Lightweight image super-resolution with expectation-maximization attention mechanism

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1109/TCSVT.2021.3078436

PUBLISHER

Institute of Electrical and Electronics Engineers

VERSION

AM (Accepted Manuscript)

REPOSITORY RECORD

Zhu, Xiangyuan, Kehua Guo, Sheng Ren, Bin Hu, Min Hu, and Hui Fang. 2021. “Lightweight Image Super-resolution with Expectation-maximization Attention Mechanism”. Loughborough University. https://hdl.handle.net/2134/14518851.v1.

Lightweight Image Super-Resolution with Expectation-Maximization Attention Mechanism

Xiangyuan Zhu, Kehua Guo, *Member, IEEE,* Sheng Ren, Bin Hu, Min Hu and Hui Fang

***Abstract*—In recent years, with the rapid development of deep learning, super-resolution methods based on convolutional neural networks (CNNs) have made great progress. However, the parameters and the required consumption of computing resources of these methods are also increasing to the point that such methods are difficult to implement on devices with low computing power. To address this issue, we propose a lightweight single image super-resolution network with an expectation-maximization attention mechanism (EMASRN) for better balancing performance and applicability. Specifically, a progressive multi-scale feature extraction block (PMSFE) is proposed to extract feature maps of different sizes. Furthermore, we propose an HR-size expectation-maximization attention block (HREMAB) that directly captures the long-range dependencies of HR-size feature maps. We also utilize a feedback network to feed the high-level features of each generation into the next generation's shallow network. Compared with the existing lightweight single image super-resolution (SISR) methods, our EMASRN reduces the number of parameters by almost one-third. The experimental results demonstrate the superiority of our EMASRN over state-of-the-art lightweight SISR methods in terms of both quantitative metrics and visual quality. The source code can be downloaded at https://github.com/xyzhu1/EMASRN.**

***Index Terms*—Image super-resolution, lightweight, progressive feature extraction, expectation-maximization attention.**

I. Introduction

SINGLE image super-resolution is intended to convert a given low-resolution (LR) image with coarse details to a corresponding high-resolution (HR) image with better visual quality and refined details. Because SISR can make the processed image have richer details, it is widely used in many applications, such as small object detection [1], surveillance and security [2], medical imaging [3], forensics [4] and astronomical images [5]. SISR is essentially an ill-posed problem, which means that for an LR image, there are countless corresponding HR images instead of a single HR image. Despite the difficulties, SISR is receiving increasing attention from researchers due to its huge academic and industrial value.

Early research on super-resolution mainly focused on traditional methods. For example, the interpolation-based method [6] is the simplest and most efficient method. However, due to the high dependence of this method on adjacent pixels, the processed images have poor performance. Later, to improve the performance of super-resolution, researchers proposed reconstruction-based methods [11], [12]. However, these reconstruction-based methods not only rely heavily on the image prior but also time-consuming.

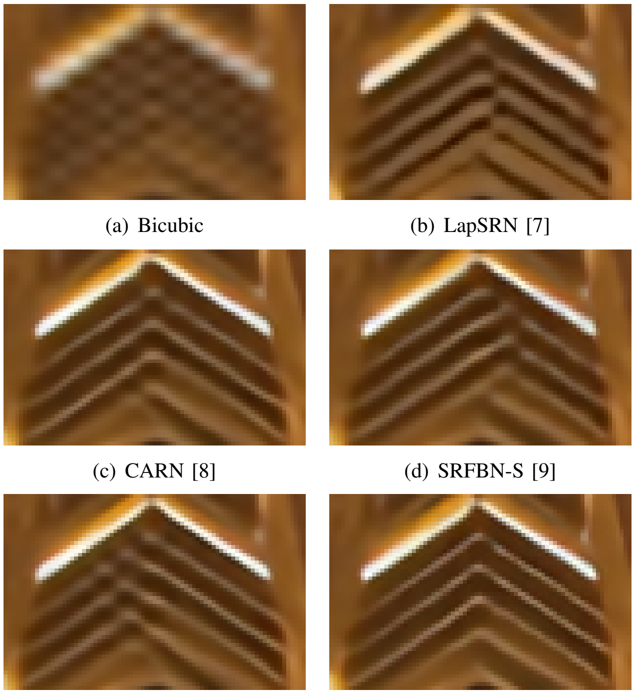
(a) Bicubic
(b) LapSRN [7]
(c) CARN [8]
(d) SRFBN-S [9]
(e) CFSRCNN [10]
(f) EMASRN (ours)

Fig. 1. Lightweight SISR results of different methods. The lines of most methods are seriously merged. Each line generated by our method is more independent and can reflect the details of the original image more accurately.

In recent years, convolutional neural networks (CNNs) have made remarkable achievements in the field of computer vision. Due to the powerful feature extraction ability and high representational capacity of CNNs, CNN-based methods [13]–[18] have surpassed traditional methods by a large margin. Through the large number of parameters and complex network structure, CNN-based methods can learn the mapping between LR and HR images from a large number of training sets. Due to the huge academic and commercial value, recently, an increasing number of CNN-based SISR methods have been proposed. Although these methods continue to make breakthroughs in performance, their parameters and complexity are increasing. This pattern makes the performance growth of existing models depend greatly on an increase in the number of parameters and the elaboration of convolutional

*Corresponding author: Kehua Guo.

Xiangyuan Zhu, Kehua Guo, Sheng Ren, Bin Hu and Min Hu are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: zhuxiangyuan@csu.edu.cn; guokehua@csu.edu.cn; rensheng@csu.edu.cn; hubincsu@csu.edu.cn; mhu1515@csu.edu.cn).

