Applied Intelligence
https://doi.org/10.1007/s10489-022-03322-9

MBMR-Net: multi-branches multi-resolution cross-projection network for single image super-resolution

Dan Zhang [1] · Binglian Zhu [1] · Yuanhong Zhong [1]

Accepted: 28 January 2022

Abstract
Deep convolutional neural networks (CNNs) have achieved significant developments in the field of single image super resolution (SISR) due to their nonlinear expression ability. However, existing architectures either rely on the representations learned from a single scale or extract deep features by cascading multiple resolutions, which unuse or underutilize the interdependence between low-resolution (LR) images and high-resolution (HR) images. In view of this trait, we propose a deep network called the multi-branches multi-resolution cross-projection network (MBMR-Net), which has multiple parallel branches, and cross-projection is performed between multiple branches to exchange information. Then, we introduce a novel attention unit that integrates second-order channel attention with spatial attention to better fuse information from multiple resolutions. Moreover, in terms of the characteristics of the model, we devise a loss function for enhancing the restoration of high-frequency details while ensuring the content information. Extensive quantitative and qualitative evaluations on benchmark datasets illustrate the effectiveness of our method and its competitive performance over state-of-the-art methods.

Keywords Multi-branches · Cross-projection · Attention mechanism · Single image super-resolution

1 Introduction

Single image super-resolution reconstruction (SISR) is the process of recovering the corresponding high-resolution image (HR) from a given low resolution (LR) image. Due to the high cost of advanced imaging equipment and the burden of data transmission, LR images generally exist in the real world. Therefore, the SISR task has received extensive attention from researchers and is widely used in the fields of facial recognition [1, 2], medical imaging [3, 4], and video surveillance [5, 6]. However, SISR is inherently ill-posed and challenging since there are always multiple HR solutions corresponding to a single LR image. With the active exploration of neural networks in this field, learning-based methods have moved to the mainstream to seek an optimal nonlinear mapping between LR and HR due to their excellent learning ability and superior reconstruction results.

✉ Yuanhong Zhong
zhongyh@cqu.edu.cn

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China

Early single image super-resolution reconstruction methods based on deep learning mainly employed information from a single scale. Dong et al. [7] first proposed the classic three-layer network SRCNN, and its excellent performance indicated the advantages of methods based on CNNs over traditional methods. Later, Dong et al. [8] attempted to accede subpixel convolution as the final layer to super-resolve feature maps with different scale factors for less calculation. Although the recovery effect was further improved, artifacts and aliasing still existed because of the limited receptive field. Deepening the network could enlarge the receptive field but make the model resist convergence. Thereafter, many efforts have been made to promote SISR accuracy by dramatically enlarging the depth of the network with some training strategies. For example, [9, 10] introduced skip connections to facilitate the training procedure, and [11] controlled the number of parameters by recursive learning. Furthermore, Tai et al. [12] extracted and fused multilevel features to enhance the visual qualities. Lim et al. [13] proposed EDSR, which removes the batch-normal layer of the conventional residual block to relieve the calculation burden and employs a data augmentation strategy. The commonality of the mentioned methods is that they are essentially deep feed-forward networks, which fulfill reconstruction by virtue of deep features

Published online: 18 February 2022

from a single scale while ignoring the feature diversity brought by multiple scales.

Previously, some works began to apply multiple scales to explore across-scale patch similarity and provide supplementary information for promoting performance. Lai et al. [14] gradually reconstructed HR images in a coarse-to-fine fashion, which achieved good adaptability even though the scale factor was large. Later, Haris et al. [15] alternately transformed the resolution of features by up/down projection blocks to refine the features of the LR scale and SR scale. Zhu et al. [16] designed a compact back-projection network to improve the DBPN. They adopted a long-skip connection to transmit low-frequency information directly; moreover, the features from both scales were utilized in the reconstruction stage, while DBPN only utilized the features from the target scale. The significant performance improvement illustrates that mining the correlations between different scales is a potential perspective for SR reconstruction.

Obviously, the information from multiple scales will greatly increase the number of parameters and features. To overcome this drawback, some studies introduced an attention mechanism [17] in this field. Structure information is naturally more important than content information in the SISR task. Low-frequency content not only contributes less to the reconstruction of image details but also leads to an increased calculation burden. Zhang et al. [18] proposed RCAN, which integrated channel attention into the proposed symmetric network architecture for adaptively applying different feature components through channel importance cues. Hu et al. [19] incorporated channel attention and spatial attention into a module and stacked it for modulating features in global and local manners. Dai et al. [20] and Gao et al. [21] held that statistics higher than first-order would further promote the discriminative ability of the network and developed a second-order attention network through global covariance pooling.

Following the successful works of cross-scale information utilization and attention mechanism in [22, 23], to make full use of multiscale features, and learn the importance of features initiatively, in this paper, we design a multi-branches multi-resolution cross-projection network (MBMR-Net). Specifically, we design a novel cross-projection SISR framework on multiple branches. The parallel streams conduct the information extraction, exchange, and fusion repeatedly. Considering that there is redundant information among multiple branches, we design an attention unit to ensure informative features. In addition, we devise a loss function for the network so that our model pays more attention to the details. Finally, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [24] are employed as metrics to evaluate the network performance.

In summary, the contributions of this paper are threefold:

1. We propose a multi-branches multi-resolution cross-projection network (MBMR-Net), which to our knowledge is the first attempt to apply the parallel structure to the SISR task. Extensive experiments illustrate the effectiveness of MBMR-Net, which is competitive with state-of-the-art methods.
2. We introduce a novel attention unit to learn channel weights and spatial correlations and then perform differentiated learning on fusion features for powerful feature representation.
3. We devise a loss function because the recovery of other branches can also facilitate target reconstruction. In addition to pixel loss, we calculate gradient maps on each branch to enhance the restoration of high-frequency details.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 describes the proposed method. Section 4 provides ablation analysis and plentiful experimental comparisons. Finally, we conclude the paper in Section 5.

2 Related work

With the progress of deep learning algorithms in recent years, the development of SR tasks has been greatly promoted. Researchers obtain high-quality reconstruction results by adjusting the network structure and introducing efficient strategies. In the following section, we briefly review the related works about SR based on CNNs.

SR based on CNN CNNs are widely applied in the SISR field owing to their powerful nonlinear representation ability. Dong et al. [7] first introduced CNN to SISR, and the impressive results proved its great contribution and inspired more research. Then, VDSR [9] and DRCN [25], the pioneering works of residual learning in SISR, were proposed successively. Assisted by residual learning [26], VDSR explicitly increases the network depth (20 layers), and DRCN implicitly grows the number of inference layers by a recursion strategy (16 layers). Recently, Lim et al. [13] refined the conventional SRResNet [27] structure by removing unnecessary modules and achieved distinguished performance, which suggests that deepening the network is conducive to reconstruction. Nevertheless, the dense connection structure [28] attracted attention as well, which encourages feature reuse and strengthens feature propagation. MemNet [12], a long-term memory model, enabled the rearward layers to receive powerful information. Tong et al. [29] proposed SRDenseNet, which applies limited features repeatedly to inner and inter

blocks, to provide richer information for reconstructing high-quality details. Moreover, given the superiority of generative adversarial networks [30], Ledig et al. [27] proposed SRGAN, which is expected to synthesize texture that conforms to human visual perception. Later, the structure of the generators and discriminators were mostly used for upgraded versions of the network, such as [31].

The utilization of different resolutions According to the utilization of different scale information, the existing SR models can be divided into four types [32]. In fact, this classification is based on the location of the upsampling layer. Foremost, pre-upsampling structures, such as those in [7, 9, 25], are shown in Fig. 1a. The input LR images are interpolated to the target scale first, and feature learning is carried out on the target scale for reconstruction. Then, the post-upsampling process is the opposite, as [29] shows in Fig. 1b, which extracts features at the original LR space and finally upsamples the deep features to the target scale for SR reconstruction. Third, progressive upsampling structures, such as in [14], are shown in Fig. 1c. When coping with larger-scale factors, such as ×4, ×8, the pyramid network enlarges the feature scale step by step. Finally, for the iterative up-and-down sampling structure shown in Fig. 1d, the representative work was provided by [15, 16]. The principle of this method is to implement up-and-down sampling iteratively on LR and SR scales and correct features through an error feedback mechanism within each projection block.

Attention mechanism The attention mechanism is a plug-and-play module that can increase the flexibility of the network and provide guidance for resource allocation. Zhang et al. [18] and Qin et al. [33] proposed the RCAN and DCAN models separately, which incorporate a channel attention mechanism with a residual network and a dense connection network, respectively. The similarity between them lies in that the subsequent information is differentiated according to hierarchical channel importance. With further research, Dai et al. [20] proposed SOCA which employs a second-order channel attention method to conduct deeper feature correlation learning. Zhang et al. [34] proposed a nonlocal attention method to capture the long-term dependence between pixels, which makes the pixels with remote spatial positions but strong semantic relevance play a role as well. To weigh spatial information, in CSAR [19], researchers combined spatial attention with channel attention to obtain semantic cues from both spatial and channel perspectives.

