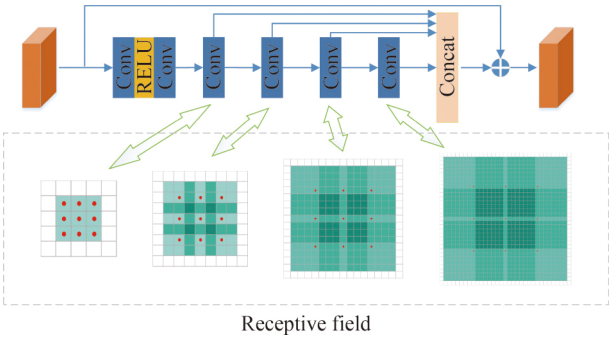

Fig. 2 The structure of CDCRB

3.2 Multi-size convolutional upsampling block

The previous deep networks were designed to reconstruct high-frequency information, but they spent a lot of unnecessary calculations to reconstruct low-frequency information. Actually, the latter can be easily reconstructed using shallow networks.

To solve this problem, this paper proposes a multi-size upsampling method (shown in Fig. 1). Here, MCUBs with different depths upsample different frequency components in the image details.

In each MCUB (see Fig. 3), the input features come from the corresponding CDCRB, then multi-scale convolution (rather than ordinary convolution) is used to extract adaptive multi-scale features, and finally the corresponding frequency component is upsampled.

The output of each MCUB is multiplied by a weighting coefficient a_i (a learnable parameter) and added to generate the final SR image by Eq. (3):

$$x = \sum_{i=1}^{N} a_i \varphi_i, \tag{3}$$

where x is the reconstructed SR image, φ_i is the component reconstructed by the ith MCUB and a_i is its corresponding weight coefficient. The coefficient a_i is learned by

$$a_i = \sum_{i=1}^{N} < x, \hat{\varphi}_i > \varphi_i, \tag{4}$$

where $< \cdot >$ represents inner product. φ and $\hat{\varphi}$ are bi-

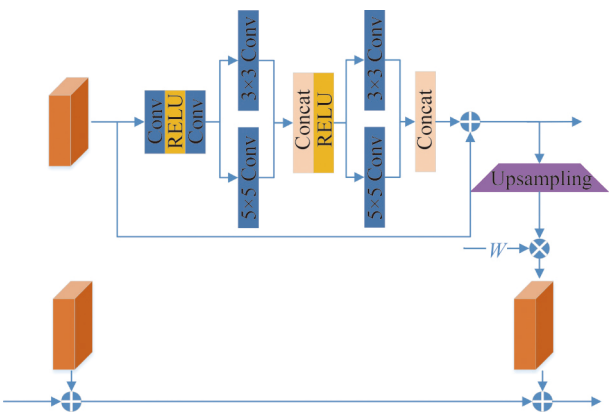

Fig. 3 The structure of MCUB

orthogonal, that is

$$< \varphi_i, \hat{\varphi}_j >= \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \tag{5}$$

and $< x, \hat{\varphi}_i >$ is the similarity between the signal x and $\hat{\varphi}_i$. The smaller the difference between x and $\hat{\varphi}_i$, the larger the value of a_i.

Any image can be regarded as a discrete two-dimensional signal $x_{(H,W)}$. After training SR-AFU, the two-dimensional tensors $\varphi_{1(H,W)}, \varphi_{2(H,W)}, \ldots, \varphi_{n(H,W)}$ of the image and the corresponding weight coefficients $a_1, a_2, \ldots, a_n$ are output and reconstruct the SR image:

$$Y_{SR(H,W)} = \sum_{i=0}^{N} a_i \varphi_{i(H,W)}. \tag{6}$$

The upsampling method proposed in this paper has the following advantages: Firstly, the low and medium frequency components can be reconstructed by the shallow blocks, thereby eliminating the need for redundant calculations in deep blocks and reducing training time. Secondly, deep blocks can focus more on the reconstruction of high-frequency details of the image. Thirdly, the network can adaptively learn the weight of each upsampling result and adjust the learning rate of each MCUB, so the overall training speed of the network is greatly improved.