Hui Fang is with the Computer Science Department, Loughborough University, Loughborough, U.K. LE11 3TU (e-mail: H.Fang@lboro.ac.uk).

neural networks. Furthermore, these complex models require tremendous computing resources and memory, which makes it difficult to apply them to real-world applications. For example, inasmuch as mobile devices such as mobile phones have limited computing power and operating memory, complex models are almost impossible to run on them. To solve this problem, many lightweight super-resolution methods have also been proposed [7]–[10]. Some lightweight SISR examples are presented in Fig. 1. These methods reduce the parameters by designing a shallow network structure or by recursive connection but cannot make full use of the expressive power of convolutional neural networks. Although their parameters are reduced compared with the previous methods, this reduction is achieved at the cost of a significant performance sacrifice. Moreover, the parameters of the existing lightweight models are still too large for real applications.

To address the above issues, we propose a lightweight single image super-resolution network EMASRN for better balancing performance against applicability. Each pixel in the image has a certain correlation with other pixels instead of being independent with each other. We propose an HR-size expectation-maximization attention mechanism to capture the long-range dependencies of HR-size feature maps. Compared with other self-attention mechanisms, our method captures the long-range dependencies indirectly by calculating the relationship between all of the pixels in the image and the basis, which makes it possible to perform on the HR-size feature maps. In addition, we propose a progressive multi-scale feature extraction block to extract the feature information of different sizes. Ablation studies are performed to test the effectiveness of the proposed blocks for SR performance improvement. Our EMASRN achieves a good balance between the number of parameters and the performance. Comparison to the existing lightweight SISR methods also shows that our EMASRN achieves state-of-the-art performance.

The major contributions of our work can be summarized as follows:

- We propose an HR-size expectation-maximization attention mechanism to directly capture the long-range dependencies of HR-size images by calculating the relationship between all pixels in the image and the basis. To the best of our knowledge, this paper is the first application of the expectation-maximization attention mechanism in image super-resolution.
- We propose a progressive multi-scale feature extraction block (PMSFE) to extract the feature information of different scales. PMSFE shares the feature information of adjacent scales in a progressive way in the early stage, which can make the fusion of different features more efficient.
- Comparison with the state-of-the-art methods shows that our method not only reduces the parameters by almost one-third but also demonstrates superiority in terms of both quantitative value and visual quality.

The rest of this paper is organized as follows. In Section II, we briefly review the related work. In Section III, we introduce the related preliminaries. In Section IV, we describe the proposed network in detail. In Section V, the experimental results are presented. Finally, we conclude this paper in Section VI.

II. Related Work

A. CNN Based Single Image Super-Resolution

With the rapid development of deep learning in recent years, many CNN-based methods have become the mainstream of SISR. Dong *et al.* [13] were the first to use convolutional neural networks in image super-resolution tasks, and they proposed the single-image super-resolution algorithm SRCNN. Later, with the help of residual learning [19], Kim *et al.* [20] proposed the 20-layer depth model VDSR, which can converge faster during the training process and achieve better results than SRCNN. Lai *et al.* [7] proposed LapSRN, which incorporates the idea of a laplacian pyramid. This model takes the original low-resolution image as input and then uses a progressive method to generate high-resolution images. Yu *et al.* [21] proposed WDSR which demonstrated that models with wider features before ReLU activation have significantly better performance. Yu *et al.* [22] applied state-of-the-art SR techniques to reconstruct CT images and introduced a coarse-to-fine and residual learning idea which is also utilized in our work. Subsequently, Lim *et al.* [23] designed EDSR by reducing unnecessary modules in the convolutional neural network and won the championship of the NTIRE2017 SR Challenge [24]. Li *et al.* [25] used the multi-scale residual network for single image super-resolution and proposed MSRN. In the same year, Haris *et al.* [26] proposed DBPN, which uses the iterative upsampling and downsampling method. DBPN uses an error feedback mechanism, which allows the model to achieve good performance in large-scale super-resolution tasks and won the championship of the NTIRE2018 SR Challenge [27]. Hu *et al.* [28] proposed CSFM to capture more informative features and maintain long-term information for image SR. Qiu *et al.* [29] processed the low-frequency information and high-frequency information of the input image separately and proposed EBRN. More recently, Zuo *et al.* [30] proposed a data-driven approach based on the deep convolutional neural network with global and local residual learning for depth map SR. Mei *et al.* [31] proposed a Cross-Scale Non-Local attention module to find the cross-scale feature correlations within the LR images. Wu *et al.* [32] designed MGAN to exploit the advantages of multi-scale and attention mechanisms in SR tasks.

Although these methods have achieved excellent performance, they have complex model structure with a large number of parameters, thus making them difficult to be used in real situations.

B. CNN Based Lightweight Single Image Super-Resolution

The applications of the SISR model to practical situations have been studied for a long time. Hui *et al.* [33] proposed an information distillation network IDN. In the proposed IDN, a feature extraction block first extracts features from the LR image. Then, multiple information distillation blocks are stacked to progressively distill residual information. Next, a reconstruction block aggregates the obtained HR residual representations to generate the residual image. Finally, an element-wise addition operation is implemented on the residual image and the upsampled LR image to obtain the HR image. Although IDN has reduced parameters compared with the previous method, this reduction is achieved at the cost of a significant performance sacrifice. Later, Ahn *et al.* [8] designed an architecture that implements a cascading mechanism upon a residual network and proposed CARN. The middle parts of CARN were designed based on ResNet. In addition to the ResNet architecture, CARN uses a cascading mechanism at both the local and global levels to incorporate the features from multiple layers. Although CARN has a better performance than IDN, its model is larger, which renders it difficult to apply in practice. Xie *et al.* [34] utilized the residual extracted from the input to predict its counterpart in the corresponding output. Li *et al.* [35] designed adaptive filters to remove the redundant low-frequency information so that the amount of memory consumption and computational cost can be reduced. Tian *et al.* [10] proposed CFSRCNN which cascaded several types of modular blocks to prevent possible training instability and performance degradation. CFSRCNN utilizes a stack of feature extraction blocks to learn the long-path and short-path features and fuses them by expending the effect of the shallower layers to the deeper layers. Although CFSRCNN has slightly fewer parameters than CARN, CFSRCNN has a reduction in PSNR, which also proves that it is difficult to achieve a good balance between parameters and performance.