Compared with the information that comes from a single-scale and gradual recovery procedure, the cross-scale scheme could refine features by patch similarity and generate reliable representations with strong position sensitivity, especially in high-frequency details. Although some pioneering works have been devoted to utilizing cross-scale, such as [14, 15], they are cascading architectures in general, and across-scale SISR in parallel multi-branches networks holds great potential because this perspective has been little explored. Based on this situation, we propose the multi-branch multi-resolution cross-projection network (MBMR-Net) to explore the correlation between multiple resolutions.

3 Method

The purpose of this paper is to craft an end-to-end trainable SR reconstruction paradigm that can generate SR images close to the genuine images. In this section, we present the multi-branch multi-resolution cross-projection network (MBMR-Net) and fully elaborate its details. Moreover, the multi-resolution feature fusion block (including multiple parallel

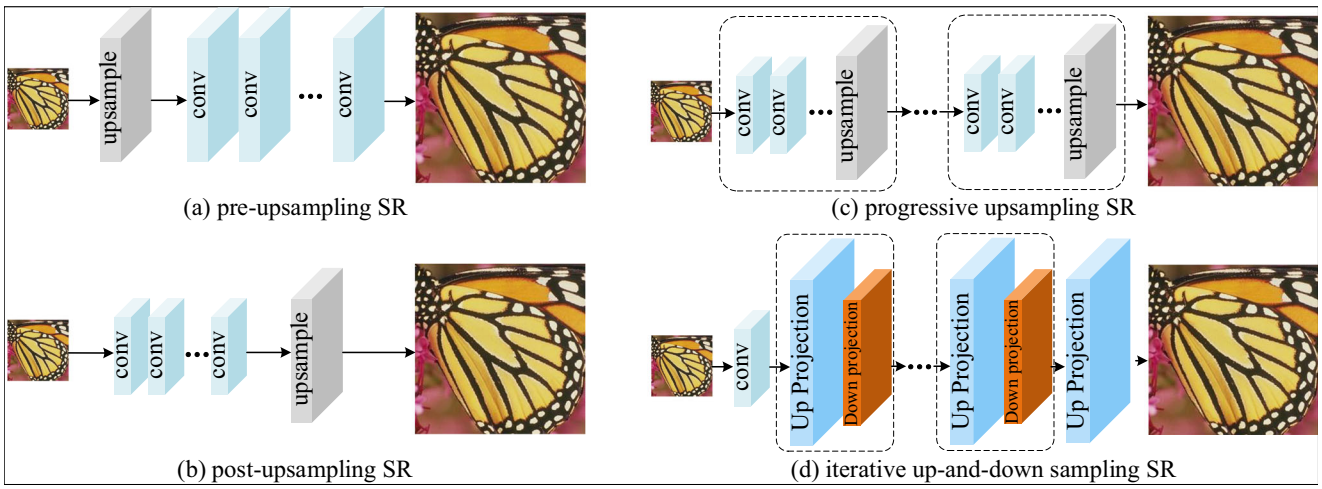

Fig. 1 The utilization of different resolutions. According to the location of upsampling layers, we enumerate four widely used frameworks in the SR task: pre-upsampling, post-upsampling, progressive upsampling, and iterative up-and-down sampling models

branches cross-projection process and an attention unit) and the loss function are discussed successively in this section.

3.1 Network architecture

Depicted as Fig. 2, MBMR-Net can be divided into shallow feature extraction, deep feature extraction, and reconstruction. Let $\boldsymbol{I}^{LR}$, $\boldsymbol{I}^{HR}$ and $\boldsymbol{I}^{SR}$ denote LR, HR, and SR images, respectively. Shallow feature extraction is performed on the input $\boldsymbol{I}^{LR}$ through two convolutional layers, with 128 and 64 output channels. The 3 × 3 convolution converts the input to feature space from image space, and the 1 × 1 convolution reduces the 128 channels to 64. This process is expressed as Eq. (1)

$$\boldsymbol{F}_{SF} = f_{SF}\left(\boldsymbol{I}^{LR}\right) \tag{1}$$

where $f_{SF}(\cdot)$ denotes shallow feature extraction and $\boldsymbol{F}_{SF}$ represents the extracted shallow features. Then, $\boldsymbol{F}_{SF}$ is input to the next stage for deep feature extraction as Eq. (2)

$$\boldsymbol{F}_{DF} = f_{DF}\left(\boldsymbol{F}_{SF}\right) \tag{2}$$

where $f_{DF}(\cdot)$ denotes deep feature extraction, and $\boldsymbol{F}_{DF}$ stands for deep features. Finally, the reconstruction layer tackles $\boldsymbol{F}_{DF}$ according to spatial resolution, the result on the target scale is the reconstructed $\boldsymbol{I}^{SR}$, and the results from other branches will be used to calculate the loss. The reconstruction process $f_{rec}(\cdot)$ can be described as Eq. (3)

$$\boldsymbol{I}^{SR}, \boldsymbol{I}^{SR}_{l-1}, \cdots, \boldsymbol{I}^{SR}_{-1} = f_{rec}\left(\boldsymbol{F}_{DF}\right) \tag{3}$$

and by filtering the outputs from other branches, the reconstruction procedure can be summarized as Eq. (4)

$$\boldsymbol{I}^{SR} = f_{rec}\left(f_{DF}\left(f_{SF}\left(\boldsymbol{I}^{LR}\right)\right)\right) = f_{MBMR}\left(\boldsymbol{I}^{LR}\right) \tag{4}$$

where $f_{MBMR}(\cdot)$ represents the proposed MBMR-Net.

We customize a loss function l_{MBMR} for network training, and the details are described in Section 3.3. Given a training set containing N image pairs $\left\{\boldsymbol{I}_i^{LR}, \boldsymbol{I}_i^{HR}\right\}_{i=1}^{N}$, the training goal is to minimize the discrepancy between $\boldsymbol{I}^{SR}$ and $\boldsymbol{I}^{HR}$, that is, minimize the loss function. For more training details, we provide a detailed description in Section 4.1.

3.2 Multi-resolution feature fusion block (MRFFB)

The multi-resolution feature fusion block (MRFFB), which performs multi-resolution feature extraction, exchange, and fusion, is the core component of deep feature extraction. To prepare for multiple parallel streams, the preset block (PB) is carried out first to generate multiple resolution branches. Then, several MRFFBs are connected in series for depth feature extraction and fusion. The MRFFB can be divided into two steps, specifically, the multiple parallel branches cross-projection and the attention unit (AU). The former is used to exchange information, and the latter is in charge of feature importance learning for better expression ability. In addition, we introduce residual connections on each branch within the MRFFB to facilitate stable training.

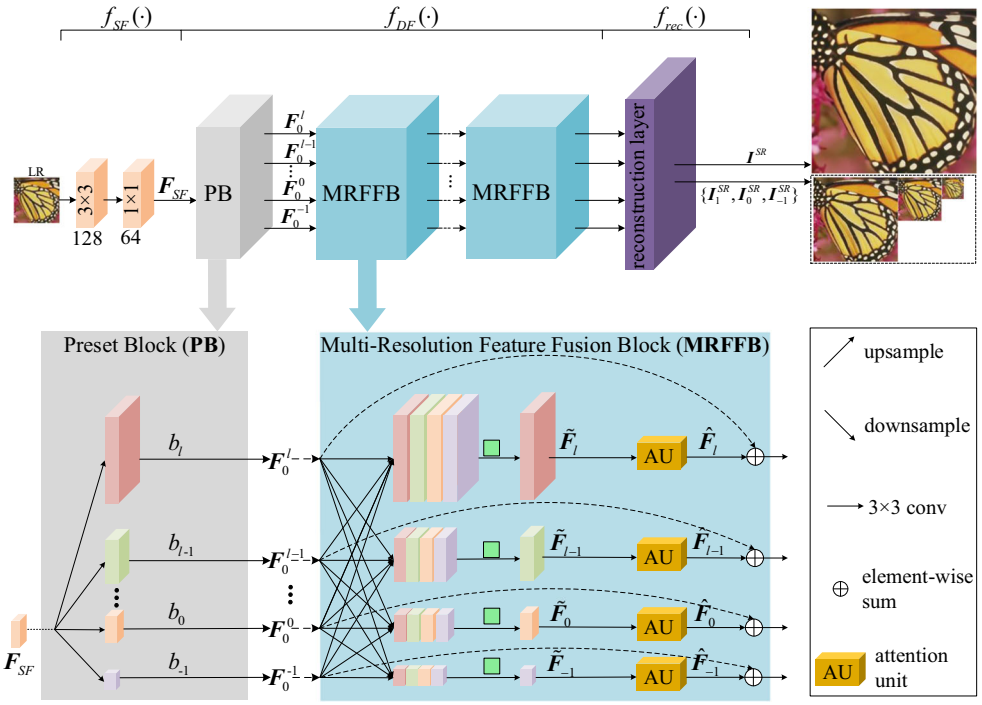

Fig. 2 The architecture of the proposed multi-branches multi-resolution cross-projection network (MBMR-Net)

3.2.1 Multiple parallel branches cross-projection

Inspired by [35], multi-resolution parallel and repeated exchange of information across resolutions makes the resulting representation semantically strong and more precise. Therefore, we design a distinctive multi-resolution fusion module, as shown in the blue background in Fig. 2. Note that our model works only if the scale factor s = 2^l, $l \in N^+$. First, PB transforms the shallow feature $\boldsymbol{F}_{SF}$ into $l + 2$ branches, which are denoted as $b_{-1}, b_0, \cdots, b_{l-1}, b_l$, and the corresponding scales are $2^{-1}, 2^0, \cdots, 2^{l-1}, 2^l$ times $\boldsymbol{F}_{SF}$, and the output channels for each branch are 64. Actually, the size of $\boldsymbol{F}_{SF}$ is the same as b_0, the transformation inside PB is shown in the gray background in Fig. 2.