The upsampling process in this paper is visualized in Fig. 4. Here, the number before the hyphen refers to the image label in the DIV2K dataset ×2 and the number i after the hyphen indicates that the image was generated by the ith MCUB of the network. The figure shows that the shallow blocks of MCUN upsample the low-frequency information, such as large smooth background areas, while the deep blocks reconstruct the high-frequency details of the image, such as contour and texture details. In short, different frequencies components of the image are reconstructed by blocks of different depths.

3.3 Wavelet-based loss function

Most previous SR methods use L1 loss (MAE) or L2 loss (MSE) as loss function [26]. However, both of them only calculate the errors between individual pixels, without considering the multi-resolution features. As a result, the reconstructed SR images are often too smooth.

This paper defines a new wavelet-based loss function, which is the weighted sum of the average absolute error of the two images after high-order wavelet decomposition in YUV space and the average absolute error of the two images in the spatial domain.

Wavelet transform can analyze the local frequency of the space-time domain and can gradually refined the signal through zooming and translation operation. The error value of the high-order wavelet transform reflects the distortion of the light and dark in the SR image, which can effectively prevent the over-smooth visual effect.

The loss function of SR-AFU is defined in Eq. (7), aiming to better reflect the multi-resolution texture and structural features of the image.

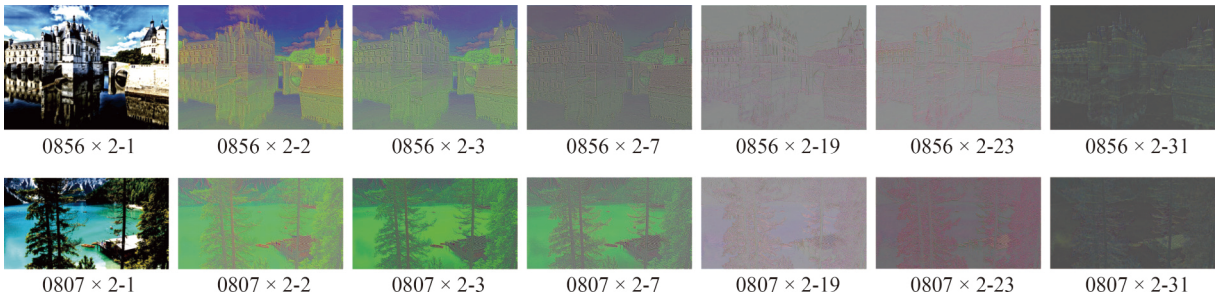

Fig. 4 The visualization of upsampling process in SR-AFU

$$\alpha * MAE((W_{\psi}^{i}, W_{\psi}^{i\,\prime})) + \beta * MAE(f, f'),\ i = \{H, V, D\}. \quad (7)$$

The first term in Eq. (7) represents the average absolute error of the high-order wavelet decomposition of two images in YUV space. The second term is used to calculate the average absolute error of two images. α and β are hyperparameters.

Equations (8) and (9) are used to decompose low-frequency components and high-frequency components from the image, respectively.

$$W_{\varphi}(0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\varphi_{0,m,n}(x,y), \quad (8)$$

$$W_{\psi}^{i}(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\psi_{j,m,n}^{i}(x,y),\ i = \{H, V, D\}, \quad (9)$$

Here, f is the image signal, j is the wavelet order and H, V, D represent horizontal, vertical and diagonal directions. Different wavelet decomposition operators are used (Eqs. (10)−(13)).

$$\varphi(x,y) = \varphi(x)\varphi(y), \quad (10)$$

$$\psi^{H}(x,y) = \psi(x)\varphi(y), \quad (11)$$

$$\psi^{V}(x,y) = \varphi(x)\psi(y), \quad (12)$$

$$\psi^{D}(x,y) = \psi(x)\psi(y), \quad (13)$$

where $\varphi(\cdot)$ and $\psi(\cdot)$ refer to the scaling function constructed by a low-pass filter and the wavelet function constructed by a high-pass filter, respectively, $\varphi(x,y)$ is the low-frequency component of the image in both horizontal and vertical directions, $\psi^{H}(x,y)$ represents the high-frequency component of the image in the horizontal direction and low-frequency component in the vertical direction, $\psi^{V}(x,y)$ represents the low-frequency component of the image in the horizontal direction and the high-frequency component in the vertical direction, and $\psi^{D}(x,y)$ is the high-frequency component of the image in both the horizontal and vertical directions.