Compared to these lightweight SISR methods, our model further reduces the number of model parameters significantly via a well-designed architecture. The proposed method achieves a better reconstruction performance with lower complexity, which is confirmed by our experimental results.

C. Attention Mechanism

The attention mechanism is widely used in high level computer vision tasks, such as object detection [36], image classification [37] and image segmentation [38]. The self-attention mechanism originally attracted widespread attention in the field of natural language processing [39]. Wang *et al.* [40] adopted the self-attention mechanism for computer vision tasks for the first time and proposed Non-local. Dai *et al.* [41] utilized a non-locally enhanced residual group structure to capture long-range spatial contextual information and proposed SAN, which obtained state-of-the-art result at that time. However, Non-local block needs to calculate the correlations between all of the pixels in the image, which leads to huge time complexity and space complexity of the method and consumes more computing resources. Li *et al.* [42] reformulated the self-attention mechanism in the manner of an expectation-maximization iteration and proposed expectation-maximization attention networks for semantic segmentation. The proposed Expectation-Maximization Attention (EMA) module is robust to the variance of input and friendly in terms of memory and computation.

Inspired by EMA, we design a lightweight single image super-resolution network EMASRN which can directly capture long-range relations on HR-size images. To the best of our knowledge, this paper is the first to introduce the expectation-maximization attention mechanism into image super-resolution.

III. Preliminaries

Before introducing our proposed method, we first introduce the definition of image super-resolution and review the expectation-maximization algorithm.

A. Definitions of Super-Resolution

The task of super-resolution aims at recovering the LR images into corresponding HR images and gives these images richer texture details and complete contour features. Generally, the LR image I_L can be modeled as the output of the following degradation:

$$I_L = \mathcal{D}\left(I_H; \theta_\delta\right), \tag{1}$$

where $\mathcal{D}(\cdot)$ represents the corresponding degradation mapping function, I_H is the HR image, and θ_δ stands for the parameters of the degradation process, such as the size of blur kernel and noise level. In practice, only LR images are provided, and the degradation process or the degradation parameters are unknown. Researchers are required to recover an approximation $\hat{I}_H$ of the ground-truth image I_H as:

$$\hat{I}_H = U\left(I_L; \theta_\gamma\right), \tag{2}$$

where $U(\cdot)$ stands for the super-resolution model and θ_γ denotes the parameters of $U(\cdot)$. The degradation mapping function is unknown and can be quite complex which makes SISR essentially an ill-posed problem. In most situations [8], [10], [25], the bicubic interpolation function is used as the degradation mapping. In this case, the degradation process can be formulated as:

$$\mathcal{D}\left(I_H; \theta_\delta\right) = \left(I_H\right) \downarrow_s , \tag{3}$$

where $\downarrow_s$ represents the downsampling operation with the scaling factor s.

Finally, the objective function of supervised super-resolution tasks can be formulated as follows:

$$\hat{\theta_\gamma} = \underset{\theta_\gamma}{\arg\min} \mathcal{L}\left(\hat{I}_H, I_H\right), \tag{4}$$

where $\mathcal{L}(\cdot)$ represents the loss function between $\hat{I}_H$ and I_H. $\hat{\theta_\gamma}$ denotes the parameter of the model when the loss value is minimized.

B. Expectation-Maximization Algorithm

The expectation maximization (EM) algorithm aims to find the maximum likelihood solution for models with latent variables. The observed dataset from K models can be denoted as $X = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$. The latent variable corresponding to the n th observed data x_n is $Z_n = \{z_{n1}, z_{n2}, \ldots, z_{nK}\}$. We call $\{X, Z\}$ the complete data and the likelihood function can be denoted as $\ln p(X, Z \mid \theta)$, where θ represents the total parameters of the model. The ultimate goal of the EM algorithm is to maximize the likelihood function and obtain all of the parameters in the model by