The inputs of the first MRFFB are the outputs of PB. Then, the features of each branch are subject to a process similar to PB called cross-projection, which can be decomposed as shown in Fig. 3. The multiple transformed representations with the same scale are concatenated directly, and a bottleneck layer (i.e., a 1 × 1 convolutional layer) is utilized to fuse the information from different resolution spaces and maintain the output dimension at 64. The formula is described as Eq. (5)

$$\widetilde{\boldsymbol{F}_k} = \mathcal{C}\left(\left[\boldsymbol{F}_{-1}^k, \boldsymbol{F}_0^k, \cdots, \boldsymbol{F}_{l-1}^k, \boldsymbol{F}_l^k\right]\right) \tag{5}$$

where $\boldsymbol{F}_i^k, i = -1, 0, \cdots, l$ denotes the feature on b_k converted from b_i, $[\cdots, \cdots]$ denotes the concatenation operation, $\mathcal{C}(\cdot)$ represents 1 × 1 convolution, and $\widetilde{\boldsymbol{F}_k}$ refers to features on b_k generated by multiple parallel branches cross-projection.

3.2.2 Attention unit (AU)

The previous multiple branches cross-projection operation generates generous features. There is no doubt that it contains redundant information. In addition, the contribution of features to reconstruction is different. Generally, features expressing image structure are expected to acquire a greater weight to highlight details while giving content information a relatively small weight. Therefore, we utilize an attention unit composed of spatial attention and channel attention for the information importance hierarchy on each branch, and the scheme is shown in Fig. 4.

Spatial attention enables network to learn the relationships among spatial locations [17]. For the $C \times H \times W$ input map, we perform two 3 × 3 convolutions in succession with the same channel reduction ratio r. The result, a $1 \times H \times W$ map, is sent to pass through the nonlinear mapping of the sigmoid function, and the range of features are transformed to [0,1] owing to its gate mechanism. The spatial attention can be described as Eq. (6)

$$\omega_{SA} = \sigma\left(\boldsymbol{W}_2 \cdot \delta\left(\boldsymbol{W}_1 \cdot \widetilde{\boldsymbol{F}_k}\right)\right) \tag{6}$$

where ω_{SA} denotes the spatial attention map, $\delta(\cdot)$ and $\sigma(\cdot)$ denote the ReLU and sigmoid functions, respectively, and $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{R}^{r\mathrm{C}\times 3\times 3}$ are the weights of the convolution operation. We adjust the input features according to the obtained spatial location importance cues as Eq. (7)

$$\boldsymbol{F}_{SA} = \widetilde{\boldsymbol{F}_k} \otimes \omega_{SA} \tag{7}$$

where $\otimes$ denotes the elementwise product and $\boldsymbol{F}_{SA}$ stands for the feature whose spatial position has been adjusted.

Spatial attention focuses on spatial characteristics while ignoring channel knowledge. As a supplement, we use the second-order channel attention scheme [20] because higher-order statistics are more informative for SR reconstruction, which has been proven by [36]. The second-order channel attention adopts global covariance pooling. The input feature $\widetilde{\boldsymbol{F}_k}$ is resized to a matrix $\boldsymbol{X} = C \times HW$ and the relevant covariance matrix is calculated through Eq. (8)

$$\boldsymbol{\Sigma} = \boldsymbol{X}\bar{\boldsymbol{I}}\boldsymbol{X}^T, \bar{\boldsymbol{I}} = \frac{1}{HW}\left(\boldsymbol{I} - \frac{1}{HW}\boldsymbol{1}\right) \tag{8}$$

where $\boldsymbol{I}$ is an identity matrix, $\boldsymbol{1}$ is a matrix with all elements 1, both sizes are $HW \times HW$, and $\boldsymbol{\Sigma}$ denotes a covariance matrix whose size is $C \times C$. As shown in [36–38], covariance normalization is necessary. First, we conduct eigenvalue

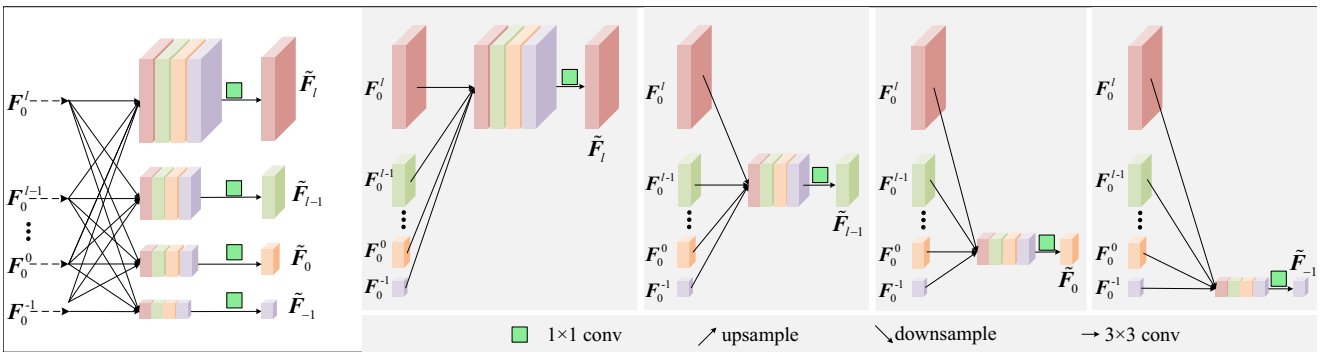

Fig. 3 The process of multiple parallel branch cross-projection for information exchange and fusion

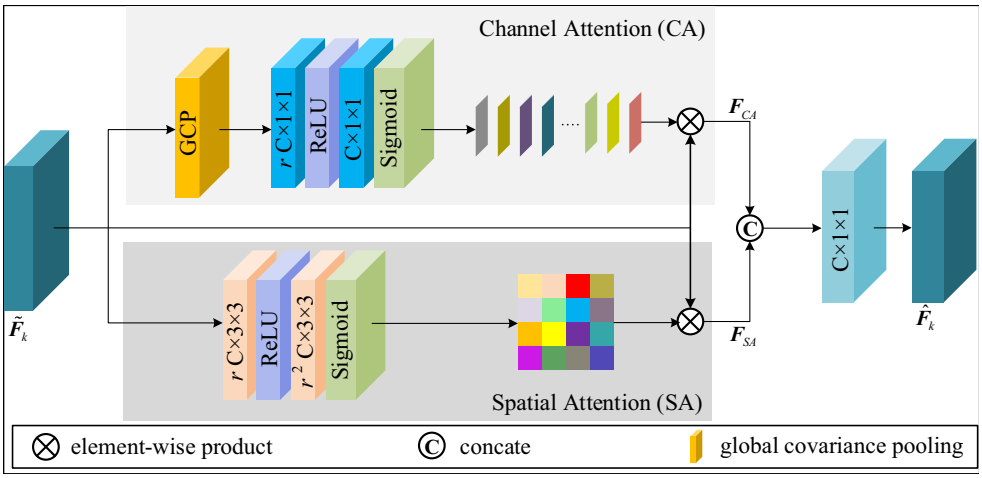

Fig. 4 Schematic diagram of the attention unit (AU) used in this paper, which is composed of spatial attention (SA) and channel attention (CA) for acquiring channel and spatial cues

decomposition for this symmetric semidefinite matrix as Eq. (9)

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \tag{9}$$

Based on the result of decomposition, the importance of eigenvectors in $\boldsymbol{U}$ is described by the corresponding eigenvalues in $\boldsymbol{\Lambda}$, and the covariance normalization is completed by Eq. (10)

$$\widehat{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{\alpha} = \boldsymbol{U}\boldsymbol{\Lambda}^{\alpha}\boldsymbol{U}^T \tag{10}$$

where $\alpha \in R^+$, if $\alpha = 1$, there is no normalization; when $\alpha < 1$, we stretch the eigenvector with an eigenvalue greater than 1 and shrink the counterpart with an eigenvalue smaller than 1. Following [20, 36], we also set $\alpha = 0.5$ for more discriminative representations. We decompose $\widehat{\boldsymbol{Y}}$ into column vectors as $[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_C]$, where $\boldsymbol{y}_i \in \mathcal{R}^{C\times 1}, i = 1, 2, \ldots, C$. Furthermore, channel-wise statistics $\mathbf{z} \in \mathcal{R}^{C\times 1}$ can be computed by Eq. (11)

$$z_c = H_{GCP}(\boldsymbol{y}_c) = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{y}_c(i) \tag{11}$$

where $H_{GCP}(\cdot)$ denotes global covariance pooling. The result $\mathbf{z}$ is passed through a 1×1 convolutional layer with a channel reduction ratio r, and then a channel upsampling layer changes the channel number back to C. Similarly, ReLU and sigmoid functions are utilized to express nonlinear relationship as Eq. (12)

$$\omega_{CA} = \sigma(\boldsymbol{W}_U \cdot \delta(\boldsymbol{W}_D \cdot \boldsymbol{z})) \tag{12}$$

where ω_{CA} denotes the channel attention map and $\delta(\cdot)$ and $\sigma(\cdot)$ are the same as those in Eq. (6); $\boldsymbol{W}_D$ and $\boldsymbol{W}_U$ are the weights in the channel down/upsampling process. Thus, the importance hierarchy of channels is distinguished, and the element-wise product is performed with the channel attention mask through Eq. (13)

$$\boldsymbol{F}_{CA} = \widetilde{\boldsymbol{F}_k} \otimes \omega_{CA} \tag{13}$$

where $\boldsymbol{F}_{CA}$ stands for the feature weighted by channel information. Finally, we concatenate the results of spatial attention and channel attention and deliver them to a 1×1 convolutional layer for information integration and dimension reduction, as shown in Eq. (14)

$$\widehat{\boldsymbol{F}_k} = \mathcal{C}([\boldsymbol{F}_{SA}, \boldsymbol{F}_{CA}]) \tag{14}$$

3.3 Loss function

Given a training set that contains N image pairs represented as $\{\boldsymbol{I}_i^{LR}, \boldsymbol{I}_i^{HR}\}_{i=1}^{N}$, we design the loss function from two aspects. Theoretically, we should strengthen the details while ensuring the content information. In terms of the characteristics of the network, we should consider the optimization of each branch.

