doi:10.3772/j.issn.1006-6748.2022.02.010

Channel attention based wavelet cascaded network for image super-resolution①

CHEN Jian(陈 健)*, HUANG Detian②*, HUANG Weiqin**
(*College of Engineering, Huaqiao University, Quanzhou 362021, P.R.China)
(**School of Information Science and Technology, Xiamen University Tan Kah Kee College, Zhangzhou 363105, P.R.China)

Abstract

Convolutional neural networks (CNNs) have shown great potential for image super-resolution (SR). However, most existing CNNs only reconstruct images in the spatial domain, resulting in insufficient high-frequency details of reconstructed images. To address this issue, a channel attention based wavelet cascaded network for image super-resolution (CWSR) is proposed. Specifically, a second-order channel attention (SOCA) mechanism is incorporated into the network, and the covariance matrix normalization is utilized to explore interdependencies between channel-wise features. Then, to boost the quality of residual features, the non-local module is adopted to further improve the global information integration ability of the network. Finally, taking the image loss in the spatial and wavelet domains into account, a dual-constrained loss function is proposed to optimize the network. Experimental results illustrate that CWSR outperforms several state-of-the-art methods in terms of both visual quality and quantitative metrics.

Key words: image super-resolution (SR), wavelet transform, convolutional neural network (CNN), second-order channel attention (SOCA), non-local self-similarity

0 Introduction

High-resolution (HR) images are able to significantly improve the accuracy of image analysis in medical diagnosis[1], remote sensing detection[2], intelligent transportation[3], facial recognition[4] and other fields. However, because of the existence of imaging equipment, atmospheric environment, noise and other factors, the captured images are usually difficult to satisfy the requirements of engineering applications. For decades, to restore a latent HR image with rich detailed information from its available low-resolution (LR) image, varieties of image super-resolution (SR) methods with excellent performance have been proposed.

With the development of deep learning, recent deep convolutional neural networks (CNNs) have been extensively exploited in image SR tasks and achieved considerable performance. To improve the feature representation ability of CNN, the methods of modifying the network structure, including increasing the network depth or/and width, have attracted extensive attention in recent years. For instance, both memory network (MemNet)[5] and residual dense network (RDN)[6] use dense blocks[7] to develop deep network models and take full advantage of hierarchical features extracted from the convolutional layers. Besides designing a deeper or wider network, some networks, such as non-local recurrent network (NLRN)[8] and squeeze-and-excitation network (SENet)[9], strengthen their performance by exploring feature interdependencies of space or channels. For image SR, most of the recent CNN-based models treat intermediate features of the each channel equally, which limits flexibility in highlighting significant features to reveal high-frequency details[5-6]. To break through this limitation, Zhang et al.[10] exploited a residual channel attention network (RCAN) for image SR by designing the residual structure and channel attention mechanism. However, RCAN only exploits the first-order feature statistics and ignores higher-order ones, which limits the representational ability of CNN. To solve this problem, Dai et al.[11]

① Supported by the National Natural Science Foundation of China (No.61901183), Fundamental Research Funds for the Central Universities (No.ZQN-921), Natural Science Foundation of Fujian Province Science and Technology Department (No.2021H6037), Key Project of Quanzhou Science and Technology Plan (No.2021C008R), Natural Science Foundation of Fujian Province (No.2019J01010561), Education and Scientific Research Project for Young and Middle-aged Teachers of Fujian Province 2019 (No.JAT191080), and Science and Technology Bureau of Quanzhou (No.2017G046).

② To whom correspondence should be addressed. E-mail: huangdetian@hqu.edu.cn.

Received on Apr. 26, 2021

built a second-order attention network (SAN) for image SR by developing a second-order channel attention (SOCA) block to learning feature relationships between intermediate layers of the network.

In recent years, some models in Refs[12-14] combining wavelet transform with CNN have also been proposed. Kang et al.[12] proved that training CNN on wavelet sub-bands is beneficial to feature learning, and then proposed a wavelet residual network (WavResNet) to restore abundant texture details. By converting the SR problem to a problem of wavelet coefficient prediction, Guo et al.[13] presented a deep wavelet prediction for image super-resolution (DWSR) to recover lost details of wavelet coefficients of original images to be reconstructed. Unfortunately, both WavResNet and DWSR just explore one-level wavelet decomposition and process each wavelet sub-band independently, which ignores the dependencies between these sub-bands. Inspired by the U-Net architecture, Liu et al.[14] developed a multi-level wavelet convolutional neural network (MWCNN), in which wavelet transform is adopted to substitute the conventional pooling layer to avoid information loss.

In addition, numerous studies on image restoration in Refs[8,15] show that capturing the interdependence of long-distance information from an image can help restore more edge and texture details. For an image, the convolution operation can handle local region information solely, and as for long-distance information, it is necessary to continuously superimpose the convolutional layer to expand the receptive field. However, such methods are inefficient and their network structure are complex and difficult to optimize. To address the issues, Wang et al.[16] proposed non-local neural networks, which calculate the weighted sum of features at each location and treat it as the response of the corresponding location to effectively extract long-distance feature dependencies. Besides, Liu et al.[8] improved the non-local neural network and combined it with recurrent neural networks (RNNs), which boosts the utilization of parameters and the robustness of the model.

Inspired by the above literatures, a channel attention based wavelet cascaded network for image super-resolution (CWSR) is proposed by fully exploiting the superiorities of the wavelet transform, CNN, SOCA and non-local self-similarity prior. The primary contributions of this paper are as follows. (1) A novel CWSR model is proposed to reconstruct as many high-frequency details as possible. Extensive experiments demonstrate that CWSR outperforms state-of-the-art methods for comparison in both visual quality and quantitative metrics. (2) SOCA is incorporated into the network to adaptively rearrange channel-wise features through exploring the inherent interdependencies between different channels. (3) The non-local module is integrated into the network to learn interdependencies between each feature and its neighborhood to enhance the quality of residual features. (4) In the spatial and wavelet domains, a dual-constrained loss function is proposed to optimize the proposed network to minimize the differences between the reconstructed image and its original HR image.

The subsequent structure of this paper is as follows. Section 1 introduces related work. Section 2 describes the detail of the proposed network. Section 3 presents the experimental results. And finally, the conclusions of the paper are drawn in Section 4.

1 Related work

Recently, the deep CNNs have achieved unprecedented success in various machine vision tasks including image SR[17-23]. However, most CNN-based models for image SR treat intermediate features of different channels equally, which limits the super-resolution performance. To deal with this problem, various attention mechanisms in Refs[7, 10, 11, 15-16, 23-25] have been explored in CNN-based approaches.