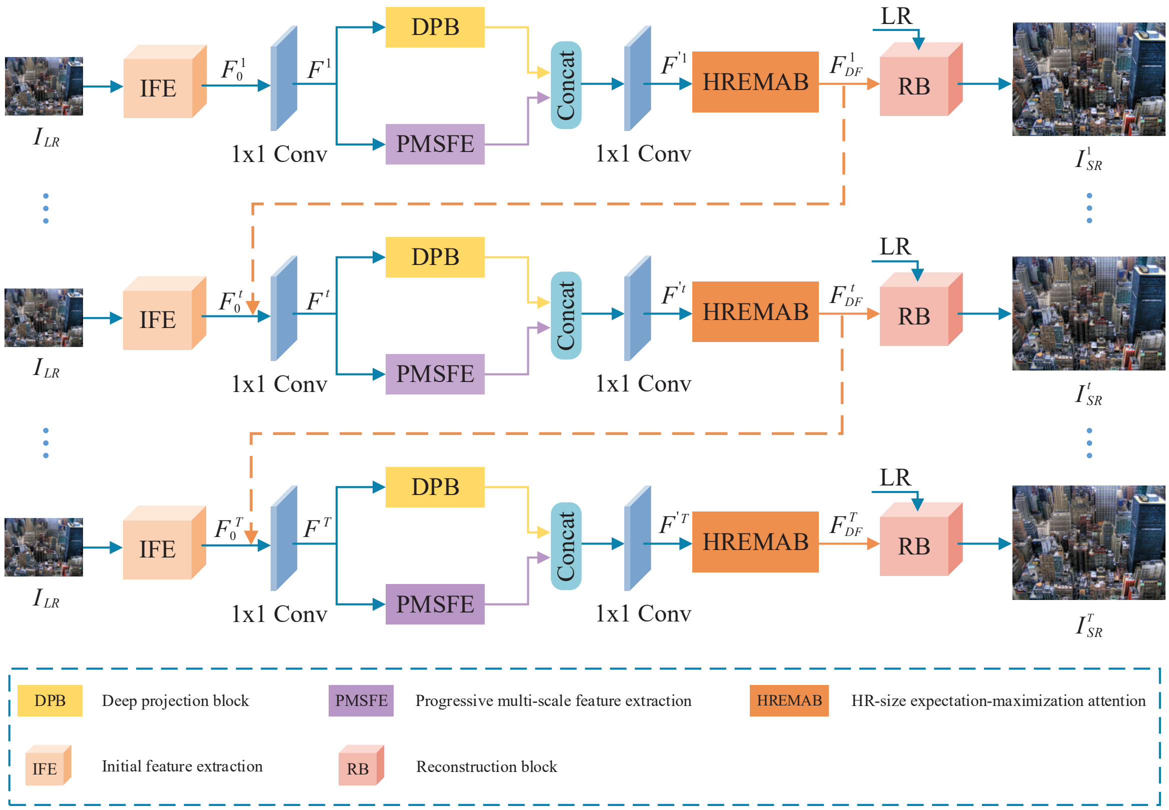

Fig. 2. An overview of our EMASRN network. EMASRN is composed of T generations as a whole. In above figure, we only show the first, tth and Tth generations, and the rest are represented by dashed lines. For each generation, the input is the same LR image, and the output is the SR image corresponding to each generation. EMASRN fuses the high-level feature obtained by the previous generation with the initial feature of the current generation. I_{SR}^t represents the output of the tth generation. The output of the Tth generation I_{SR}^T is selected as the total output of the entire network.

two steps, i.e., the E step and the M step. In step E, the posterior distribution of Z is obtained by using the current model parameter θ^{old}. Then, the posterior distribution is used to find the complete likelihood function:

$$Q\left(\theta, \theta^{old}\right)=E\left[\log p(X, Z \mid \theta) \mid X, \theta^{old}\right]. \tag{5}$$

In step M, the likelihood function is maximized to update the model parameters:

$$\theta^{\text{new}}=\arg \max_{\theta} Q\left(\theta, \theta^{old}\right), \tag{6}$$

where θ^{new} represents the updated parameters of the model. Step E and Step M are executed alternately until the convergence condition is met.

IV. Proposed Method

A. Network Architecture

As shown in Fig. 2, our EMASRN mainly consists of of five parts: initial feature extraction (IFE), deep projection block (DPB), progressive multi-scale feature extraction (PMSFE), HR-size expectation-maximization attention block (HREMAB) and reconstruction block (RB). DPB, PMSFE and HREMAB constitute the deep feature extraction block (DFEB) of our network. Our EMASRN is composed of T generations in total. By using the feedback networks [43], EMASRN fuses the high-level feature obtained by the $(t-1)$th generation with the initial feature of the tth generation. In the tth generation, the LR image is used as an input, and the output is the corresponding super-resolution image (I_{SR}^t). The loss of each generation is calculated by I_{SR}^t and HR and added in a certain proportion as the total loss of the model. The final SR image is obtained by the output of the Tth generation.

For the tth generation, the input image is denoted as I_{LR}. For I_{LR}, the initial feature extraction block is used to extract the initial feature:

$$F_0^t=H_{IFE}\left(I_{LR}\right), \tag{7}$$

where F_0^t represents the shallow feature extracted from the tth generation. $H_{IFE}(\cdot)$ denotes the initial feature extraction block. The initial feature extraction block consists of two convolutional layers with a convolution kernel size of 3 × 3. Then the extracted shallow feature is sent to the deep feature extraction block:

$$F_{DF}^t=H_{DFEB}\left(\operatorname{conv}\left[F_0^t, F_{DF}^{t-1}\right]\right)=H_{DFEB}\left(F^t\right), \tag{8}$$

where $H_{DFEB}(\cdot)$ stands for the deep feature extraction block, F_{DF}^t represents the deep feature extracted by the tth generation, and F_{DF}^{t-1} denotes the deep feature extracted by the $(t-1)$th generation. Furthermore, F^t is obtained by concatenating F_0^t and F_{DF}^{t-1} according to the channel dimension.

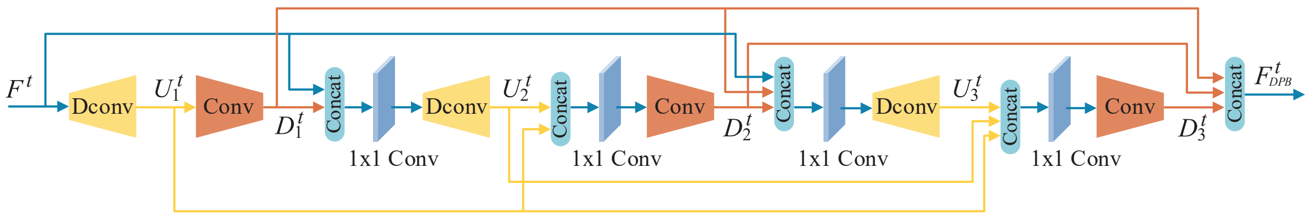

Fig. 3. The structure of the deep projection block, where U_i^t denotes the upsampling feature of the ith group and D_i^t stands for the downsampling feature of the ith group.