Content loss On the target scale branch, the restored SR images should be close to the HR images. Pixel loss has been employed to measure the pixelwise difference between two images. Considering the inferior nature of l_2 that imposes a penalty on larger errors and accepts small errors to some extent [39], the reconstructed image encounters oversmoothing. Hence, we choose l_1 loss to address the discrepancy between the super-resolved image and its HR counterpart as Eq. (15)

$$l_1^{SR} = \frac{1}{N}\sum_{i=1}^{N} \left\| \boldsymbol{I}_i^{SR} - \boldsymbol{I}_i^{HR} \right\|_1 \tag{15}$$

Detail loss Inspired by [40, 41], optimizing the products of the intermediate process can also play an auxiliary role to facilitate target reconstruction. This reminds us to pay attention to the optimization of other branches. The restored images $\boldsymbol{I}_{-1}^{SR}$, $\boldsymbol{I}_0^{SR}, \cdots, \boldsymbol{I}_{l-1}^{SR}$ from b_{-1}, b_0, $\cdots$, b_{l-1} are relatively low-

resolution for $\boldsymbol{I}^{SR}$ from b_l. Due to the lack of clear images on other branches, we reshape $\boldsymbol{I}^{HR}$ to the corresponding size of each branch by bicubic with different sampling factors and mark the reshaped images as $\boldsymbol{I}_{-1}^{HR\downarrow}, \boldsymbol{I}_{0}^{HR\downarrow}, \cdots, \boldsymbol{I}_{l-1}^{HR\downarrow}$. In contrast, we exploit only the detailed components at multiple scales with the intent of providing more detailed references for reconstruction instead of comparing the differences of the whole image. The Sobel operator is applied to calculate the gradient maps of reconstructed images $\boldsymbol{I}_{-1}^{SR}, \boldsymbol{I}_{0}^{SR}, \cdots, \boldsymbol{I}_{l-1}^{SR}, \boldsymbol{I}^{SR}$ and reshaped reference images $\boldsymbol{I}_{-1}^{HR\downarrow}, \boldsymbol{I}_{0}^{HR\downarrow}, \cdots, \boldsymbol{I}_{l-1}^{HR\downarrow}, \boldsymbol{I}^{HR}$ as Eq. (16)

$$G_{i,j} = \sqrt{\left[\nabla_x \boldsymbol{I}(i,j)\right]^2 + \left[\nabla_y \boldsymbol{I}(i,j)\right]^2} \tag{16}$$

where $\nabla_x \boldsymbol{I}(i, j)$ and $\nabla_y \boldsymbol{I}(i, j)$ denote the gradients along the horizontal and vertical directions at pixel (i, j) in image $\boldsymbol{I}$, respectively, and $G_{i,j}$ represents the value of the gradient map at position (i, j), normalizing $G_{i,j}$ as Eq. (17)

$$\boldsymbol{G}_{i,j} = \frac{G_{i,j} - min\left(G_{i,j}\right)}{max\left(G_{i,j}\right) - min\left(G_{i,j}\right)} \tag{17}$$

where $min\left(G_{i,j}\right)$ and $max\left(G_{i,j}\right)$ denote the minimum and maximum in G. The results are expressed as $\boldsymbol{G}_{-1}^{SR}, \boldsymbol{G}_{0}^{SR}, \cdots, \boldsymbol{G}_{l-1}^{SR}, \boldsymbol{G}^{SR}$ and $\boldsymbol{G}_{-1}^{HR\downarrow}, \boldsymbol{G}_{0}^{HR\downarrow}, \cdots, \boldsymbol{G}_{l-1}^{HR\downarrow}, \boldsymbol{G}^{HR}$. Finally, the gradient loss l_1^G shown as Eq. (18)

$$l_1^G = l_{target}^G + l_{others}^G = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_l * \left\|\boldsymbol{G}_i^{SR} - \boldsymbol{G}_i^{HR}\right\|_1 + \sum_{j=-1}^{l-1}\lambda_j * \left\|\boldsymbol{G}_{i,j}^{SR} - \boldsymbol{G}_{i,j}^{HR\downarrow}\right\|_1\right) \tag{18}$$

The former l_{target}^G is exploited to depict the high-frequency differences between I^{SR} and I^{HR}, where λ_l is the tradeoff parameter. The latter l_{others}^G imitates cycle loss [40, 41], and the gradient maps of reshaped images are compared with their restored counterparts. If the difference in the comparison is small enough, the reconstruction process from the current branch to the target scale may be simulated approximately through the reverse process of reshaping. In Eq. (19), λ_j, $j = -1, 0, \cdots, l-1$ is the tradeoff parameter used to distinguish the contribution of different branches. Ultimately, the loss function of the network consists of l_1^{SR} and l_1^G, shown as Eq. (19)

$$l_{MBMR} = l_1^{SR} + l_1^G \tag{19}$$

4 Experiments

In this section, we demonstrate the settings of the experiment first, including the training set, test sets, and other implementation details. Second, we analyze the effectiveness of each component according to ablation experiments. Finally, we compare the proposed model with state-of-the-arts on subjective and objective aspects.

4.1 Implementation details

Following [13, 14], we train our model with the training set of the DIV2K dataset from the NTIRE challenge [42], which contains 800 LR-HR pairs. Furthermore, we conduct horizontal flipping and random rotation (including 90 ° , 180 ° 270°) on training images for data augmentation. In each minibatch, 8 LR color patches with a size of 48 × 48 are provided as inputs. For testing, we evaluate our model on 5 widely used standard benchmark datasets including Set5 [43], Set14 [44], B100 [45], Urban100 [46] and Manga109 [47], with PSNR and SSIM. We select the Adam optimizer [48] during training and initialize its parameters as $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The learning rate is initialized to 10^{-4} and halved every 100 epochs. Our proposed model has been implemented using PyTorch and on an NVIDIA 1080Ti GPU.

In the proposed network, there are 7 MRFFBs in the deep feature extraction stage. To reduce the difficulty of training, we use the parameter sharing strategy to ease the training. Inner PB and the MRFFBs, pixel-shuffle [49] is adopted to obtain different upsampling factors when the resolution of the target branch is greater than the current resolution. Standard 3 × 3 convolution is utilized to extract features at the current resolution. For downsampling, we employ stride convolution to obtain a relatively reliable expression at a smaller resolution, and the kernel size and stride vary with the downsampling factors. Specifically, when the downsampling factors are 2, 4, 8, and 16, the corresponding strides are 2, 4, 8, and 16, the kernel size is 4 × 4, 8 × 8, 12 × 12, and 20 × 20, and proper padding should be carried out if necessary. The reduction ratio r in the attention unit is set to 0.125. According to the statement in Section 3.2.1, the branches that exist in the model vary with the recovery task. As a part of a total loss, the hyperparameter $\lambda_j = -1, 0, \cdots, l$ in Eq. (19) would affect the performance of the model. We adjust the parameters of different tasks based on the settings of [50], $\lambda_{-1,\,0,\,1} = 0.05,0.1,0.2$ for task s = 2, $\lambda_{-1,\,0,\,1,\,2} = 0.05,0.1,0.2,0.3$ for task s = 4, and $\lambda_{-1,\,0,\,1,\,2,\,3} = 0.05,0.1,0.2,0.3,0.4$ for target scale s = 8.

4.2 Effect of multi-resolution feature fusion block

To our knowledge, MBMR-Net is the first attempt to apply a parallel framework to the SISR task. As the core component of the network, we design a series of experiments to prove the effectiveness of the multi-resolution fusion scheme. For a fair comparison, all evaluations are based on the same configurations and the results on Set5 (×4) are shown in Table 1.