1.1 CNN-based SR approaches

In the image SR community, Dong et al.[17] developed a three-layer lightweight CNN for image SR (SRCNN), which implements end-to-end mapping between LR and HR images. To reduce the computational burden of SRCNN, Dong et al.[18] further proposed a fast super-resolution CNN (FSRCNN) by employing the deconvolution layer to enlarge the image and utilizing smaller filter sizes and more mapping layers. Similarly, Shi et al.[19] built an efficient sub-pixel CNN (ESPCN) by developing a sub-pixel convolution layer to enlarge feature maps extracted from LR images. Compared with SRCNN, FSRCNN and ESPCN achieved significant improvement in both super-resolution performance and computational efficiency. To further promote the super-resolution performance, Kim et al.[20]. utilized residual learning[21] to exploit a very deep convolutional network (VDSR). Subsequently, by designing deeper and wider residual modules, Lim et al.[23] built an enhanced deep SR network (EDSR), which achieves great success in both reconstruction accuracy and computational efficiency.

1.2 Attention mechanism

Although the residual module facilitates to im-

prove the super-resolution performance by increasing the network depth, the network becomes difficult to converge after it reaches a certain depth. To tackle this issue, the methods of embedding attention mechanisms, such as spatial attention and channel attention, into CNN-based models in Refs[10,11,15-16,24-25] have received more and more attention. Wang et al.[16] built a non-local neural network by developing non-local operation to capture long-range dependency. It is worth noting that non-local operations can be easily incorporated into various computer vision networks and boost their performance. Liu et al.[8] also designed non-local operations and incorporated them into the RNN for end-to-end training to capture feature correlation between each location and its neighborhood. The difference between Ref. [16] and Ref. [8] is that the former measures feature correlation of each location throughout the whole image, while the later measures feature correlation of each location only in its neighborhood. Zhang et al.[15] built residual non-local attention networks (RNAN), in which local and non-local attention blocks were designed to capture the long-range dependency between pixels to promote the representation ability of the network. Except for local and non-local attention, channel attention is developed to explore the dependency between network channels. RCAN[10] and SAN[11] respectively employed the first- and second-order feature statistics to develop different channel attention mechanisms to enhance the representational ability of CNNs.

2 Proposed network

Considering that most existing CNN-based models do not make full use of the information of the original LR images and treat each channel-wise feature equally, a channel attention based wavelet cascaded network for image super-resolution is proposed to further improve the super-resolution performance.

2.1 Network framework

Wavelet transform possesses multi-resolution decomposition characteristics, it is able to effectively decompose the 'contour' and 'detail' features of the image. With this advantage of the wavelet transform, the image SR is performed in the wavelet domain, rather than in the spatial domain, to overcome the shortcoming that local features are hard to be well represented in the spatial domain[12-13].

CWSR, as shown in Fig. 1, is essentially a U-Net architecture network, and each level of the network includes a contracting sub-network and an expanding sub-network. Considering that the discrete wavelet transform (DWT) is a reversible operation and can simultaneously capture the frequency and position information of features, it is incorporated into the contracting sub-network to replace the conventional pooling operation to preserve the edge and texture features of the input image and avoid the loss of information. In the expanding sub-network, the inverse discrete wavelet transform (IDWT) is used to implement the upsampling operation to achieve the mapping from LR to HR features. At the same time, to explore and utilize the feature interdependencies between the four wavelet sub-bands, a convolutional layer is introduced in the proposed model to strengthen the image details after each level of DWT operation. Such methods, which have been exploited in the exiting models in Refs[12-14], can effectively improve the super-resolution performance.

As shown in Fig. 1, CWSR consists of 3 parts. Each group of DWT and IDWT of the same size constitute a part of the network. That is, DWT1 and IDWT1, DWT2 and IDWT2, DWT3 and IDWT3 constitute the first, second and third part of the network, respectively. In a certain part, three CNN units and a CSOCA module are connected after each level of DWT, where each CNN unit contains a 4-layer fully convolutional network (FCN) and all sub-bands are provided as inputs; the CSOCA module consists of a convolutional (Conv) layer and a SOCA block, and the Conv layer of CSOCA is used for feature selection. Specifically, each layer of the CNN unit contains three Conv filters, batch normalization (BN), and rectified linear unit (ReLU). It is noteworthy that the last CNN unit (CNN18) in the network utilizes only one Conv layer (without BN and ReLU) to compress the channel number. CSOCA module consists of Conv, global covariance block, Conv, ReLU and Sigmoid. After features are input into the Sigmoid, it will output weights f (as shown in SOCA module in Fig. 1) ranging from 0 to 1, which are used to measure the importance of the features among channels. To realize the mapping from shallow features to deep features and network training, the features obtained by the CSOCA1 and CSOCA2 modules and the features obtained by the CNN25 and CNN15 modules are added with the element-wise summation, respectively. In addition, in the final stage of obtaining the reconstructed image, a non-local module was added before the CNN18 block to enhance residual features.