4 Experiments

4.1 Datasets and settings

Following the previous methods [23,13], we use the DIV2K dataset [27] for training and Set5 [28], Set14 [29], BSDS100 [30], and Urban100 [31] datasets for testing. The 800 images from DIV2K are used for training. Each training image is randomly split with the size of 128×128 (×2), 85×85 (×3) and 64×64 (×4), rotated by 90º, 180º and 270º, and flipped horizontally and vertically. The batch size is 16. For training, we use the Adam optimizer [32] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ and weighted normalization. The initial learning rate is set to 10^{-3} and the learning rate is halved at every 2×10^5 iterations.

The SR-AFU is implemented by TensorFlow in a NVIDIA GeForce RTX 2080Ti GPU. The number of blocks in SR-AFU is set to 32, and each CDCRB uses 32 filters (the number of convolution kernels in the block). The weight of each MCUB is initialized as 0.01. In our loss function, α is set to 2, β is set as 0.5, and sym4 wavelet basis is used. We conduct experiments with Bicubic (BI) degradation model [13].

The SR results of all methods are evaluated with PSNR, SSIM [33] and visualization.

4.2 Evaluation of AFU

In order to verify the effectiveness of AFU, we remove all the upsampling blocks (MCUB) except the last one in the SR-AFU model and then compare it with the original SR-AFU model. Figures 5(a) and 5(b) shows the comparison of PSNR and loss values of two models trained on the DIV2K ×4 dataset with a learning rate of 10^{-3}, respectively.

The experimental results show that using AFU, the convergence speed of the SR network is significantly improved, the training time is greatly shortened, and the PSNR of the model with AFU is still higher than that of the model without AFU. When iterating to about 17000 batches, the loss of the model without AFU surges and the PSNR drops from above 30 to almost 15 due to noise interference. In contrast, the PSNR of the model with AFU remains stable and the loss shows a steady decline. Even if the iteration reaches 48000 batches, the PSNR of the model with AFU is still 0.9 dB higher than that without AFU.

The main reasons for the above results are as follows. The ordinary residual structure simply adds the front-layer feature maps directly during initialization, and the data distribution of the back-layer feature maps is unstable, resulting in a slower gradient decline at the initial stage of training. The AFU method avoids this direct addition, but adjusts the ratio of the output of each layer to make the distribution of the output layer data relatively uniform, so the gradient declines faster in the initial training. As the training progresses, the network begins to converge, and the ratio adjustment of the output layer of SR-AFU also affects the gradient propagation, making the model more stable.

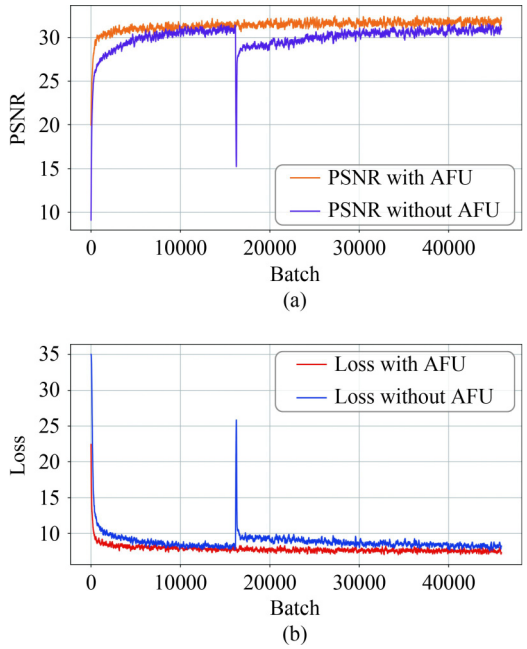

Fig. 5 Comparison of SR models with and without AFU trained on the DIV2K ×4 dataset with a learning rate of 1e−3. (a) PSNR; (b) loss

4.3 Evaluation of loss function based on wavelet transform

To evaluate the loss function defined in this paper, we compare the SR-AFU with the wavelet-based loss (SR-AFU$_{wav}$) and the SR-AFU with L1 loss (SR-AFU$_{L1}$) in visual effects.