Table 1 Ablation experiments on the effectiveness of multiple parallel streams measured with PSNR/SSIM on Set5 (×4)

branch l algorithm	−1	0	1	2	PSNR/SSIM
MBMR-Net-0	✗	✓	✗	✗	31.274/0.8783
MBMR-Net-01	✗	✓	✓	✗	32.025/0.8898
MBMR-Net-02	✗	✓	✗	✓	32.446/0.8946
MBMR-Net-012	✗	✓	✓	✓	32.471/0.8951
MBMR-Net	✓	✓	✓	✓	32.484/0.8959

Each row represents an experiment, ✓ means that this branch is used in the current experiment, and unused branches are represented by ✗

Multiple parallel streams We added relevant branches progressively to confirm that each branch plays a positive role in SISR. First, MBMR-Net-0 only utilizes the depth features on b_0. From the perspective of network structure, the framework is a post-upsampling process and is analogous to EDSR [13]. Although the PSNR score of 31.274 dB is far lower than 32.46 dB of EDSR, the inferior results are understandable because the depth of the network (40 layers) is far less than that of EDSR (70 layers). Second, there are two distinct cases when feature fusion is carried out on two branches, represented as MBMR-Net-01 and MBMR-Net-02. MBMR-Net-01 retains b_0 and b_1 in the network, which can be regarded as an enhanced LapSRN [14]. The low-resolution representations do not receive information from high-resolution representations in LapSRN, but we repeatedly fuse the representations from two-resolution streams throughout the process. The result of 32.03 dB shows the superiority of iterative information exchange, while LapSRN scored 31.74 dB. The other case, MBMR-Net-02, contains the features from b_0 and b_l (i.e., b_2 when s = 4). Although DBPN [15] also utilizes the information from the LR scale and HR scale, their utilization strategies are different. The DBPN mines information in series and scores at 32.44 dB, while MBMR-Net-02 obtains 32.45 dB in parallel. It can be seen from these two cases that both the intermediate resolution and target resolution information are helpful for reconstruction. Enlightened by the experimental results, we further studied the MBMR-Net-012 algorithm by combining three parallel streams b_0, b_1 and b_2, and the score was improved to 32.47 dB. Finally, considering that b_0 enjoys the benefit of bottom-up for revising the low-resolution representations, we introduce b_{-1} to the network. Thus, the MBMR-Net is formed and obtains the best score of 32.48 dB, which strongly proves the efficient and stable performance of the multi-branches architecture.

Attention unit To certify the effectiveness of the attention mechanism, we design experiments, and the results on Set5 (×4) are elaborated in Table 2. We apply CA or SA separately respecting that the method of information fusion in the multiple parallel branches cross-projection process is ordinary and crude. Compared with the basic framework without an attention mechanism, which scored 32.484 dB, the restoration effect was improved to 32.527 dB/32.532 dB with the action of CA/SA, which illuminates that CA/SA is conducive to improving performance. Through this observation, we can determine that CA/SA does play a role in channel/spatial location. What will happen if CA and SA are used together? Furthermore, the cooperation of CA and SA (i.e., attention unit, AU) achieves a better result of 32.589 dB, as shown in the last column, which implies that the combination of CA and SA contributes to reconstruction because of their complementary functions.

4.3 Effect of the loss function

To study the impact of losses on our model, we train MBMR-Net with different losses including l_1, l_2, $l_1 + \lambda * l^G_{target}$, which represent the weighted sum of the l_1 loss and the gradient loss on b_l, and the devised l_{MBMR}. The convergence processes on Set5 (×4) are depicted in Fig. 5. From the outcomes, the gap between l_1 and l_2 is small, but l_2 is unstable and recovers slowly after falling into local optimization during the convergence process, and l_1 performs better in this regard. In addition, l_1 is slightly inferior to $l_1 + \lambda * l^G_{target}$, which indicates that it is recommended to treat low-frequency and high-frequency information differently in SISR tasks. Compared with $l_1 + \lambda * l^G_{target}$, l_{MBMR} proposed according to the network characteristics obtains the best performance. The whole training process is relatively stable and gets the highest score in the evaluation.

Figure 6 shows the quantitative and qualitative results of Set14 "comic.png" (×4), where the upper row shows the recovered details and the bottom row shows the restored SR results. It can be seen in the first row that the image restored by our method contains more details, and our method works well with the regions that have sharper details. However, it is

Table 2 The results of the investigation of two attention strategies (CA, SA, and the combination of CA and SA) measured with PSNR/SSIM on Set5 (×4)

Attention Unit	CA	✗	✓	✗	✓
	SA	✗	✗	✓	✓
PSNR/SSIM on Set5 ×4		32.484/0.8959	32.527/0.8968	32.532/0.8972	32.589/0.8997

Each column represents an experiment, ✓ means it is used in the current experiment, and the unused is represented by ✗

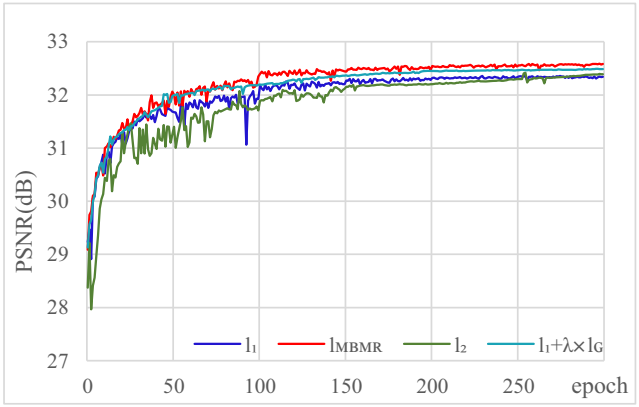

Fig. 5 Convergence processes on Set5 (×4) with different loss functions

still not ideal for areas with dense but not sharp details such as "silver headdress" in the "comic.png". From the qualitative results, the comparisons of the second row indicate that l_1 improves the blurred result caused by l_2. Model training with l_{MBMR} further promotes the quality of the restored image, and the visual effect is more realistic to the ground truth than other models.

4.4 Comparison with state-of-the-art approaches

We compare our model on ×2, ×4, ×8 scale factors with some classic and state-of-the-art approaches, including Bicubic, SRCNN [7], FSRCNN [8], VDSR [9], DRCN [25], LapSRN [14], MemNet [12], EDSR [13], DBPN [15], MSRN [51], FilterNet [52], MRFN [53], and SeaNet [50], 13 SISR models in total. The quantitative evaluation results are shown in Table 3. It can be read intuitively from the table that our model is quite competitive. Overall, the classic EDSR, which won first place in the NTIRE 2017 super-resolution challenge [54], remains a powerful approach.

In general, the difficulty of reconstruction has a positive correlation with the scale factor, and PSNR/SSIM decreases with increasing scale factor. MBMR-Net has varied improvements on different scale factors by comparison. It can be easily found that the score of MBMR-Net on the ×2 task is close to that of EDSR on Set5, while MBMR-Net is ahead of the second-place 0.1 dB on the ×4 scale, and the advantage is enlarged to 0.2 dB on the ×8 scale further. The possible reason should be analyzed from the model itself. In fact, the LR images corresponding to the ×2 scale still retain many details. At this time, the advantages of the model have not been fully yielded. While raising the scale, there is less and less information left in LR images, and the superiority of MBMR-Net gradually comes to light. The advantages can be attributed to two aspects through our analysis: (i) our scheme can be considered a compromise between the width and depth of the network. (ii) Multiscale feature transformation brings feature diversity. Based on this finding, we speculate that MBMR-Net may have greater potential in dealing with large-scale tasks, and the fairly good results on other datasets confirm this conclusion.

Recently, SR algorithms have been consistently proposed, such as MSRN [51], FilterNet [52], MRFN [53], and SeaNet [50]. MBMR-Net shows a certain advantage over these methods. Among them, SeaNet noticed edge-assisted image reconstruction and used a modified MSRN as a plug-and-play module. It is worth noting that the multiscale concept between MBMR-Net and MSRN is different. It refers to the size of the convolution kernel in MSRN, while it indicates that the spatial size of the feature map varies from branch to branch in

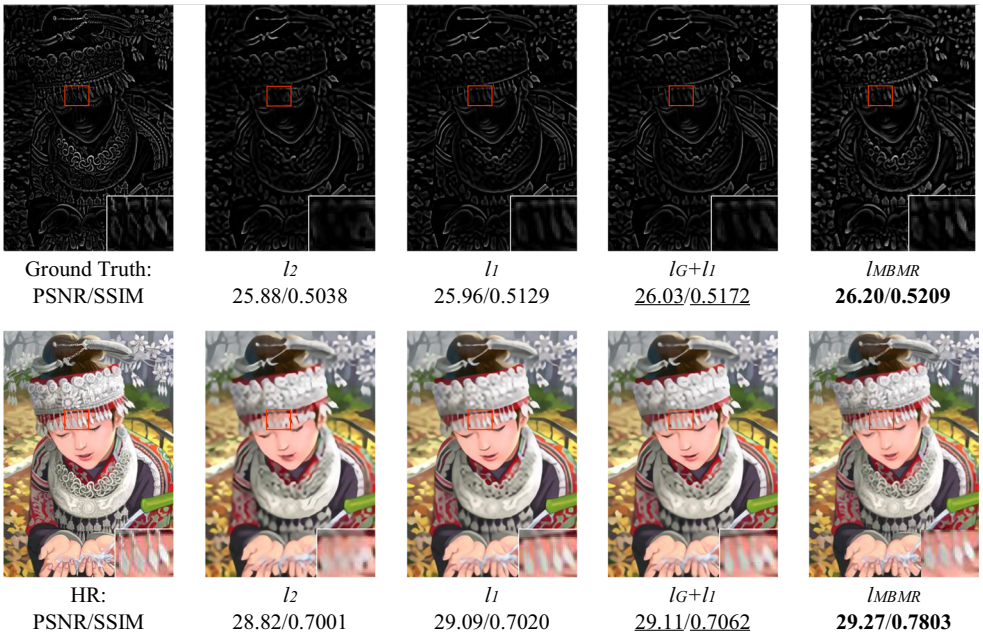

Fig. 6 Impact of different losses on reconstruction performance. The upper row shows the recovered details, the bottom row displays the recovered SR image, PSNR/SSIM are marked at the bottom of the pictures, and the first performance is marked with bold and the second is underlined