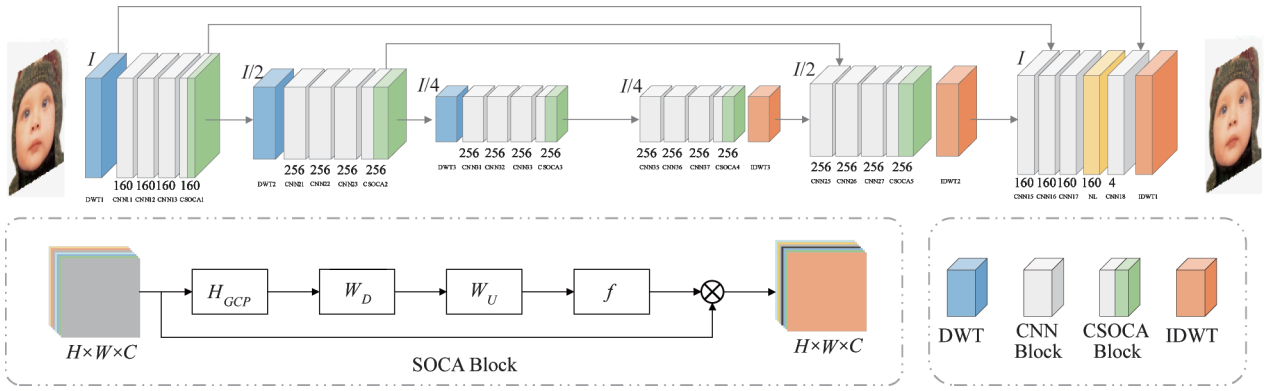

Fig. 1 The architecture of CWSR

The workflow of CWSR is as follows. Firstly, to achieve both odd and even magnification, an image of the same size as the HR image are fed, which is obtained by up-sampling the image to be reconstructed with bicubic interpolation, into CWSR, instead of directly feeding the original LR image. Then, the LR image is first decomposed into four sub-bands by performing DWT1 operation, and then four sub-bands are separately fed into CNN11 as four channels to investigate the relationship of the sub-bands. Subsequently, after performing each level of DWT operation or before each level of IDWT operation, SOCA is employed to explore interdependencies between channel-wise features of four wavelet sub-bands. Next, the non-local module is used to enhance the quality of the residual features before performing the last level IDWT (IDWT1) operation. Finally, four wavelet sub-bands obtained by DWT1 are individually added to the corresponding residual images obtained by CNN18, and then IDWT1 is performed to obtain the final reconstructed image. It is worth noting that each time a DWT operation is performed, the size of the feature map will be reduced to 1/4 of the original size, and the number of the corresponding channel will be increased to four times that of the original one; on the other hand, all feature maps obtained by DWT are input into the CNN module for training, instead of training each sub-band separately.

As with DWSR[13], the Haar kernels are selected as the wavelet basis function. In CWSR, assuming that the size of the input LR image is $4I$ and the initial number of channels is n, the feature map size obtained by performing a one-level DWT on the input is I, and the corresponding number of channels is increased to $4n$. Among them, the feature corresponding to $[1:n]$ channels is the low-frequency approximation sub-band LL, the features corresponding to $[n+1:2n]$, $[2n+1:3n]$ and $[3n+1:4n]$ channels are the high-frequency detail sub-band LH, HL and HH in the horizontal, vertical and diagonal direction, respectively.

2.2 Second-order channel attention

Compared with traditional CNN, the advantage of DWT is that its frequency and location characteristics facilitate the preservation of edges and textures because of its biorthogonal property[14]. Moreover, since DWT is reversible, down-sampling the image with DWT ensures that all image information will be preserved. Considering that there is a meaningful relationship among the four wavelet sub-bands obtained by DWT decomposition, the CNN module is utilized to exploit their relationship.

For CNN-based SR models, if the low-frequency and high-frequency information of the input image are treated indiscriminately in each channel, the powerful representation ability of CNN will be suppressed[10-11]. To effectively explore dependencies between channel-wise features, SOCA is incorporated into the proposed CWSR, which enables the network to learn more high-frequency features to boost its representational ability. As shown in Fig. 1, SOCA block is fused with the last Conv layer after each level of DWT operation, and the fused module is named as CSOCA. Where, SOCA is employed to learn feature dependencies adaptively by utilizing second-order feature statistics for a more discriminative representation. Assuming a $W \times H \times C$ feature map $F = [F_1, F_2, \cdots, F_C]$ with C feature maps with the size of $W \times H$, then the feature map F is reshaped to a feature matrix $\boldsymbol{X}$ with $S \times C$, (where, $S = W \times H$), subsequently, the covariance matrix $\boldsymbol{\Sigma}$ is obtained:

$$\boldsymbol{\Sigma} = \boldsymbol{X}\bar{\boldsymbol{I}}\boldsymbol{X}^{\mathrm{T}} \tag{1}$$

where, $\bar{\boldsymbol{I}} = \frac{1}{S}\left(\boldsymbol{I} - \frac{1}{S}\boldsymbol{1}\right)$, $\boldsymbol{I}$ is the $S \times S$ identity matrix, $\boldsymbol{1}$ is the $S \times S$ matrix of all ones. Since $\boldsymbol{\Sigma}$ is symmetric positive semi-definite, covariance normalization (CN) is performed on $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} \tag{2}$$

where, $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ are an orthogonal matrix and diagonal

matrix, respectively, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_C)$. Subsequently, CN can be converted to the power of eigenvalues.

$$\boldsymbol{Y} = \boldsymbol{\Sigma}^{\alpha} \tag{3}$$

where, $\alpha = \frac{1}{2}$. Let $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_c, \cdots, \boldsymbol{y}_C]$, the channel-wise statistics $\boldsymbol{z} \in R^{C\times 1}$ can be calculated by shrinking $\boldsymbol{Y}$, so the cth dimension of $\boldsymbol{z}$ is as

$$z_c = H_{GCP}(\boldsymbol{y}_c) = \frac{1}{c}\sum_{i}^{C} \boldsymbol{y}_c(i) \tag{4}$$

where, $H_{GCP}(\cdot)$ stands for the global covariance pooling (GCP) function. Then, the channel attention map $\boldsymbol{w}$ can be calculated by

$$\boldsymbol{w} = \mathrm{sigmoid}(conv2(\mathrm{ReLU}(conv1(z)))) \tag{5}$$

where, $conv1$ and $conv2$ respectively compress and expand the channel number of the input, whose purpose is to increase the non-linear representation. In Fig. 1, W_D and W_U separately denote the set weights of $conv1$ and $conv2$. Finally, the obtained $\boldsymbol{w}$ can be used to rescale the input feature.

$$\hat{\boldsymbol{f}}_c = \omega_c \cdot \boldsymbol{f}_c \tag{6}$$

where, ω_c and $\boldsymbol{f}_c$ represent the scaling factor and the cth channel of the feature map, respectively.

2.3 Non-local module

The non-local operation can be used to capture long-range dependency, i. e., capturing the dependency between a location and its neighborhood, which breaks through the limitation of the local operation of traditional CNNs[15-16]. Then, the obtained dependency can be used as a weight to represent the similarity between other locations and the current location to be calculated.

Inspired by Ref. [26], the non-local operation is wrapped as a non-local module by adding a skip connection, as shown in Fig. 2, and then the non-local module is incorporated into the network to produce reliable feature dependencies to enhance the quality of residual features. Furthermore, to decrease the amount of calculation, a bottleneck structure is also introduced. Assuming the channel number of input feature is $4C$, its channel number is reduced to 1/4 of the original channel number after it is operated by θ, φ and g, where, θ, φ and g all represent $1 \times 1 \times C$ convolution operation, ν denotes a $1 \times 1 \times 4C$ convolution operation, $\otimes$ and $\oplus$ stand for dot multiplication and addition.