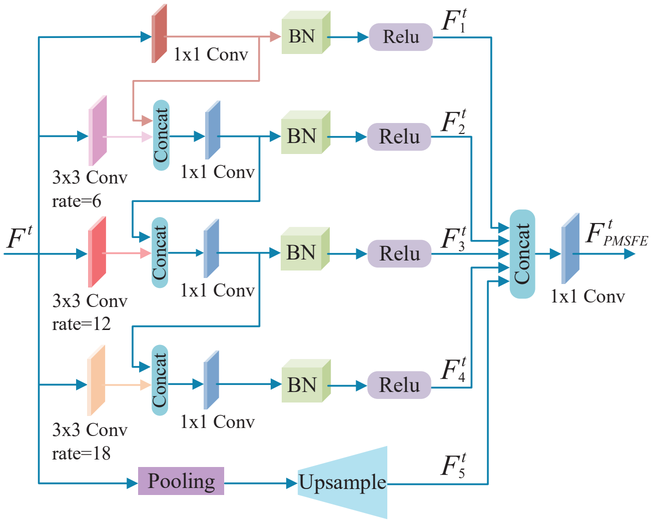

Fig. 4. The structure of progressive multi-scale feature extraction, where F_1^t, F_2^t, F_3^t , F_4^t and F_5^t are the features of different scales.

Finally, F_{DF}^t and I_{LR} are sent to the reconstruction block through the residual connection:

$$\begin{aligned} I_{SR}^t &= H_{RB}\left(F_{DF}^t, I_{LR}\right) \\ &= I_{\uparrow} + F_R^t \\ &= H_{EMASRN}\left(I_{LR}\right), \end{aligned} \tag{9}$$

where $H_{RB}(\cdot)$ represents the reconstruction block. In the reconstruction block, I_{LR} is upsampled bilinearly and $I_{\uparrow}$ is obtained, which has the same size as the HR image. At the same time, a deconvolution and a convolution operation are conducted on F_{DF}^t successively to obtain F_R^t which also has the same size as the HR image. Finally, $I_{\uparrow}$ is added to F_R^t via a residual connection, and the output image I_{SR}^t of the tth generation is obtained. In general, $H_{EMASRN}(\cdot)$ represents the overall model we proposed, which converts I_{LR} to I_{SR}^t.

B. Deep Projection Block

Timofte *et al.* [44] pointed out that back-projection can effectively improve the performance of super-resolution tasks. There are also some methods [9], [26] that utilized iterative upsampling and downsampling to perform back projection and achieved state-of-the-art result at that time. Similarly, we use the deep projection block to extract the deep feature with continuous upsampling and downsampling.

As shown in Fig. 3, the deep projection block is composed of three projection groups via dense connection. The upsampling feature of the ith group can be expressed as:

$$U_i^t = B_i^{\uparrow}\left(\left[F^t, D_1^t, \ldots, D_{i-1}^t\right]\right), \tag{10}$$

where U_i^t denotes the upsampling feature of the ith group and $B_i^{\uparrow}(\cdot)$ stands for the upsampling layer of the ith projection group. F^t is obtained by concatenating F_0^t and F_{DF}^{t-1} according to the channel dimension followed by a convolutional layer, which is mentioned in the previous section. D_1^t represents the downsampling feature. With the exception of the first projection group, each upsampling layer undergoes a convolution operation with a convolution kernel size of 1×1. The corresponding downsampling feature of the ith group can be computed as:

$$D_i^t = B_i^{\downarrow}\left(\left[U_1^t, U_2^t, \ldots, U_i^t\right]\right), \tag{11}$$

where D_i^t stands for the downsampling feature of the ith group and $B_i^{\downarrow}(\cdot)$ represents the downsampling layer of the ith projection group. After obtaining the downsampling feature of each group, these features are concatenated according to the channel dimension:

$$F_{DPB}^t = \left[D_1^t, D_2^t, \ldots, D_I^t\right], \tag{12}$$

where F_{DPB}^t denotes the feature extracted by the deep projection block in the tth generation.

C. Progressive Multi-Scale Feature Extraction

Convolution has a fixed size, and the features extracted by convolution are often within the size range of the convolution kernel. If only a certain fixed-size convolution kernel is used for feature extraction, then the feature information obtained will be local, and this local feature information will limit the reconstruction ability of the network. Many methods have proved that multi-scale features are beneficial for improving the performance of computer vision tasks, such as optical flow estimation [45], cardiac motion estimation [46] and style transfer [47]. For example, Chen *et al.* [48] utilized dilated convolution to perform multi-scale feature extraction, and proposed the encoding-decoding image segmentation method DeepLabV3+. However, this method concatenates the feature maps of different scales directly, which makes it difficult to merge this information. To solve this problem, we propose a progressive multi-scale feature extraction block (PMSFE), which concatenates the adjacent scale feature information progressively in the early stage. Our PMSFE is shown in

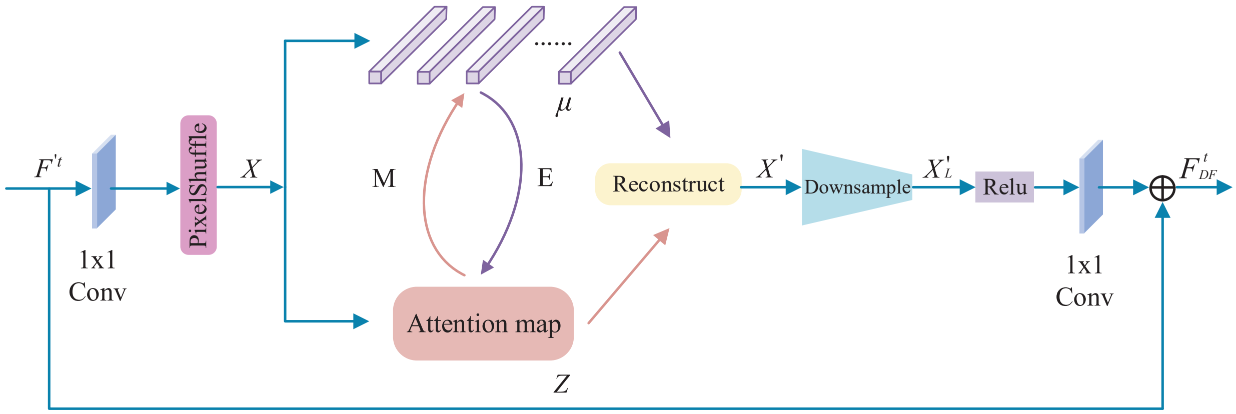

Fig. 5. The structure of HR-size expectation-maximization attention, where μ represents the base and Z represents the contribution of the basis to the pixels. Step E and Step M are executed alternately until convergence.