Table 3 Comparison with state-of-the-art methods

Method	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	×2	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN [7] (2014)	×2	36.66/0.9542	32.45/0.9067	31.53/0.8920	29.50/0.8946	35.60/0.9663
FSRCNN [8] (2016)	×2	37.05/0.9560	32.66/0.9090	31.53/0.8920	29.88/0.9020	36.67/0.9710
VDSR [9] (2016)	×2	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
DRCN [25] (2016)	×2	37.63/0.9584	33.06/0.9108	31.85/0.8947	30.76/0.9147	37.63/0.9723
LapSRN [14] (2017)	×2	37.52/0.9591	33.08/0.9130	31.80/0.8950	30.41/0.9101	37.27/0.9740
MemNet [12] (2017)	×2	37.78/0.9597	33.28/0.9142	32.08/0.8978	31.31/0.9195	37.72/0.9740
EDSR [13] (2017)	×2	38.11/0.9602	**33.92**/0.9195	32.32/0.9013	32.93/**0.9351**	39.10/0.9773
DBPN [15] (2018)	×2	38.09/0.9600	33.85/0.9190	32.27/0.9000	**33.02**/0.9310	**39.32**/0.9780
MSRN [51] (2018)	×2	38.08/0.9605	33.73/0.9170	32.22/0.9002	32.22/0.9326	38.82/**0.9868**
FilterNet [52] (2019)	×2	37.86/0.9610	33.34/0.9150	32.09/0.8990	31.24/0/9200	—/—
MRFN [53] (2019)	×2	37.98/0.9611	33.41/0.9159	32.14/0.8997	31.45/0.9221	38.29/0.9759
SeaNet [50] (2020)	×2	38.08/0.9609	33.75/0.9190	32.27/0.9008	32.50/0.9318	38.76/0.9774
MBMR-Net (Ours)	×2	**38.19/0.9613**	33.87/**0.9203**	**32.35/0.9073**	32.98/0.9334	39.26/0.9799
Bicubic	×4	28.42/0.8104	26.00/0.7027	25.95/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [7] (2014)	×4	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
FSRCNN [8] (2016)	×4	30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280	27.90/0.8610
VDSR [9] (2016)	×4	31.35/0.8830	28.02/0.7680	27.29//0.7251	25.180.7524	28.83/0.8870
DRCN [25] (2016)	×4	31.56/0.8810	28.15/0.7627	27.24/0.7150	25.15/0.7530	28.98/0.8816
LapSRN [14] (2017)	×4	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
MemNet [12] (2017)	×4	31.74/0.8894	28.26/0.7723	27.40/0.7281	25.50/0.7630	29.42/0.8942
EDSR [13] (2017)	×4	32.46/0.8968	28.80/0.7876	27.71/0.7420	**26.64/0.8033**	**31.02/0.9148**
DBPN [15] (2018)	×4	32.47/0.8870	28.39/0.7780	27.48/0.7330	25.71/0.7720	30.22/0.9020
MSRN [51] (2018)	×4	32.07/0.8903	28.60/0.7751	27.52/0.7273	26.04/0.7896	30.17/0.9034
FilterNet [52] (2019)	×4	31.74/0.8900	28.27/0.7730	27.39/0.7290	25.53/0.7680	—/—
MRFN [53] (2019)	×4	31.90/0.8916	28.31/0.7746	27.43/0.7309	25.46/0.7654	29.57/0.8962
SeaNet [50] (2020)	×4	32.33/0.8970	28.72/0.7855	27.65/0.7388	26.32/0.7942	30.74/0.9129
MBMR-Net (Ours)	×4	**32.59/0.8997**	**28.86/0.7880**	**27.75/0.7431**	26.07/0.7912	30.83/0.9129
Bicubic	×8	24.40/0.6580	23.10/0.5660	23.76/0.5480	20.74/0.5160	21.47/0.6500
SRCNN [7] (2014)	×8	25.33/0.6900	23.76/0.5910	24.13/0.5660	21.29/0.5440	22.46/0.6950
FSRCNN [8] (2016)	×8	20.13/0.5520	19.75/0.4820	24.21/0.5680	21.32/0.5380	22.39/0.6730
VDSR [9] (2016)	×8	25.93/0.7240	24.26/0.6140	24.49/0.5830	21.70/0.5710	23.16/0.7250
DRCN [25] (2016)	×8	25.93/0.6743	24.25/0.5510	24.49/0.5168	21.71/0.5289	23.20/0.6686
LapSRN [14] (2017)	×8	26.15/0.7380	24.35/0.6200	24.54/0.5860	21.81/0.5810	23.39/0.7550
MemNet [12] (2017)	×8	26.16/0.7414	24.38/0.6199	24.58/0.5842	21.89/0.5825	23.56/0.7387
EDSR [13] (2017)	×8	26.43/0.7480	24.39/0.6230	24.60/0.5890	22.01/0.5920	23.97/0.7560
DBPN [15] (2018)	×8	26.96/0.7762	24.91/0.6420	**24.81/0.5985**	22.51/0.6221	**24.69/0.7841**
MSRN [51] (2018)	×8	26.59/0.7254	24.88/0.5961	24.70/0.5410	22.37/0.5977	24.28/0.7517
MBMR-Net (Ours)	×8	**27.02/0.7842**	**24.96/0.6517**	24.74/0.5761	**22.63/0.6251**	24.33/0.7738

The PSNR/SSIM are used as evaluations. The first performance is marked in bold, and the second performance is underlined

MBMR-Net. Comparing the concepts of the two methods, the main reasons for the differences may be as follows: (i) Edge-Net uses a modified MSRN as the structure, whose motivation is to expand the receptive field with convolution kernels of different sizes. There will be a certain error when using the features on the single LR scale for upsampling, which will result in the limited quality of the generated features. MBMR-Net adjusts the feature maps by top-down and bottom-up at the same time to optimize globally and locally. (ii) Edge-Net, used in front of SeaNet, can be regarded as a plug-and-play module for generating high-quality and high-

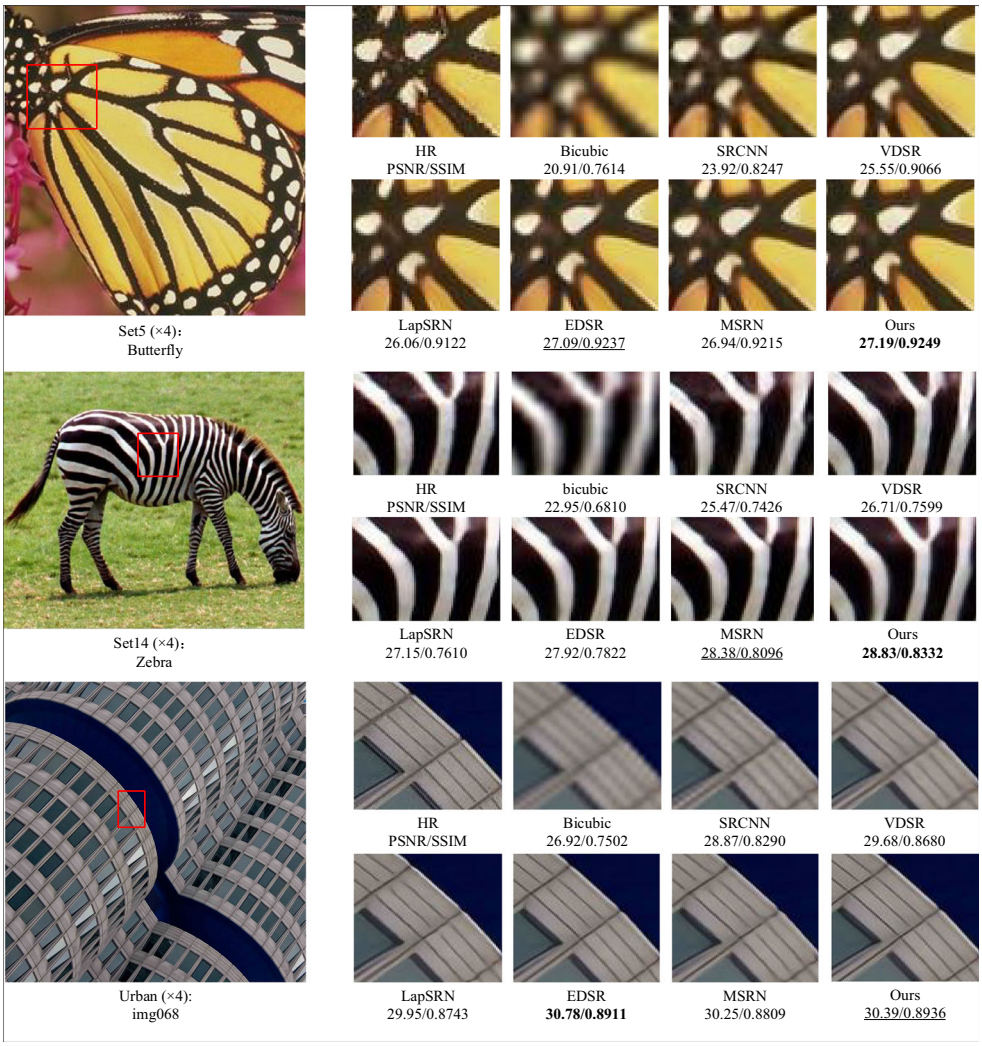

Fig. 7 Intuitive comparison results on the set5, set14, and urban100 datasets. The qualitative results are marked at the bottom of the corresponding images, and the first performance is marked with bold and the second is underlined

frequency features. The enhancement of detail restoration is slackened in the back of the network.