Figure 6 shows two image examples from the DIV2K ×4 dataset. The leftmost in the figure is the original LR image, the middle is the local area of the SR image reconstructed by SR-AFU$_{L1}$, and the rightmost is the same local area of the SR image reconstructed by SR-AFU$_{wav}$.

By comparing the output images of the two SR-AFU models, it can be seen that the details of the SR images reconstructed by SR-AFU$_{L1}$ are relatively blurred, while the contrast of light and dark of the SR images reconstructed by SR-AFU$_{wav}$ is more obvious. The results indicate that the SR-AFU$_{wav}$ has more advantages in image visual effects.

4.4 Ablation study

In order to further explore the influence of the three factors of dilated convolution, AFU, and the loss based on wavelet transform (Loss$_{wav}$) on the SR-AFU method, we conducted a ablation study. We use the PSNR Set5 (2X) for comparison. The experimental results are shown in Table 1. The tick indicates that the proposed method is used, and the cross indicates that an alternative benchmark method is used, which are ordinary convolution, last layer upsampling and L1 loss, respectively.

The results in Table 1 show that when none of the three methods are used, the PSNR of the model is 38.12, which is equivalent to EDSR [23] (its PSNR is 38.11). When the three methods are used simultaneously (i.e., SR-AFU), the PSNR of the model is 38.27, which is the same as that of RCAN [14].

4.5 Results with bicubic (BI) degradation model

Experiment on the method comparison is based on the Bicubic (BI) degradation model, which is the most commonly used model.

● Comparison by PSNR/SSIM

We compare SR-AFU with ten state-of-the-art methods (Bicubic, SRCNN [34], FSRCNN [18], VSDR [11], LapSRN [20], MemNet [35], ESDR [23], DBPN [36], RDN [13], and SRFBN [37]) by PSNR and SSIM values.

All quantitative comparisons for SR with the scale ×2, ×3, and ×4 are reported in Table 2. The results of most previous methods are cited from their papers. The results show that SR-AFU outperforms most of the previous methods on all four datasets with three scaling factors, which verifies the effectiveness of our model in PSNR and SSIM.

● Comparison of SR-AFU and RCAN by PSNR, parameter size and prediction speed

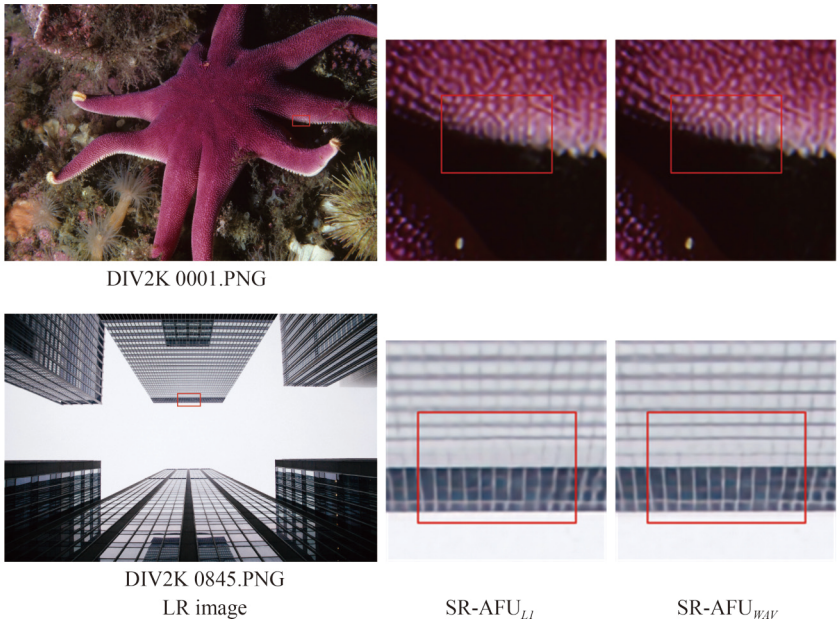

Fig. 6 Two images from DIV2K ×4 and their local area reconstructed by SR – AFU$_{L1}$ and SR – AFU$_{wav}$

Table 1 Investigations of Dilated Conv., AFU and $Loss_{wav}$. The best PSNR values on Set5 (2×) in 4×10^4 iterations are listed