Fig. 4. For the input feature F^t, conventional convolution with a kernel size of 1×1 and dilated convolution with dilation rates of 6, 12 and 18 are performed to extract the features of different scales. Then, the adjacent scale feature is concatenated progressively to ensure the effectiveness of the fusion. Batch normalization and rectified linear units are subsequently performed to obtain F^t_1, F^t_2, F^t_3 and F^t_4 respectively. At the same time, a pooling operation is performed on F^t to obtain F^t_5. F^t_1, F^t_2, F^t_3, F^t_4 and F^t_5 are concatenated according to the channel dimension, and then a convolution layer is performed to obtain the output of the progressive multi-scale feature extraction block:

$$\begin{aligned} F^t_{PMSFE} &= H_{PMSFE}\left(\text{conv}\left[F^t_1, F^t_2, F^t_3, F^t_4, F^t_5\right]\right) \\ &= H_{PMSFE}\left(F^t\right), \end{aligned} \tag{13}$$

where F^t_{PMSFE} stands for the multi-scale feature of tth generation and $H_{PMSFE}(\cdot)$ denotes the operation of progressive multi-scale feature extraction.

D. HR-Size Expectation-Maximization Attention

Each pixel in the image has a certain correlation with other pixels instead of being independent of each other. However, the convolution operation can only capture the relationship between the pixels in the area of the convolution kernel and cannot establish the relationship between any pixel in the image and all other pixels. To solve this problem, Wang *et al.* [40] utilized the self-attention mechanism for computer vision tasks for the first time in 2018 and proposed a method named Non-local. Although Non-local can capture the long-range dependencies between the pixels of the feature map, this method needs to calculate the relationship between all of the pixels in the image, which entails an unacceptable computational burden. On the one hand, the input LR images usually have a relatively larger size, which needs to consume huge computing resources, on the other hand, capturing the long-range dependencies only on the LR size images cannot greatly improve the performance of super-resolution tasks.

Li *et al.* [42] proposed a new type of attention mechanism, namely expectation-maximization attention, which computes a more compact basis set by iteratively executing the EM algorithm. Inspired by this, we design an HR-size expectation-maximization attention block (HREMAB) that can directly capture the long-range dependencies on the HR-size feature map. As shown in Fig. 5, the input feature of this block is $F'^t \in \mathbb{R}^{N \times C}$, where $N = H \times W$, H and W represent the height and width of the feature map, respectively. To directly capture the long-range dependencies on the feature map that have the same size as the HR image, we upsample the feature map following [49] and obtain $X \in \mathbb{R}^{N' \times C}$, where $N' = (rH) \times (rW)$, and r is the upsampling scale factor. We select $\mu \in \mathbb{R}^{K \times C}$ as the initial bases, and the latent variable $Z \in \mathbb{R}^{N' \times K}$ can be formulated as:

$$z_{nk} = \frac{\mathcal{K}\left(\mathbf{x}_n, \boldsymbol{\mu}_k\right)}{\sum_{j=1}^{K} \mathcal{K}\left(\mathbf{x}_n, \boldsymbol{\mu}_j\right)}, \tag{14}$$

where the latent variable z_{nk} represents the contribution of the kth basis to the nth pixel, and $\mathcal{K}(\cdot)$ represents the general kernel function. For simplicity, we take the exponential inner dot $\exp\left(\mathrm{a}^{\mathrm{T}}\mathrm{b}\right)$ as $\mathcal{K}(\cdot)$. In practice, the above formula can be implemented as a matrix multiplication plus one softmax layer. Therefore, in the tth iteration, step E can be formulated as:

$$Z^{(t)} = \text{softmax}\left(X\left(\mu^{(t-1)}\right)^{\mathrm{T}}\right), \tag{15}$$

where $\mu^{(t-1)}$ represents the updated bases in the $(t-1)$th iteration. Subsequently, $\mu^{(t)}$ is updated in the tth iteration of step M:

$$\mu^{(t)}_k = \frac{\sum_{n=1}^{N} z^{(t)}_{nk} x_n}{\sum_{n=1}^{N} z^{(t)}_{nk}}, \tag{16}$$

where $\mu^{(t)}_k$ represents the kth basis in the tth iteration. After the E step and M step are executed alternately T times, bases μ and latent variable Z are used to reconstruct feature map X':

$$X' = Z\mu\,. \tag{17}$$

Compared with X, each pixel in X' is associated with all other pixels, which makes the feature map more truly reflect the image's internal information. More importantly, in contrast to the method that captures the long-range dependencies on the LR size feature maps, our method directly captures the

TABLE I
COMPARISON RESULTS ACHIEVED BY OUR METHOD AND ITS VARIANTS. ALL OF THE VARIANTS ARE USED TO TEST THE PSNR VALUE ON THE SET5 DATASET FOR ×4 SR.

	1	2	3	4	5
DPB	✓	✓	✓	✓	✓
PMSFE		✓		✓	✓
HREMAB			✓		✓
LREMAB				✓	
Params	472K	531K	504K	515K	546K
PSNR on Set5 (×4)	31.9209	31.9763	31.9731	31.9260	32.0178

TABLE II
COMPARATIVE RESULTS BY EMASRN FOR ×4 SR WITH DIFFERENT SETTINGS, WHERE "CONV" REPRESENTS THE ADDED CONVOLUTION LAYER TO INCREASE THE NUMBER OF PARAMETERS.