In addition, we show a visual comparison of the ×4 scale from Set5, Set14, Urban100 in turn, and the qualitative results are marked at the bottom of the corresponding images, as shown in Fig. 7. We can observe that MBMR-Net performs better reconstruction than the other methods. Taking "Butterfly" in Set5 as an example, the edges are continuous and clear, which is consistent with the original image. The "cluttered" textures in the HR image are also restored because the multi-resolution fusion strategy grasps the content learning while strengthening the learning of texture; however, these details are directly replaced by content in EDSR and MSRN. Similarly, for "img068" in Urban, the image restored by MBMR-Net is more respectful to the HR image. For the case with good visual effects but poor scores, the possible causes are listed as follows: (i) PSNR only describes the difference of corresponding pixels instead of visual perception, while there is a shortage of accepted perceptual metrics. (ii) Although the high and low frequencies coexist harmoniously in the recovery results evaluated on the three-channel RGB, there may be slight differences in pixel values because the loss function emphasizes the high-frequency part.

Through the above analysis, our network has advantages in large-scale factors. To support this conclusion, we make a visual comparison on the ×8 scale, as shown in Fig. 8. From the fact shown about "img043" from Urban100 and "GarakutayaManta" from Manga109, the images recovered by our model are the best in both subjective feelings and objective scores. MBMR-Net has significantly improved the recovery through the corresponding evaluation results displayed at the bottom of the images. From the perspective of visual perception, MBMR-Net has a particularly outstanding reconstruction effect on clean and sharp edges, such as obvious line segments. Moreover, the results generated by MBMR-Net have the least blur and are most in line with the HR images.

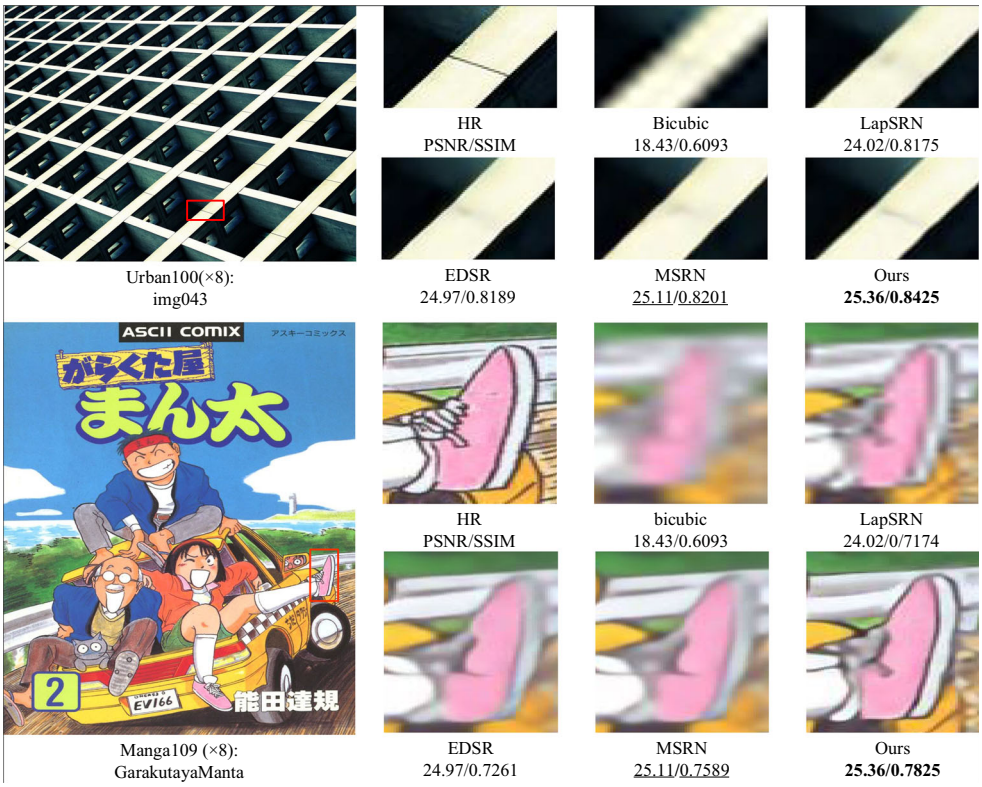

Fig. 8 For visual comparison on the ×8 scale from Urban100 and Manga109. The qualitative results are marked at the bottom of the corresponding images, and the first performance is marked with bold and the second is underlined

4.5 Model analysis

Study about the number of MRFFB As a hyperparameter, the number of MRFFB is bound to affect the performance of the model. We empirically adjust it during model training, and the curve is shown in Fig. 9. It is easy to find that the model does not work to the maximum extent when the amount of MRFFB is small. The performance is continuously improved when the number of modules increases from 5 to 7. The running time reveals that the behavior of increasing modules will lead to a calculation burden. However, continuously increasing the number is counterproductive. The performance of the network will decline slightly when the number increases from 7 to 9. There are two reasons we analyze this in terms of the structural characteristics. One is that a large number of modules result in a large number of parameters and lead to difficult optimization, and the other reason may be that the restored edges are incompatible with the image content. Ultimately, we set the number of MRFFBs to 7 according to the experimental results.

Study of model size We compare the model size of MBMR-Net with other SR methods, and Fig. 10 shows the relationship between parameters and performance on Set5 (×4). The

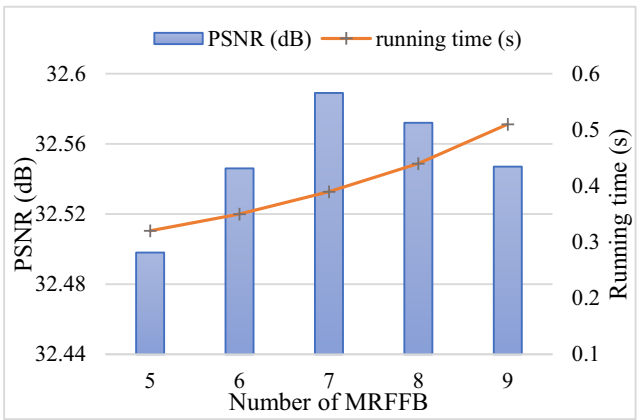

Fig. 9 The study about the number of MRFFB. We evaluate the network performance on Set5 (×4) by PSNR (dB) and running time (s)

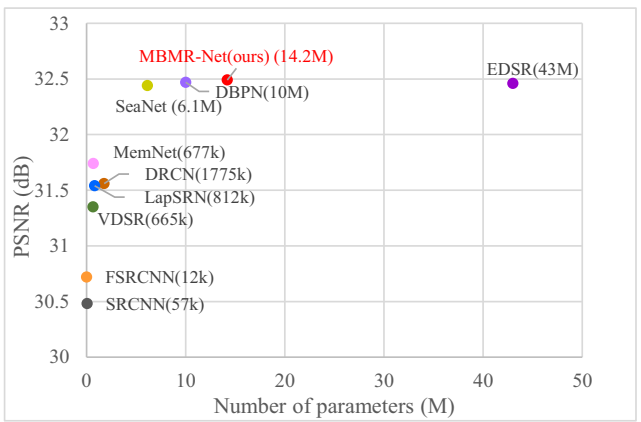

Fig. 10 The comparison of model performance and parameters (Set5 ×4)

improvement of performance is accompanied by an increase in parameters. MBMR-Net significantly improves the performance and slightly exceeds the performance of EDSR with fewer parameters, which is less than one third of EDSR. The reason lies in two aspects: (i) During the repeated process of MRFFB, the parameter sharing generalization is adopted to keep the parameter quantity stable. (ii) Inside MRFFB, the multi-resolution parallel process is similar to group convolution [55], the difference is that group convolution performs regular convolution at the same spatial resolution, while our scheme works at different spatial resolutions. The connection between them suggests that the multi-resolution parallel process enjoys the advantages of group convolution.

Actually, MBMR is superior to most models in terms of parameters and performance. As mentioned above, the parameter sharing strategy brings parameter advantages, which means the generalization ability is sacrificed to a certain extent. This statement is confirmed by the results in the last column of Table 3. The training set DIV2K contains limited scenes which are animals, buildings, and landscapes, while Manga109 is an animation dataset. The results on this dataset can reflect the generalization ability of the model. Although the second-place score indicates that MBMR-Net has strong superiority in generalization compared with most models, but there is still room for improvement. In future research, we will pay more attention to improving the generalization ability.