Method	Dilated Conv.	AFU	$Loss_{wav}$	PSNR/dB
SR-AFU variants	×	×	×	38.12
	√	×	×	38.19
	×	√	×	38.21
	√	√	×	38.25
	×	×	√	38.15
	√	×	√	38.23
	×	√	√	38.24
	√	√	√	38.27

Table 2 Quantitative results with BI degradation model

Scale	Method	Set5		Set14		BSDS100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2	Bicubic	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403
	SRCNN [34]	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
	FSRCNN [18]	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020
	VDSR [11]	37.53	0.9590	33.05	0.9130	31.90	0.8960	30.77	0.9140
	LapSRN [20]	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9140
	MemNet [35]	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195
	EDSR [23]	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351
	DBPN [36]	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9326
	IMDN [38]	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283
	PAN [39]	38.00	0.9605	33.59	0.9181	32.18	0.8997	32.01	0.9273
	AWSRN [40]	38.11	0.9608	33.78	0.9189	32.26	0.9006	32.49	0.9316
	RDN [13]	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353
	SR-AFU (ours)	**38.27**	**0.9615**	**34.03**	**0.9215**	**32.34**	**0.9016**	**33.12**	**0.9361**
3	Bicubic	30.39	0.8682	27.55	0.7742	27.21	0.7386	24.46	0.7340
	SRCNN [34]	32.75	0.9090	29.30	0.8215	28.14	0.7863	26.24	0.7989
	FSRCNN [18]	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080
	VDSR [11]	33.67	0.9210	29.78	0.8320	28.82	0.7980	27.07	0.8280
	LapSRN [20]	33.82	0.9227	29.87	0.8350	28.96	0.8001	27.56	0.8376
	MemNet [35]	34.09	0.9248	30.01	0.8350	28.96	0.8001	27.56	0.8376
	EDSR [23]	34.65	0.9280	30.52	0.8462	28.97	0.8025	27.57	0.8398
	IMDN [38]	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519
	PAN [39]	34.40	0.9271	30.36	0.8423	29.11	0.8050	28.11	0.8511
	AWSRN [40]	34.52	0.9281	30.38	0.8426	29.16	0.8069	28.42	0.8580
	RDN [13]	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653
	SR-AFU (ours)	**34.74**	**0.9293**	**30.60**	**0.8471**	**29.28**	**0.8097**	**28.91**	**0.8665**
4	Bicubic	28.42	0.8103	26.00	0.7027	25.96	0.6676	23.14	0.6576
	SRCNN [34]	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221
	FSRCNN [18]	30.72	0.8660	27.61	0.7500	26.98	0.7150	24.62	0.7280
	VDSR [11]	31.35	0.8830	28.02	0.7680	27.29	0.7026	25.18	0.7540
	LapSRN [20]	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560
	MemNet [35]	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630
	EDSR [23]	32.46	0.8969	28.81	0.7875	27.71	0.7421	26.62	0.8033
	DBPN [36]	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946
	SRFBN-S	31.98	0.8920	28.45	0.7780	27.44	0.7310	25.71	0.7720
	SRFBN [37]	32.39	0.897	28.77	0.7860	27.68	0.740	26.47	0.798
	IMDN [38]	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838
	PAN [39]	32.13	0.8948	28.61	0.7822	27.59	0.7363	26.11	0.7854
	AWSRN [40]	32.27	0.8960	28.69	0.7843	27.64	0.7385	26.29	0.7930
	RDN [13]	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028
	SR-AFU (ours)	**32.47**	**0.8987**	**28.82**	**0.7879**	**27.73**	**0.7420**	**26.65**	**0.8042**

We further compare the PSNR (Urban100), the number of parameters and the average prediction speed of SR-AFU and RCAN [14] under different experimental settings (Table 3).

The PSNR of SR-AFU is slightly lower than RCAN because the latter uses the attention mechanism, which can successfully improve the PSNR. However, the channel attention used by RCAN is a second-order module, requiring the matrix multiplication after enlarging the image. It consumes a lot of calculation and memory. This will bring some difficulties to use RCAN in practical applications, for example, the prediction speed is too slow or the memory capacity cannot be met. In contrast, our model can be trained by only using a laptop with a GTX1050Ti video card.