Model	Params	Set5	Set14	B100	Urban100	Manga109
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
EMASRN with Non-local	515K	31.7845/0.8899	28.3531/0.7756	27.3697/0.7293	Out of Memory	Out of Memory
EMASRN with Non-local+Conv	551K	31.5630/0.8867	28.2032/0.7717	27.3089/0.7269	Out of Memory	Out of Memory
EMASRN with LREMAB	515K	31.9260/0.8912	28.4452/0.7778	27.4544/0.7312	25.7421/0.7735	30.0187/0.9022
EMASRN with LREMAB+Conv	551K	31.8500/0.8901	28.4192/0.7766	27.4302/0.7302	25.6891/0.7710	29.9732/0.9011
EMASRN with HREMAB	546K	32.0178/0.8922	28.4685/0.7784	27.4681/0.7319	25.7475/0.7740	30.0344/0.9022

long-range dependencies on the feature maps that have the same size as HR images. This makes the correlation between pixels in the feature map more consistent with the HR image. After X' is obtained, downsampling is performed to obtain X'_L which has the same size as the LR image. Then, a rectified linear unit and convolution are performed on X'_L to obtain the output:

$$F_{DF}^{t} = F'^{t} + \text{conv}\left(\text{relu}\left(X'_L\right)\right), \tag{18}$$

where F_{DF}^{t} represents the output of the HR-size expectation-maximization attention block.

V. EXPERIMENTS

A. Datasets and Metrics

Following [23], [50], [51], we use the high-quality (2K resolution) images from the DIV2K [52] dataset as our training set. We adopt several standard benchmark datasets during testing, including Set5, Set14, BSD100, Urban100, Manga109 and DIV2K (100 validation images in total), and each dataset has different characteristics. For qualitative evaluation, we convert the image to YCbCr channels and calculate the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) on the Y channel as the evaluation metrics.

B. Implementation Details

To prepare the training data, we synthesize the LR images by downsampling the training HR images using bicubic interpolation. The training images are augmented by randomly rotating 90°, 180°, 270° and horizontally flipping. In each min batch, 16 LR color patches are provided as inputs. We train our model for 1000 epochs and set T=4. Following [35], we choose ADAM as our optimizer and L1 loss to optimize our model. The learning rate is initialized as 0.0001 and then reduced to half after every 200 epochs. We implement our networks with the PyTorch framework and train them on NVIDIA 2080Ti GPUs.

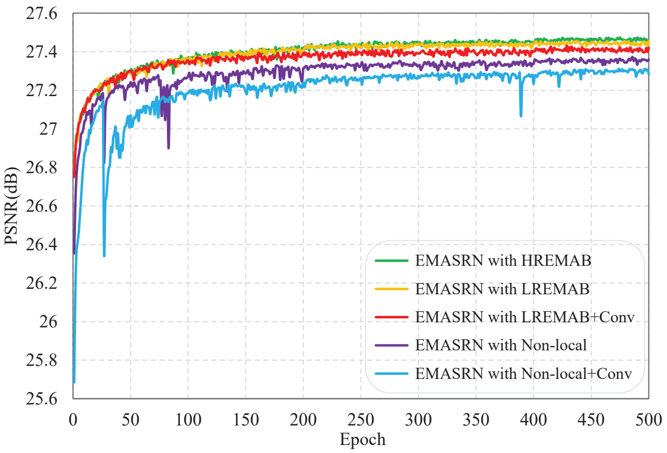

Fig. 6. Comparisons of five different settings on the B100 dataset for ×4 SR, where "Conv" represents the added convolution layer that increases the number of parameters.

C. Model Analysis

To quickly study the effectiveness of the deep projection block (DPB), progressive multi-scale feature extraction (PMSFE) and HR-size expectation maximization attention block (HREMAB), we set the total epoch to 500 and T=2 to train the corresponding model. We remove the corresponding blocks and train the model, and the results are shown in Table I, Table II, Table III and Fig. 6.

1) Progressive Multi-scale Feature Extraction: A progressive multi-scale feature extraction block (PMSFE) is proposed to extract the feature information of different scales in the image. In order to test the effectiveness of PMSFE, we added PMSFE to model 1 and formed model 2. As we can see from Table I, the performance of the model benefits from PMSFE. If the PMSFE is added, the PSNR value is increased from 31.9209 to 31.9763. In addition, If PMSFE is removed from model 5, the PSNR value decreases from 32.0178 to 31.9731. This change occurs because that multi-scale feature extraction

TABLE III
TRADE-OFF BETWEEN T AND THE PERFORMANCE ON DIFFERENT DATASETS DURING TESTING FOR ×4 SR, WHERE T STANDS FOR THE NUMBER OF FEEDBACK ITERATIONS.

	Set5	Set14	B100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
T=4	32.17/0.8948	28.57/0.7809	27.55/0.7351	26.01/0.7838	30.41/0.9076
T=3	32.13/0.8944	28.55/0.7808	27.53/0.7348	25.95/0.7818	30.32/0.9067
T=2	32.00/0.8926	28.48/0.7799	27.48/0.7340	25.78/0.7765	30.06/0.9037
T=1	31.53/0.8852	28.22/0.7718	27.30/0.7256	25.34/0.7592	29.28/0.8919

TABLE IV
COMPARISON RESULTS BETWEEN OUR EMASRN AND THE STATE-OF-THE-ART METHODS. WE COMPARE THE PSNR AND SSIM ON THE SET5, SET14, B100, URBAN100, MANGA109 AND DIV2K (100 VALIDATION IMAGES) DATASETS.