5 Conclusion

In this paper, we propose a multi-branch multi-resolution cross-projection network (MBMR-Net) to fully explore feature dependencies and detailed information for better single-image super-resolution. Concretely, a multi-resolution feature fusion block (MRFFB) is designed to employ multiple parallel branch cross-projection for feature diversity. We propose an attention unit (AU) to distinguish the importance of features during the process of feature fusion. Its specific implementation is to combine second-order channel attention with spatial attention. For better reconstruction details, we customize a loss function for the network by utilizing high-frequency information while bypassing the low-frequency part. Sufficient experiments on benchmarks are carried out, the results strongly certificate the feasibility of the parallel framework, and the comparison with state-of-the-art methods demonstrates its effectiveness.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Bulat A, Tzimiropoulos G (2018) Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 109–117
2. Chen Y, Tai Y, Liu X, Shen C, Yang J (2018) FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2492–2501
3. Pham CH, Ducournau A, Fablet R, Rousseau F (2017) Brain MRI Super-Resolution using Deep 3D Convolutional Networks. In: IEEE International Symposium on Biomedical Imaging, pp 197–200
4. Lyu Q, Shan H, Steber C, Helis C, Wang G (2020) Multi-Contrast Super-Resolution MRI Through a Progressive Network. IEEE Transactions on Medical Imaging 39(9):2738–2749
5. Rasti P, Uiboupin T, Escalera S, Anbarjafari G (2016) Convolutional Neural Network Super Resolution for Face Recognition in Surveillance Monitoring. In: International Conference on Articulated Motion and Deformable Objects, pp 175–184
6. Xun Y, Zhou P, Wang M (2018) Person Reidentification via Structural Deep Metric Learning. IEEE Transactions on Neural Networks and Learning System 30(10):2987–2998
7. Chao D, Chen CL, He K, Tang X (2014) Learning a Deep Convolutional Network for Image Super-Resolution. In: European Conference on Computer Vision (ECCV), pp 184–199
8. Chao D, Chen CL, Tang X (2016) Accelerating the Super-Resolution Convolutional Neural Network. In: European Conference on Computer Vision (ECCV), pp 391–407
9. Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: IEEE conference on computer vision and pattern recognition, pp 1646–1654
10. Mao XJ, Shen C, Yang YB (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Advances in neural information processing systems (NIPS)
11. Ying T, Jian Y, Liu X (2017) Image Super-Resolution via Deep Recursive Residual Network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2790–2798
12. Tai Y, Yang J, Liu X, Xu C (2017) MemNet: A Persistent Memory Network for Image Restoration. In: IEEE International Conference on Computer Vision (ICCV), pp 4549–4557
13. Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced Deep Residual Networks for Single Image Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1132–1140
14. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5835–5843
15. Haris M, Shakhnarovich G, Ukita NJA (2018) Deep Back-Projection Networks For Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1664–1673
16. Zhu F, Zhao Q (2019) Efficient Single Image Super-Resolution via Hybrid Residual Feature Learning with Compact Back-Projection Network. In: IEEE International Conference on Computer Vision Workshop (ICCVW), pp 2453–2460
17. Jie H, Li S, Gang S, Albanie S Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(8, 2017):2011–2023
18. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: European Conference on Computer Vision, pp 294–310

19. Hu Y, Li J, Huang Y, Gao X (2019) Channel-Wise and Spatial Feature Modulation Network for Single Image Super-Resolution. In: IEEE Transactions on Circuits and Systems for Video Technology, pp 3911–3927
20. Dai T, Cai J, Zhang Y, Xia ST, Zhang L (2019) Second-order Attention Network for Single Image Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 11057–11066
21. Gao Z, Xie J, Wang Q, Li P (2020) Global Second-Order Pooling Convolutional Networks. In: IEEEConference on Computer Vision and Pattern Recognition (CVPR), pp 3019–3028
22. Fe W, Jiang IM, Chen Q, Yang S, Tang X (2017) Residual Attention Network for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6450–6458
23. Li K, Wu Z, Peng KC, Ernst J, Fu Y (2018) Tell Me Where to Look: Guided Attention Inference Network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 9215–9223
24. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
25. Kim J, Lee JK, Lee KM (2016) Deeply-Recursive Convolutional Network for Image Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5016–5023
26. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
27. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z (2016) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 105–114
28. Huang G, Liu Z, Pleiss G, Van DML, Weinberger K (2019) Convolutional networks with dense connectivity. In: IEEE transactions on pattern analysis machine intelligence
29. Tong T, Li G, Liu X, Gao Q (2017) Image Super-Resolution Using Dense Skip Connections. In: IEEE International Conference on Computer Vision (ICCV), pp 4809–4817
30. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS), pp 2672–2680
31. Haris M, Shakhnarovich G, Ukita N (2021) Deep Back-ProjectiNetworks for Single Image Super-Resolution. IEEE Transactions on Pattern Analysis Machine Intelligence 43(12): 4323–4337
32. Wang Z, Chen J, Hoi S (2020) Deep Learning for Image Super-resolution: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10):3365–3387
33. Qin D, Gu X (2019) Deep Residual-Dense Attention Network for Image Super-Resolution. In: International Conference on Neural Information Processing, pp 3–10
34. Y. Zhang, K. Li, K. Li, B. Zhong, Y. Fu, Residual non-local attention networks for image restoration, 2019. https://arxiv.org/pdf/1903.10082v1.pdf
35. Wang J, Sun K, Cheng T, Jiang B, Xiao B (2019) Deep High-Resolution Representation Learning for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10): 3349–3364
36. Wang Q, Li P, Zuo W, Lei Z (2016) RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian with Application to Material Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4433–4441
37. Koniusz P, Fei Y, Gosselin PH, Mikolajczyk K (2017) Higher-order Occurrence Pooling for Bags-of-Words: Visual Concept Detection. IEEE Transactions on Pattern Analysis Machine Intelligence 39(2):313–326
38. Li P, Xie J, Wang Q, Zuo W (2017) Is second-order information helpful for large-scale visual recognition? In: IEEE international conference on computer vision (ICCV), pp 2089–2097
39. Zhao H, Gallo O, Frosio L, Kautz J (2017) Loss Functions for Image Restoration With Neural Networks. IEEE Transactions on Computational Imaging 3(1):47–57
40. Yuan Y, Liu S, Zhang J, Zhang Y, Dong C, Lin L (2018) Unsupervised Image Super-Resolution using Cycle-in-Cycle Generative Adversarial Networks. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 814–823
41. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: IEEE International Conference on Computer Vision (ICCV), pp 2242–2251
42. Agustsson E, Timofte R, NTIRE (2017) Challenge on single image super-resolution: dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), vol 2017, pp 1122–1131
43. Bevilacqua M, Roumy A, Guillemot C, Morel A (2012) Low-Complexity Single Image Super-Resolution Based on Nonnegative Neighbor Embedding. In: British Machine Vision Conference, pp 1–10
44. Zey De R, Elad M, Protter M (2010) On Single Image Scale-Up Using Sparse-Representations. Springer, pp 711–730
45. Martin D, Fowlkes C, Tal D, Malik J (2002) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: IEEE International Conference on Computer Vision (ICCV), pp 416–423
46. Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5197–5206
47. Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based Manga Retrieval using Manga109 Dataset. Multimedia Tools and Applications 76(20):21811–21838
48. Kingma D, Ba JJCS (2014) Adam: a method for stochastic optimization. In: International conference on learning representations (ICLR)
49. Shi W, Caballero J, Huszár F, Totz J, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1874–1883
50. Fang F, Li J, Zeng T (2020) Soft-edge Assisted Network for Single Image Super-Resolution. IEEE Transactions on Image Processing: 4656–4668
51. Li J, Fang F, Mei K, Zhang G (2018) Multi-scale Residual Network for Image Super-Resolution. In: European Conference on Computer Vision (ECCV), pp 527–542
52. Li F, Bai H, Zhao Y (2020) FilterNet: Adaptive Information Filtering Network for Accurate and Fast Image Super-Resolution. IEEE Transactions on Circuits Systems for Video Technology 30(6):1511–1523
53. He Z, Cao Y, Du L, Xu B, Zhuang Y (2019) MRFN: Multi-Receptive-Field Network for Fast and Accurate Single Image Super-Resolution. IEEE Transactions on Multimedia 22(4):1042–1054
54. R. Timofte, E. Agustsson, L.V. Gool, e. al., NTIRE 2017 Challenge on single image super-resolution: methods and results, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1110–1121

Dan Zhang is a postgraduate in the School of Microelectronics and Communication Engineering at Chongqing University, Chongqing, China. Her research interests include image processing, machine learning, and computer vision.

Binglian Zhu received her Ph.D. degree in automatic control from Chongqing University, Chongqing, China, in 1997. She is a Professor at the School of Microelectronics and Communication Engineering, Chongqing University. Her research interests include image processing and communication signal processing.

Yuanhong Zhong received his Ph.D. degree in commutation engineering from Chongqing University, Chongqing, China, in 2011. He is currently an Associate Professor with the School of Microelectronics and Communication Engineering, Chongqing University. His research interests include image processing, machine learning, and computer vision.