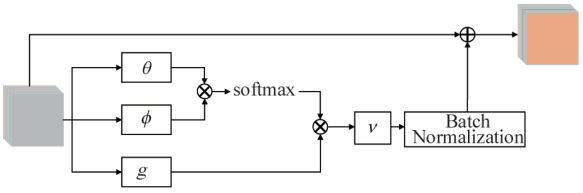

Fig. 2 Non-local module

Concretely, the non-local module takes a multi-channel input $\boldsymbol{M}$ as the image feature and generates an output feature $\boldsymbol{O}$. Its expression is

$$\begin{aligned}\boldsymbol{h}_i &= \mathrm{softmax}(\theta(\boldsymbol{m}_i)^{\mathrm{T}}\varphi(\boldsymbol{m}_j))g(\boldsymbol{m}_j)\\ &= \frac{1}{\sum_{\forall j} e^{\theta(\boldsymbol{m}_i)^{\mathrm{T}}\varphi(\boldsymbol{m}_j)}} e^{\theta(\boldsymbol{m}_i)^{\mathrm{T}}\varphi(\boldsymbol{m}_j)} W_g \boldsymbol{m}_j\end{aligned} \tag{7}$$

$$\boldsymbol{o}_i = BN(\boldsymbol{\nu}\boldsymbol{h}_i) + \boldsymbol{m}_i \tag{8}$$

where, $\boldsymbol{m}_i$ represents the features of the current location i concerned, $\boldsymbol{m}_j$ represents the neighboring location of $\boldsymbol{m}_i$; $\theta(\boldsymbol{m}_i) = W_\theta \boldsymbol{m}_i$, $\varphi(\boldsymbol{m}_j) = W_\varphi \boldsymbol{m}_j$, $g(\boldsymbol{m}_j) = W_g \boldsymbol{m}_j$, W_θ, W_φ and W_g individually represent the weight matrix to be learned, and can be obtained by 1×1 convolution; $\sum_{\forall j} e^{\theta(\boldsymbol{m}_i)^{\mathrm{T}}\varphi(\boldsymbol{m}_j)}$ stands for the normalization factor; $e^{\theta(\boldsymbol{m}_i)^{\mathrm{T}}\varphi(\boldsymbol{m}_j)}$ denotes calculating the similarity of $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$, and the superscript T stands for transpose; $\boldsymbol{o}_i$ denotes the output feature of the location i, $BN(\cdot)$ denotes batch normalization, and $\boldsymbol{\nu}$ denotes the weight matrix of $\boldsymbol{o}_i$ to be learned. From Eq. (7) and Eq. (8), it can be seen that the non-local operation calculates the normalized correlation between each feature and its neighborhood for the current feature map, and its output is the weighted average of its neighborhood.

2.4 Loss function

IDWT operation is performed to generate a reconstructed image as a final result from a series of wavelet sub-bands. On the one hand, the wavelet domain loss is used to constrain the proposed model to recover more high-frequency details. On the other hand, the spatial domain loss is utilized to constrain the proposed model to achieve a balance between edge texture features and smooth features. Eventually, a dual-constrained loss function is proposed to optimize the proposed CWSR. Thus, the total loss is composed of the wavelet domain loss $loss_{\mathrm{wav}}$ and the spatial domain one $loss_{\mathrm{img}}$.

Since the L_2 norm can not only be used to measure the difference between two vectors, but also prevent overfitting in model training, which greatly improves the generalization ability of the model. Therefore, a novel loss function is proposed based on the L_2 norm, and its expression is as

$$loss_{\mathrm{total}} = \lambda loss_{\mathrm{wav}} + (1-\lambda) loss_{\mathrm{img}} \tag{9}$$

where, λ and $1-\lambda$ represent the weights of the wavelet and spatial domain loss, respectively. Then, the loss function can be obtained as

$$loss = \frac{1}{2N}\sum_{i=1}^{N} \| I_{SR} - I_{HR} \|^2 \tag{10}$$

where, N represents the number of training samples, I_{SR} represents the reconstructed image, and I_{HR} represents the original HR image.

Wavelet domain loss. Due to taking full use of the relationships between four sub-bands, the proposed network can avoid information loss, which is conducive to recovering more detailed information. Suppose $\boldsymbol{y}_{LR}$ represents the test LR image, and $\boldsymbol{x}_{HR}$ represents the corresponding original HR image. The input of the proposed network is a middle resolution (MR) image $\boldsymbol{y}_{MR}$ obtained by up-sampling $\boldsymbol{y}_{LR}$. It is necessary to learn the relationship between the wavelet coefficients obtained by feeding $\boldsymbol{y}_{MR}$ and $\boldsymbol{x}_{HR}$ to a one-level DWT, so that the network output will be as close as possible to the wavelet coefficients obtained by performing a one-level DWT on the corresponding HR image.

Calculating the wavelet domain loss $loss_{\text{wave}}$ involves solving $loss_1$ in Fig. 3 and $loss_3$ in Fig. 4, where $loss_{\text{wav}} = loss_1 + loss_3$. The residuals obtained by the CSOCA1 module are respectively added to the four sub-bands to form $DCS1(\boldsymbol{LL}_1, \boldsymbol{LH}_1, \boldsymbol{HL}_1, \boldsymbol{HH}_1)$. Suppose that the four sub-bands, represented as $DWT_{HR}(\boldsymbol{LL}, \boldsymbol{LH}, \boldsymbol{HL}, \boldsymbol{HH})$, are obtained by feeding the original HR image $\boldsymbol{x}_{HR}$ to a one-level DWT, then $loss_1$ between $DCS1$ and DWT_{HR} can be solved according to L_2 norm. Concretely, each sub-band of $DCS1$ first solves the loss with its corresponding sub-band of DWT_{HR}, and then the losses of four sub-bands are summed to obtain the final loss $loss_1$, that is, $loss_1 = loss_{LL1} + loss_{LH1} + loss_{HL1} + loss_{HH1}$. The purpose of calculating $loss_1$ is to constrain the extracted features in the shallow network so that the edge and texture features of the reconstructed image are as close as possible to that of the original HR image.