Moreover, the parameter size of RCAN increases exponen-

Table 3 Comparison of parameters and prediction speed of SR-AFU and RCAN

Image shape	Index	SR-AFU	RCAN
(2×) (1,320,480,3)	Parameters	2,357,264	15,513,283
	Time/s	1.9074778	3.2828535
	PSNR	33.12	33.34
(3×) (1,160,240,3)	Parameters	2,497,124	15,882,563
	Time/s	0.5087615	1.096696
	PSNR	28.91	29.09
(4×) (1,80,120,3)	Parameters	2,692,928	16,399,555
	Time/s	0.1530186	0.4655488
	PSNR	26.65	26.82

tially (16M), but some parameters may actually be redundant. For example, the image recognition accuracy of GoogleLeNet with 20 million parameters is higher than that of VGG with 138 million parameters, many of which are redundant. The AFU method in this paper can remove unnecessary parameters in the model, making it run faster with a still high PSNR. Compared with EDSR (using 43M parameters), RDN (using 22.3M parameters) and RCAN (using 16M parameters), our model has less (2.6M) parameters.

● Comparison by visual effects

We finally compare SR-AFU with Bicubic, SRCNN [34], ESDR [23], DBPN [36], RCAN [14] and SAN [25] in visual effects.

Figure 7 shows visual comparisons on scale ×4 for two images in Urban100 dataset. It can be seen that our model is more sensitive to light and dark changes of images. The main reason is that SR-AFU uses the loss function based on the discrete wavelet transform, which allows the model to better process images with high frequencies in the horizontal and vertical directions. Moreover, the CDCRB allows the model to observe images from multiple receptive fields, thereby enabling SR-AFU to recover the structural features more effectively.

Taking the image "img_076" as an example, we observed that the results of Bicubic, SRCNN and EDSR lose details and produce blurring structures. The details of the images reconstructed by DBPN, RCAN and SAN are clearer than the previous three methods. They can recover most horizontal lines, but it is difficult to recover vertical lines well. SR-AFU can reconstruct details in different directions very well, and the reconstructed image has the best visual effect.

5 Conclusion

This paper proposes a super-resolution network based on adaptive frequency component upsampling (SR-AFU). Specifically, the cascaded dilated convolution residual block (CDCRB) can expand receptive field to understand image semantics. Meanwhile, the multi-size convolutional upsampling block (MCUB) adaptively upsample different frequency components, so that the deep network can focus more on high-frequency details. The loss based on wavelet transform allows to generate more realistic SR images. Comparative experi-