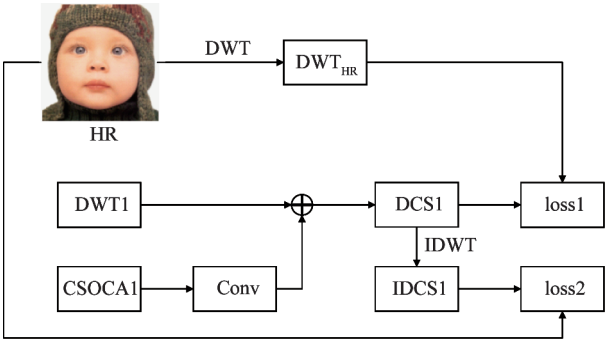

Fig. 3 Loss of the shallow network

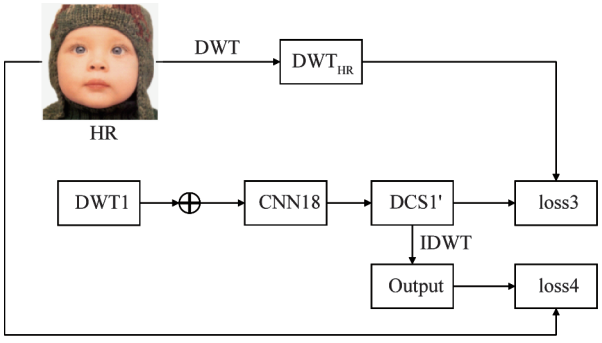

Fig. 4 Loss of the deep network

Similarly, $loss_3$ in Fig. 4 can be solved by $DCS1'$ and DWT_{HR} according to L_2 norm, where, $DCS1'$ represents the residual images obtained by adding the four wavelet sub-bands obtained by DWT1 to the corresponding residuals obtained by CNN18, as shown in Fig. 1, which can be used as a supplement to the detailed information of the MR image $\boldsymbol{y}_{MR}$ during the reconstruction process to boost the super-resolution performance. Similarly, each sub-band of $DCS1'$ first calculates the loss with its corresponding sub-band of the HR image, and then the losses of four sub-bands are summed to obtain the final loss $loss_3$, that is, $loss_3 = loss_{LL3} + loss_{LH3} + loss_{HL3} + loss_{HH3}$.

Spatial domain loss. To achieve a balance between detail and smooth features, in addition to the wavelet domain loss, the spatial domain loss of the image is also concerned. Calculating the spatial domain loss $loss_{\text{img}}$ involves solving $loss_2$ in Fig. 3 and $loss_4$ in Fig. 4, where $loss_{\text{img}} = loss_2 + loss_4$. In Fig. 3, $IDCS1$ represents the result obtained after the IDWT is performed on $DCS1(\boldsymbol{LL}_1, \boldsymbol{LH}_1, \boldsymbol{HL}_1, \boldsymbol{HH}_1)$. Since $IDCS1$ is essentially a rough reconstructed image to be further optimized, $loss_2$ between $IDCS1$ and the original HR image $\boldsymbol{x}_{HR}$ can be solved according to L_2 norm. In Fig. 4, since the reconstructed image $\boldsymbol{x}_{SR}$ obtained by performing a one-level IDWT on $DCS1'$ is equivalent to the output of the proposed CWSR, $loss_4$ between $\boldsymbol{x}_{SR}$ and $\boldsymbol{x}_{HR}$ can be calculated according to L_2 norm.

3 Experiment

3.1 Parameters, datasets and metrics

To train CWSR, a large training consists of the images from the following dataset, including BSD[27], DIV2K[28] and WED[29]. Specifically, 200 HR images were selected from BSD, 800 ones were selected from DIV2K, and 4744 ones were selected from WED. During training, 24×6000 patches with the size of 240×240 were cropped from the training images. For the network parameters, their initialization is the same as Ref. [21]. ADAM optimizer[30] was employed to train the proposed CWSR. A min-batch size was 32, and other hyper parameters of ADAM are set to default values. During the iteration, the learning rate decays from

0.01 to 0.0001. Moreover, the neighborhood size is 45 × 45 in the non-local module. The proposed CWSR has been implemented on two Nvidia Titan RTX 24GB GPUs.

To ensure the objectivity of the experiments, four benchmark sets including Set5[31], Set14[32], BSD100 (100 images derived from BSDS 300[27]) and Urban100[33] were selected as test datasets. Moreover, the reconstructed images are gauged with subjective visual perception and objective evaluation metrics including peak signal-to-noise ratio (PSNR) and structural similarity (SSIM)[34].

3.2 Experimental results

To quantitatively evaluate the reconstructed image, the MR images obtained by up-sampling LR ones are regarded as the input images to be reconstructed, and the original HR images are regarded as the reference ones.

3.2.1 Analysis of different modules

To explore the feasibility and effectiveness of different modules, the proposed CWSR model is compared with two intermediate models, i.e., WCN is the model using only wavelet transform and CNN modules, and WCN + SOCA are the model using wavelet transform, CNN and SOCA modules. Essentially, CWSR is the model using wavelet transform, CNN, SOCA and non-local modules. Table 1 lists the average PSNR and SSIM of reconstructed images obtained by three models mentioned above. As can be seen from Table 1, although WCN + SOCA performs much better than WCN, CWSR achieves the optimal super-resolution performance with the highest average PSNR and SSIM. These results indicate that almost all different modules of CWSR have positive significance.

Table 1 Comparison of the average PSNR (dB) and SSIM obtained by different modules with scale factor ×4

Images	PSNR/SSIM	WCN	WCN + SOCA	CWSR
Set5	PSNR	32.01	32.10	32.24
	SSIM	0.8942	0.8943	0.8947
Set14	PSNR	28.22	28.39	28.50
	SSIM	0.7819	0.7821	0.7828
B100	PSNR	27.43	27.56	27.67
	SSIM	0.7354	0.7357	0.7360
Urban100	PSNR	26.20	26.30	26.39
	SSIM	0.7896	0.7926	0.7929

3.2.2 Analysis of network structure

CWSR can essentially be extended to different levels of DWT. However, higher level of DWT directly means deeper network and higher computational complexity. Accordingly, an appropriate level of DWT is required to balance super-resolution performance and computational efficiency. This experiment compared the performance of the models (i.e., CWSR-1, CWSR-2, CWSR-3 and CWSR-4) with 1-, 2-, 3- and 4-level DWT with scale factor 4.

Table 2 presents the average PSNR, SSIM and computational time of different models with the level of 1 to 4. As can be seen from Table 2, in terms of both PSNR and SSIM metrics, CWSR-3 is significantly superior to CWSR-1 and CWSR-2, while CWSR-4 has only a negligible improvement over CWSR-3; then, combined with the computational time, it can be seen that the CWSR-3 model, which is the default CWSR model, has a better tradeoff between super-resolution performance and computational efficiency than the other three models. The main reason is that $\boldsymbol{LL}_n$ contains scarcely effective low-frequency information for image SR after an appropriate level of DWT.

Table 2 Comparison of the average PSNR (dB), SSIM and computational time (seconds) of CWSRs with different levels of DWT with scale factor ×4

Images	PSNR/SSIM/Time	CWSR-1	CWSR-2	CWSR-3	CWSR-4
Set5	PSNR	31.66	32.05	32.24	32.25
	SSIM	0.8901	0.8930	0.8947	0.8948
	Time	0.306	0.403	0.517	0.649
Set14	PSNR	28.02	28.23	28.50	28.51
	SSIM	0.7758	0.7802	0.7828	0.7830
	Time	1.017	1.262	1.548	1.815
B100	PSNR	27.46	27.60	27.67	27.68
	SSIM	0.7344	0.7349	0.7360	0.7363
	Time	0.265	0.336	0.431	0.536
Urban100	PSNR	25.77	26.15	26.39	26.40
	SSIM	0.7724	0.7866	0.7929	0.7932
	Time	3.610	4.919	6.803	8.546

3.2.3 Comparison with state-of-the-art methods

To further validate the effectiveness of CWSR, it is compared with state-of-the-art methods, i.e., SRCNN[17], VDSR[21], LapSRN[35], DRRN[36], IDN[37], DPSR[38], IMDN[39] and MWCNN[14]. Table 3 lists the average PSNR and SSIM of reconstructed images obtained by various methods with different scale factors (i.e., ×2, ×3 and ×4) on Set5, Set14, B100 and Urban 100. From Table 3, it can be seen that the average PSNR and SSIM of reconstructed images obtained by CWSR are higher than that of most methods for

comparison, which indicates that the proposed method produces the leading result in overall super-resolution performance. Specifically, as can be seen from Table 3, for the case of scale factor with 2, the average PSNR of CWSR is only slightly lower than that of IMDN on the Set5 dataset, but the average SSIM is still higher than IMDN; at the same time, both the average PSNR and SSIM of the proposed CWSR are the highest on other three datasets. And for the case of scale factor with 3, although the average PSNR of the proposed method is rarely insufficient on the Set5 and Set14 datasets, the average SSIM is also the highest on all datasets. For the case of scale factor with 4, the proposed method almost achieves optimal or suboptimal performance in terms of PSNR and SSIM metrics.

Table 3 Comparison of the average PSNR (dB) and SSIM obtained by different methods with different scale factors

Images	PSNR/SSIM	Scale	SRCNN	VDSR	LapSRN	DRRN	IDN	DPSR	IMDN	MWCNN	CWSR
Set5	PSNR	×2	36.66	37.53	37.52	37.74	37.83	37.78	38.00	37.91	37.98
	SSIM		0.9542	0.9587	0.9590	0.9591	0.9600	0.9600	0.9605	0.9600	0.9608
Set14	PSNR		32.45	33.03	33.08	33.23	33.30	33.58	33.63	33.70	33.81
	SSIM		0.9067	0.9124	0.9130	0.9136	0.9148	0.9171	0.9177	0.9182	0.9213
B100	PSNR		31.36	31.90	31.80	32.05	32.08	32.17	32.19	32.23	32.29
	SSIM		0.8879	0.8960	0.8950	0.8973	0.8985	0.8993	0.8996	0.8999	0.9003
Urban 100	PSNR		29.50	30.76	30.41	31.23	31.27	31.93	32.17	32.30	32.45
	SSIM		0.8946	0.9140	0.9100	0.9188	0.9196	0.9264	0.9283	0.9296	0.9330
Set5	PSNR	×3	32.75	33.66	-	34.03	34.11	34.33	34.36	34.17	34.25
	SSIM		0.9090	0.9213	-	0.9244	0.9253	0.9266	0.9270	0.9261	0.9272
Set14	PSNR		29.30	29.77	-	29.96	29.99	30.33	30.32	30.16	30.26
	SSIM		0.8215	0.8314	-	0.8349	0.8354	0.8424	0.8417	0.8414	0.8424
B100	PSNR		28.41	28.82	-	28.95	28.95	29.11	29.09	29.12	29.16
	SSIM		0.7863	0.7976	-	0.8004	0.8013	0.8057	0.8046	0.8060	0.8088
Urban 100	PSNR		26.24	27.14	-	27.53	27.42	28.11	28.17	28.13	28.27
	SSIM		0.7989	0.8279	-	0.8378	0.8359	0.8514	0.8519	0.8514	0.8545
Set5	PSNR	×4	30.48	31.35	31.54	31.68	31.82	32.19	32.21	32.12	32.24
	SSIM		0.8628	0.8838	0.8850	0.8888	0.8903	0.8954	0.8946	0.8941	0.8947
Set14	PSNR		27.50	28.01	28.19	28.21	28.25	28.65	28.58	28.41	28.50
	SSIM		0.7513	0.7674	0.7720	0.7720	0.7730	0.7833	0.7811	0.7816	0.7828
B100	PSNR		26.90	27.29	27.32	27.38	27.41	27.59	27.56	27.62	27.67
	SSIM		0.7101	0.7251	0.7280	0.7284	0.7297	0.7360	0.7353	0.7355	0.7360
Urban 100	PSNR		24.52	25.18	25.21	25.44	25.41	26.14	26.04	26.27	26.39
	SSIM		0.7221	0.7524	0.7560	0.7638	0.7632	0.7867	0.7838	0.7890	0.7929

To intuitively compare reconstructed images of different SR methods from subjective visual perception, Fig. 5 – Fig. 8 illustrate the enlarged results of reconstructed images at the same area with scale factor 4, and the corresponding original HR image is given as a reference image. From Fig. 5 – Fig. 8, it can be seen that most compared methods are unable to accurately reconstruct the edge and texture details, and even suffer from serious artifacts. However, CWSR reconstructs a sharper image and restores more high-frequency details.

Fig. 5 presents the visual comparisons of various methods on the image 'butterfly' from Set5. It can be seen from Fig. 5 that the images reconstructed by SRCNN, VDSR and LapSRN are not clear enough; the image reconstructed by DRRN even has aliasing; IDN and DPSR can only recover the main contour of the image, but not more detailed information; in contrast, IMDN, MWCNN and the proposed method can recover more details and achieve better super-resolution results. However, compared with IMDN and MWCNN, the reconstructed image of CWSR has sharper contours and better preserves the edge and texture information of the butterfly wings.

Fig. 6 shows the visual comparisons of various methods on the image 'Baboon' from Set14. As can

be seen from Fig. 6, the reconstructed image of CWSR retains the details of the beard well, and the images reconstructed by DPSR, IMDN and MWCNN are slightly blurry. This is mainly because CWSR makes full use of wavelet transform, channel attention and non-local prior to jointly recover more high-frequency details.