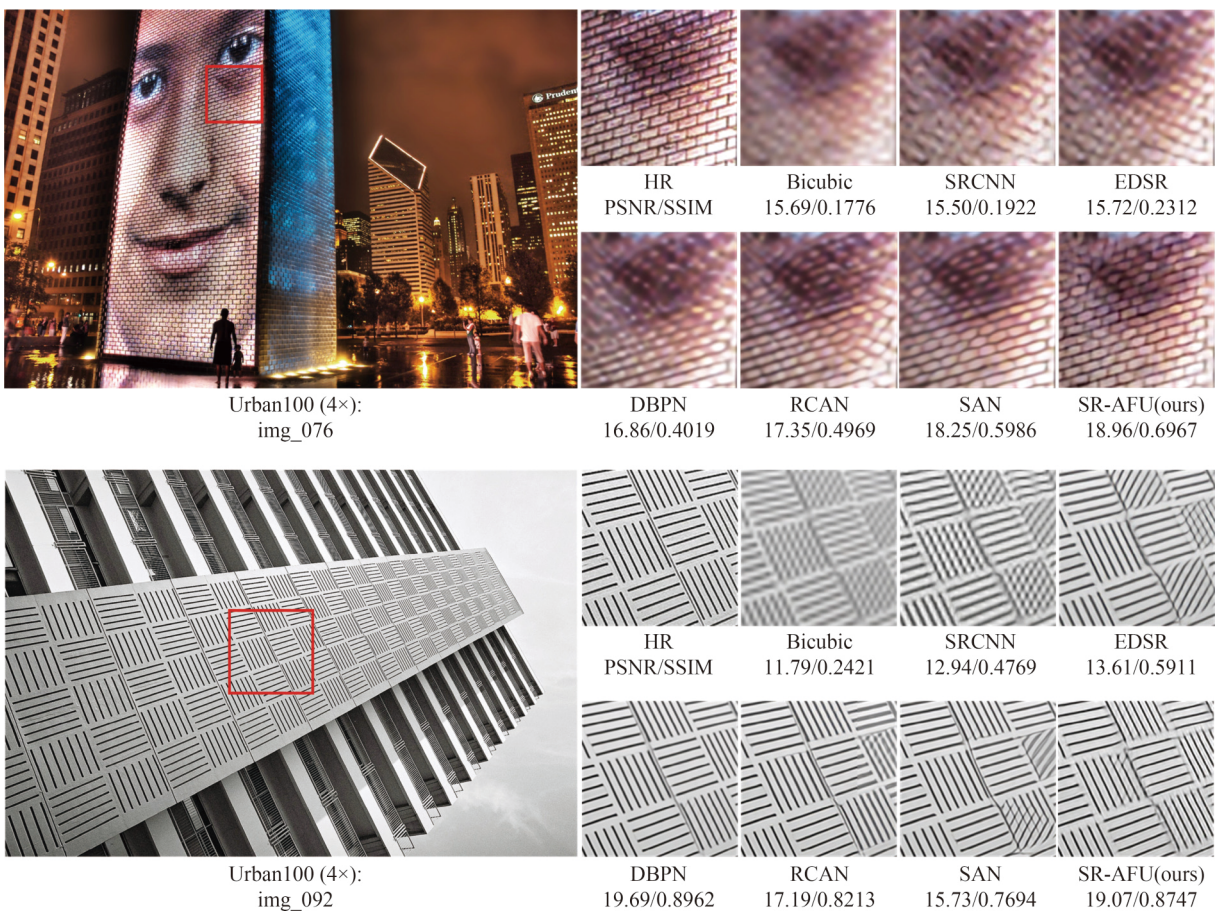

Fig. 7 Visual comparison for 4× SR with BI model on two datasets

ments show that the use of AFU can accelerate the convergence speed of training while effectively maintaining a high PSNR.

In the future, we will study the performance of AFU module in different network structures and on different image generation tasks. In addition, considering that the inverse process of the AFU module can retain both the shallow features and abstract semantics of the image, we will combine AFU and its inverse process to build a new image autoencoder for solving more difficult computer vision tasks.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant Nos. 61603197 and 61772284), Natural Science Foundation of Nanjing University of Posts and Telecommunications (NY221071).