Fig. 7 illustrates the visual comparisons of various methods on the image 'Zebra' from B100. As can be seen from Fig. 7, although the reconstructed image of LapSRN algorithm is better than that of SRCNN, VDSR and IDN in edge sharpening, many artificial details appear in its reconstructed images; compared with LapSRN, the images reconstructed by DRRN, DPSR and IMDN have only a few artificial details and contain sharper edge; compared with the previous SR algorithms, MWCNN can obviously restore more edge details, and there are almost no artificial details in its reconstructed images; however, thanks to the attention

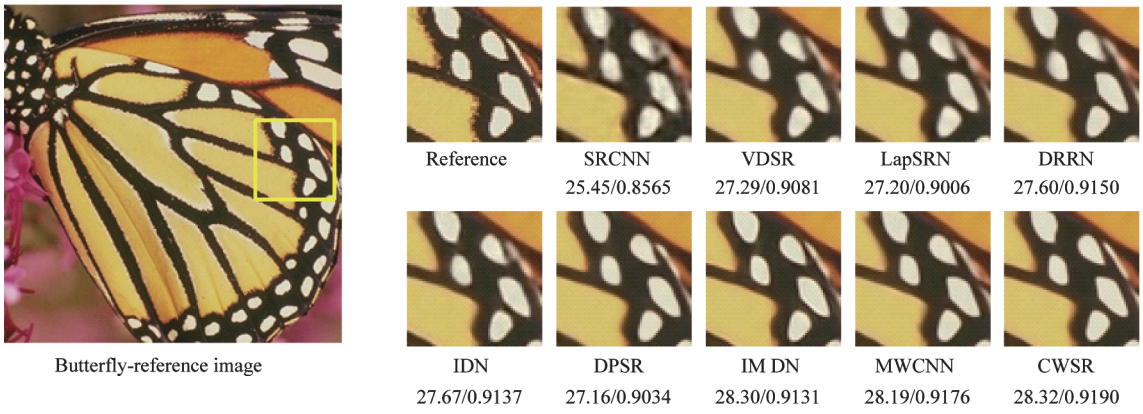

Fig. 5 Visual comparison of super-resolution results of 'Butterfly' (Set5) obtained by different algorithms with scale factor ×4

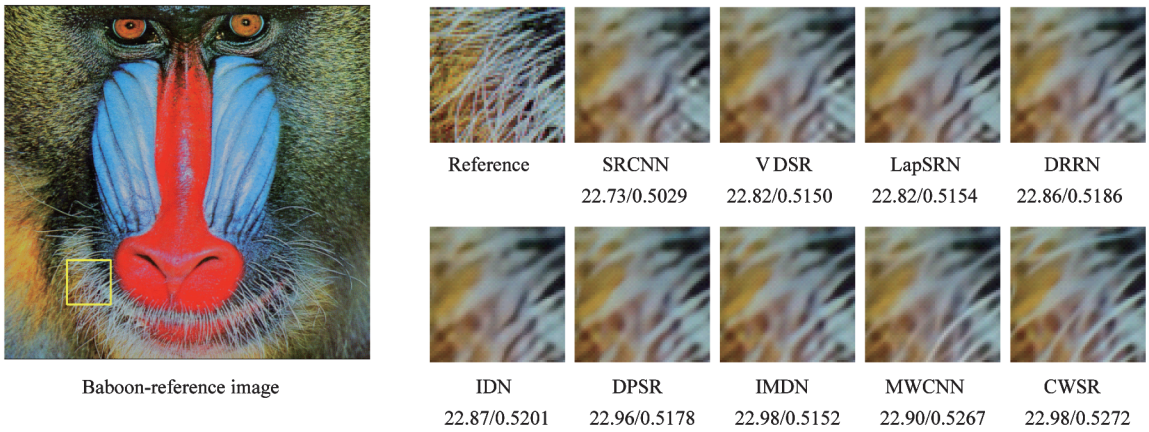

Fig. 6 Visual comparison of super-resolution results of 'Baboon' (Set14) obtained by different algorithms with scale factor ×4

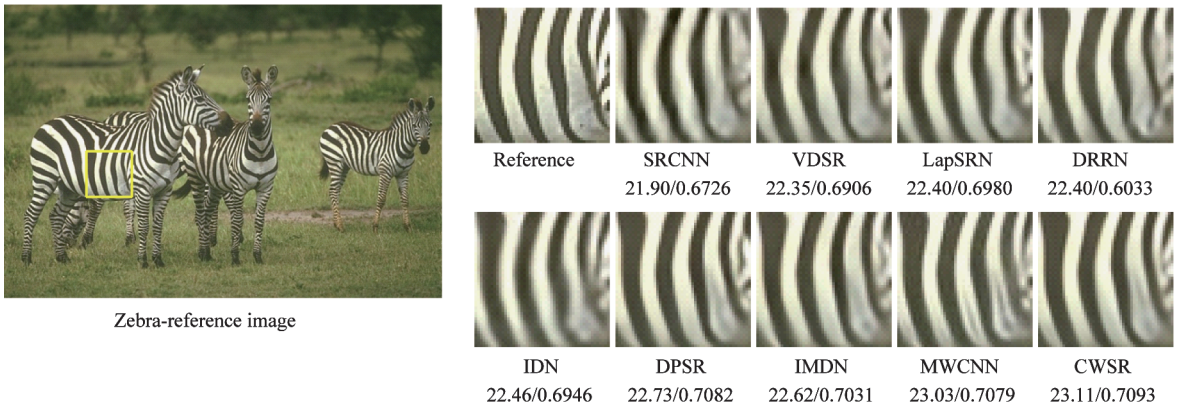

Fig. 7 Visual comparison of super-resolution results of 'Zebra' (B100) obtained by different algorithms with scale factor ×4

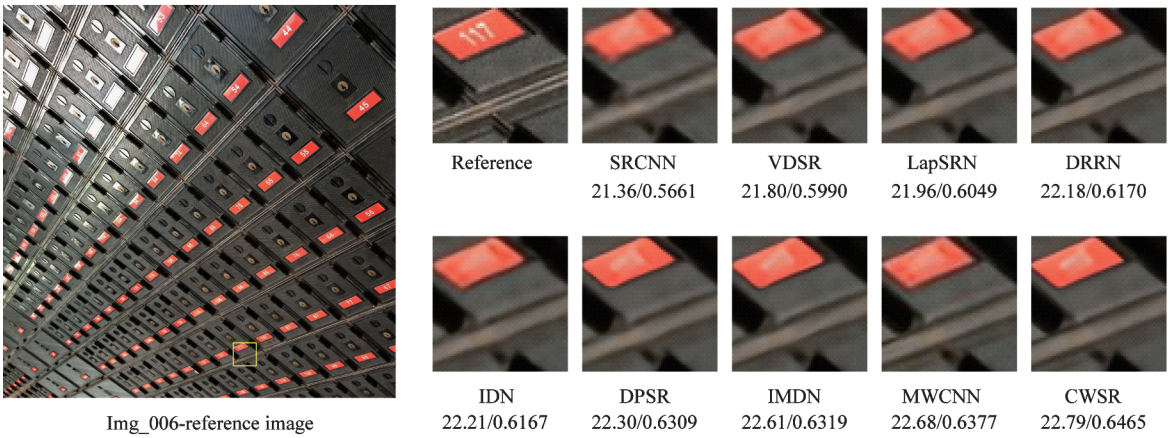

Fig. 8 Visual comparison of super-resolution results of 'Img_006' (Urban100) obtained by different algorithms with scale factor ×4

mechanism that can explore dependencies between channel-wise features and the non-local module that can further enhance the residuals, CWSR is slightly better than MWCNN in restoring the edge details.