References

1. Freeman W T, Pasztor E C, Carmichael O T. Learning low-level vision. International Journal of Computer Vision, 2000, 40(1): 25–47
2. Shi W, Caballero J, Ledig C, Zhuang X, Bai W, Bhatia K, de Marvao A M S M, Dawes T, O'Regan D, Rueckert D. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: Proceedings of the 16th International Conference on Medical Image Computing and Computer-Assisted Intervention. 2013, 9−16
3. Zhang S, Liang G, Pan S, Zheng L. A fast medical image super resolution method based on deep learning network. IEEE Access, 2018, 7: 12319–12327
4. Oh J, Lee C, Seo D C. Automated HRSI georegistration using orthoimage and SRTM: focusing KOMPSAT-2 imagery. Computers & Geosciences, 2013, 52: 77–84
5. Nogueira K, Penatti O A B, Dos Santos J A. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognition, 2017, 61: 539–556
6. Hu X, Ma P, Mai Z, Peng S, Yang Z, Wang L. Face hallucination from low quality images using definition-scalable inference. Pattern Recognition, 2019, 94: 110–121
7. Chen L C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the 15th European Conference on Computer Vision. 2018, 833−851
8. Li P, Wang Q, Zuo W, Zhang L. Log-Euclidean kernels for sparse representation and dictionary learning. In: Proceedings of IEEE International Conference on Computer Vision. 2013, 1601−1608
9. Lee Y, Choe Y. Neighbor embedding based single image super-resolution using hybrid feature and adaptive weight decay regularization. In: Proceedings of the 4th IEEE International Conference on Consumer Electronics Berlin. 2014, 185−187
10. Dong C, Loy C C, He K, Tang X. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295–307
11. Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1646−1654
12. Kim J, Lee J K, Lee K M. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1637−1645
13. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y. Residual dense network for image super-resolution. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 2472−2481
14. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the 15th European Conference on Computer Vision. 2018, 294−310
15. Rad M S, Bozorgtabar B, Marti U V, Basler M, Ekenel H K, Thiran J P. SROBB: targeted perceptual loss for single image super-resolution. In: Proceedings of IEEE/CVF International Conference on Computer Vision. 2019, 2710−2719
16. Ng M K, Shen H, Lam E Y, Zhang L. A total variation regularization based super-resolution reconstruction algorithm for digital video. EURASIP Journal on Advances in Signal Processing, 2007, 2007: 074585
17. Jiang K, Wang Z, Yi P, Jiang J. Hierarchical dense recursive network for image super-resolution. Pattern Recognition, 2020, 107: 107475
18. Dong C, Loy C C, Tang X. Accelerating the super-resolution convolutional neural network. In: Proceedings of the 14th European Conference on Computer Vision. 2016, 391−407
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016, 770−778
20. Lai W S, Huang J B, Ahuja N, Yang M H. Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017, 5835−5843
21. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017, 105−114
22. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Loy C C. ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018, 63−79
23. Lim B, Son S, Kim H, Nah S, Lee K M. Enhanced deep residual networks for single image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017, 1132−1140
24. Yu J, Fan Y, Yang J, Xu N, Wang Z, Wang X, Huang T. Wide activation for efficient and accurate image super-resolution. 2018, arXiv preprint arXiv: 1808.08718
25. Dai T, Cai J, Zhang Y, Xia S T, Zhang L. Second-order attention network for single image super-resolution. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 11057−11066
26. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for neural networks for image processing. 2018, arXiv preprint arXiv: 1511.08861
27. Timofte R, Agustsson E, Van Gool L, Yang M H, Zhang L, et al. NTIRE 2017 challenge on single image super-resolution: methods and results. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017, 1110−1121
28. Bevilacqua M, Roumy A, Guillemot C, Morel M L A. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of British Machine Vision Conference. 2012
29. Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. In: Proceedings of the 7th International Conference on Curves and Surfaces. 2010, 711−730
30. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th IEEE International Conference on Computer Vision. 2001, 416−423
31. Huang J B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2015, 5197−5206
32. Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations. 2015
33. Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612
34. Dong C, Loy C C, He K, Tang X. Learning a deep convolutional network for image super-resolution. In: Proceedings of the 13th

European Conference on Computer Vision. 2014, 184−199
35. Tai Y, Yang J, Liu X, Xu C. MemNet: a persistent memory network for image restoration. In: Proceedings of IEEE International Conference on Computer Vision. 2017, 4549−4557
36. Haris M, Shakhnarovich G, Ukita N. Deep back-projection networks for super-resolution. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 1664−1673
37. Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W. Feedback network for image super-resolution. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 3862−3871
38. Hui Z, Gao X, Yang Y, Wang X. Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th ACM International Conference on Multimedia. 2019, 2024−2032
39. Zhao H, Kong X, He J, Qiao Y, Dong C. Efficient image super-resolution using pixel attention. In: Proceedings of the European Conference on Computer Vision. 2020, 56−72
40. Wang C, Li Z, Shi J. Lightweight image super-resolution with adaptive weighted learning network. 2019, arXiv preprint arXiv: 1904.02358

Ke-Jia Chen is an associate professor in Nanjing University of Posts and Telecommunications, China. She received her PhD in Université de Technologie de Compiègne, France and her master's degree in Nanjing University, China. She joined Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, China in 2017. Her current research focuses on machine learning and its applications in complex network analysis.

Mingyu Wu received the BS degree in Electronic Information Engineering from Nanjing University of Posts and Telecommunications, China in 2021. He is working toward the master degree in Signal and information processing at Nanjing University of Posts and Telecommunications, China. His current research interests include casual inference, sequence modeling and cross-modal analysis.

Yibo Zhang received the BS degree in Electronic Information Engineering from Nanjing University of Posts and Telecommunications, China in 2021. His research interests include computer vision, UAV system and automatic driving.

Zhiwei Chen received the BS degree in Electronic Information Engineering from Nanjing University of Posts and Telecommunications, China in 2021. He is working toward the master degree in Electronic information at the South China Normal University, China. His current research interests include machine learning, computer vision and Neuromorphic computation.