Fig. 8 presents the visual comparisons of various methods on the image 'Img_006' from Urban100. It can be seen from Fig. 8 that the reconstructed image of SRCNN has some distortion, while the edge and texture of reconstructed images of VDSR, IDN, LapSRN, DPSR and DRRN are blurred; obviously, IMDN and MWCNN outperform the previous methods, but also are still inferior to CWSR. The image reconstructed by the proposed CWSR has better visual effects than that by MWCNN, with sharper edges and textures.

4 Conclusion

To obtain more high-frequency information, a channel attention based wavelet cascaded network for image super-resolution is proposed. A SOCA module is incorporated into the proposed network to adaptively learn the inherent correlations of channel-wise features, and then the non-local module is utilized to capture interdependencies between each feature location and its neighborhoods to boost residual features, finally a novel dual-constrain loss function based on the spatial and wavelet domains is proposed to strengthen the constraints on network training. Experimental results demonstrate the superiority of CWSR in comparison with several state-of-the-art super-resolution methods.

References

[1] ZENG K, DING S, JIA W. Single image super-resolution using a polymorphic parallel CNN[J]. *Applied Intelligence*, 2019, 49(1): 292-300

[2] LEI S, SHI Z, ZOU Z. Coupled adversarial training for remote sensing image super-resolution[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(5): 3633-3643

[3] ZHU J, ZENG H, HUANG J, et al. Vehicle re-identification using quadruple directional deep learning features[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, 21(1): 410-420

[4] CHEN J, CHEN J, WANG Z, et al. Identity-aware face super-resolution for low-resolution face recognition[J]. *IEEE Signal Processing Letters*, 2020, 27: 645-649

[5] TAI Y, YANG J, LIU X, et al. MemNet: a persistent memory network for image restoration[C]// Proceedings of IEEE International Conference on Computer Vision, Honolulu, USA, 2017: 4549-4557

[6] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 2472-2481

[7] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2261-2269

[8] LIU D, WEN B, FAN Y, et al. Non-local recurrent network for image restoration[C]// Proceedings of Advances in Neural Information Processing Systems, Montreal, Canada, 2018: 1680-1689

[9] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(8): 2011-2023

[10] ZHANG Y, LI K, LI K, et al. Image super-resolution using very deep residual channel attention networks[C]// Proceedings of European Conference on Computer Vision, Warsaw, Poland, 2018: 294-310

[11] DAI T, CAI J, ZHANG Y, et al. Second-order attention network for single image super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2019: 11057-11066

[12] KANG E, MIN J, YE J C. Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction [EB/OL]. https://arxiv.org/abs/1703.01383: arXiv, (2017-03-04), [2021-04-05]

[13] GUO T, MOUSAVI H S, VU T H, et al. Deep wavelet prediction for image super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1100-1109

[14] LIU P, ZHANG H, ZHANG K, et al. Multi-level wavelet-CNN for image restoration[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, USA, 2018: 886-895
[15] ZHANG Y, LI K, LI K, et al. Residual non-local attention networks for image restoration[C]// Proceedings of International Conference on Learning Representations, New Orleans, USA, 2019: 1-18
[16] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7794-7803
[17] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]// Proceedings of European Conference on Computer Vision, Zurich, Switzerland, 2014: 184-199
[18] DONG C, LOY C C, TANG X. Accelerating the super-resolution convolutional neural network[C]// Proceedings of European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 391-407
[19] 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 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1874-1883
[20] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1646-1654
[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778
[22] KIM J, LEE J K, LEE K M. Deeply-recursive convolutional network for image super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1637-1645
[23] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1132-1140
[24] LI F, BAI H, ZHAO Y. Learning a deep dual attention network for video super-resolution[J]. *IEEE Transactions on Image Processing*, 2020, 29: 4474-4488
[25] LI J, CUI R, LI B, et al. Hyperspectral image super-resolution by band attention through adversarial learning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(6): 4304-4318
[26] BUADES A, COLL B, MOREL J. A non-local algorithm for image denoising[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 60-65
[27] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]// Proceedings of IEEE International Conference on Computer Vision, Vancouver, Canada, 2001: 416-423
[28] AGUSTSSON E, TIMOFTE R. NTIRE 2017 challenge on single image super-resolution: dataset and study[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1122-1131
[29] MA K, DUANMU Z, WU Q, et al. Waterloo exploration database: new challenges for image quality assessment models[J]. *IEEE Transactions on Image Processing*, 2017, 26(2): 1004-1016
[30] KINGMA D, BA J. Adam: a method for stochastic optimization[C]// Proceedings of International Conference on Learning Representations, San Diego, USA, 2015: 1-15
[31] BEVILACQUA M, ROUMY A, GUILLEMOT C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[C]// Proceedings of British Machine Vision Conference, Guildford, England, 2012: 1-10
[32] ZEYDE R, ELAD M, PROTTER M. On single image scale-up using sparse-representations[C]// Proceedings of International Conference on Curves and Surfaces, Berlin, Germany, 2012: 711-730
[33] HUANG J, SINGH A, AHUJA N. Single image super-resolution from transformed self-exemplars[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5197-5206
[34] HUANG D, HUANG W, GU P, et al. Image super-resolution reconstruction based on regularization technique and guided filter[J]. *Infrared Physics and Technology*, 2017, 83: 103-113
[35] LAI W, HUANG J, AHUJA N, et al. Deep laplacian pyramid networks for fast and accurate super-resolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 5835-5843
[36] TAI Y, YANG J, LIU X. Image super-resolution via deep recursive residual network[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2790-2798
[37] HUI Z, WANG X, GAO X. Fast and accurate single image super-resolution via information distillation network[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 723-731
[38] ZHANG K, ZUO W, ZHANG L. Deep plug-and-play super-resolution for arbitrary blur kernels[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2019: 1671-1681
[39] HUI Z, GAO X, YANG Y, et al. Lightweight image super-resolution with information multi-distillation network[C]// Proceedings of ACM International Conference on Multimedia, Nice, France, 2019: 2024-2032

CHEN Jian, born in 1996. He is currently pursuing the M. S. degree in computer science and technology from the College of Engineering, Huaqiao University, China. His research interests include computer vision, image super-resolution and deep learning.