Model	Scale	Params	Set5	Set14	B100	Urban100	Manga109	DIV2K
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	3	-	30.39/0.8682	27.55/0.7742	27.21/0.7349	24.46/0.7349	26.95/0.8556	28.22/0.8906
SRCNN [13]	3	57K	32.75/0.9090	29.28/0.8209	28.41/0.7863	26.24/0.7989	30.59/0.9107	29.64/0.9138
FSRCNN [53]	3	12K	33.16/0.9140	29.43/0.8242	28.53/0.7910	26.43/0.8080	30.98/0.9212	-
VDSR [20]	3	665K	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279	32.01/0.9310	30.09/0.9208
DRCN [54]	3	1,774K	33.82/0.9226	29.76/0.8311	28.80/0.7963	27.15/0.8276	32.31/0.9328	-
LapSRN [7]	3	502K	33.81/0.9220	29.79/0.8325	28.82/0.7980	27.07/0.8275	32.21/0.9350	-
DRRN [55]	3	297K	34.03/0.9244	29.96/0.8349	28.95/0.8004	27.53/0.8378	32.74/0.9390	-
MemNet [56]	3	677K	34.09/0.9248	30.00/0.8350	28.96/0.8001	27.56/0.8376	-	-
SelNet [57]	3	974K	34.27/0.9257	30.30/0.8399	28.97/0.8025	-	-	-
IDN [33]	3	553K	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359	32.71/0.9381	-
SRMDNF [50]	3	1,530K	34.12/0.9250	30.04/0.8370	28.97/0.8030	27.57/0.8400	33.01/0.9399	-
CARN [8]	3	1,592K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.43/0.9427	32.33/0.8860
CARN-M [8]	3	412K	33.99/0.9236	30.08/0.8367	28.91/0.8000	27.55/0.8385	32.79/0.9383	31.99/0.8815
SRFBN-S [9]	3	349K	34.20/0.9255	30.10/0.8372	28.96/0.8010	27.66/0.8415	33.02/0.9404	32.11/0.8826
CFSRCNN [10]	3	1,495K	34.24/0.9256	30.27/0.8410	29.03/0.8035	28.04/0.8496	33.33/0.9423	-
FilterNet [35]	3	1,249K	34.08/0.9250	30.03/0.8370	28.95/0.8030	27.55/0.8380	-	-
EMASRN	3	427K	34.36/0.9264	30.30/0.8411	29.05/0.8035	28.04/0.8493	33.43/0.9433	32.31/0.8859
EMASRN+	3	427K	**34.48/0.9275**	**30.38/0.8422**	**29.11/0.8046**	**28.17/0.8514**	**33.71/0.9450**	**32.39/0.8869**
Bicubic	4	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866	26.66/0.8521
SRCNN [13]	4	57K	30.48/0.8628	27.49/0.7503	26.90/0.7101	24.52/0.7221	27.66/0.8505	27.78/0.8753
FSRCNN [53]	4	12K	30.71/0.8657	27.59/0.7535	26.98/0.7150	24.62/0.7280	27.90/0.8517	-
VDSR [20]	4	665K	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8809	28.17/0.8841
DRCN [54]	4	1,774K	31.53/0.8854	28.02/0.7670	27.23/0.7233	25.14/0.7510	28.98/0.8816	-
LapSRN [7]	4	502K	31.54/0.8852	28.09/0.7700	27.32/0.7275	25.21/0.7562	29.09/0.8900	-
DRRN [55]	4	297K	31.68/0.8888	28.21/0.7720	27.38/0.7284	25.44/0.7638	29.46/0.8960	-
MemNet [56]	4	677K	31.74/0.8893	28.26/0.7723	27.40/0.7281	25.50/0.7630	-	-
SelNet [57]	4	1,417K	32.00/0.8931	28.49/0.7783	27.44/0.7325	-	-	-
IDN [33]	4	553K	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8942	-
SRMDNF [50]	4	1,555K	31.96/0.8930	28.35/0.7770	27.49/0.7340	25.68/0.7730	30.12/0.9018	-
SRDenseNet [58]	4	2,015K	32.02/0.8934	28.50/0.7782	27.53/0.7337	26.05/0.7819	-	-
CARN [8]	4	1,592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.42/0.9070	30.41/0.8363
CARN-M [8]	4	412K	31.92/0.8903	28.42/0.7762	27.44/0.7304	25.62/0.7694	29.80/0.8988	30.11/0.8305
SRFBN-S [9]	4	427K	31.98/0.8923	28.45/0.7779	27.44/0.7313	25.71/0.7719	29.91/0.9008	30.19/0.8316
CFSRCNN [10]	4	1,458K	32.06/0.8920	28.57/0.7801	27.52/0.7333	26.03/0.7824	30.32/0.9050	-
FilterNet [35]	4	1,249K	31.74/0.8900	28.27/0.7730	27.39/0.7290	25.53/0.7680	-	-
EMASRN	4	546K	32.17/0.8948	28.57/0.7809	27.55/0.7351	26.01/0.7838	30.41/0.9076	30.37/0.8363
EMASRN+	4	546K	**32.31/0.8964**	**28.66/0.7828**	**27.61/0.7364**	**26.15/0.7868**	**30.69/0.9105**	**30.46/0.8379**

can help the model fully mine various information in the image, which is of great significance to the reconstruction of the image.

2) HR-Size Expectation Maximization Attention Block: An HR-size expectation maximization attention block (HREMAB) is proposed to capture the long-range dependencies on the feature map. Comparing model 3 with model 1 in Table I, we find that the PSNR gain brought by HREMAB is almost the same as that brought by PMSFE. Moreover, compared with model 5, HREMAB is removed from model 2, which makes the PSNR value decrease drastically from 32.0178 to 31.9763. This change also implies that effectively using the correlation between pixels within the image can greatly assist image reconstruction.

3) HR-Size Self-Attention vs Non-local: Non-local is the first method to apply the self-attention mechanism in computer vision tasks; however, it suffers because it involves huge number of calculations and great complexity. To test the performance of Non-local, we also replace HREMAB with Non-local and conduct experiments. From Table II and Fig. 6 we can see that compared with HREMAB, the performance of Non-local is reduced significantly. It is worth noting that due to the large consumption of computing resources by Non-local, it exceeds the maximum memory range on Urban100 and Manga109. This also means that Non-local is too expensive for image super-resolution tasks.

4) HR-Size Self-Attention vs LR-Size Self-Attention: The self-attention mechanism can capture the correlation between

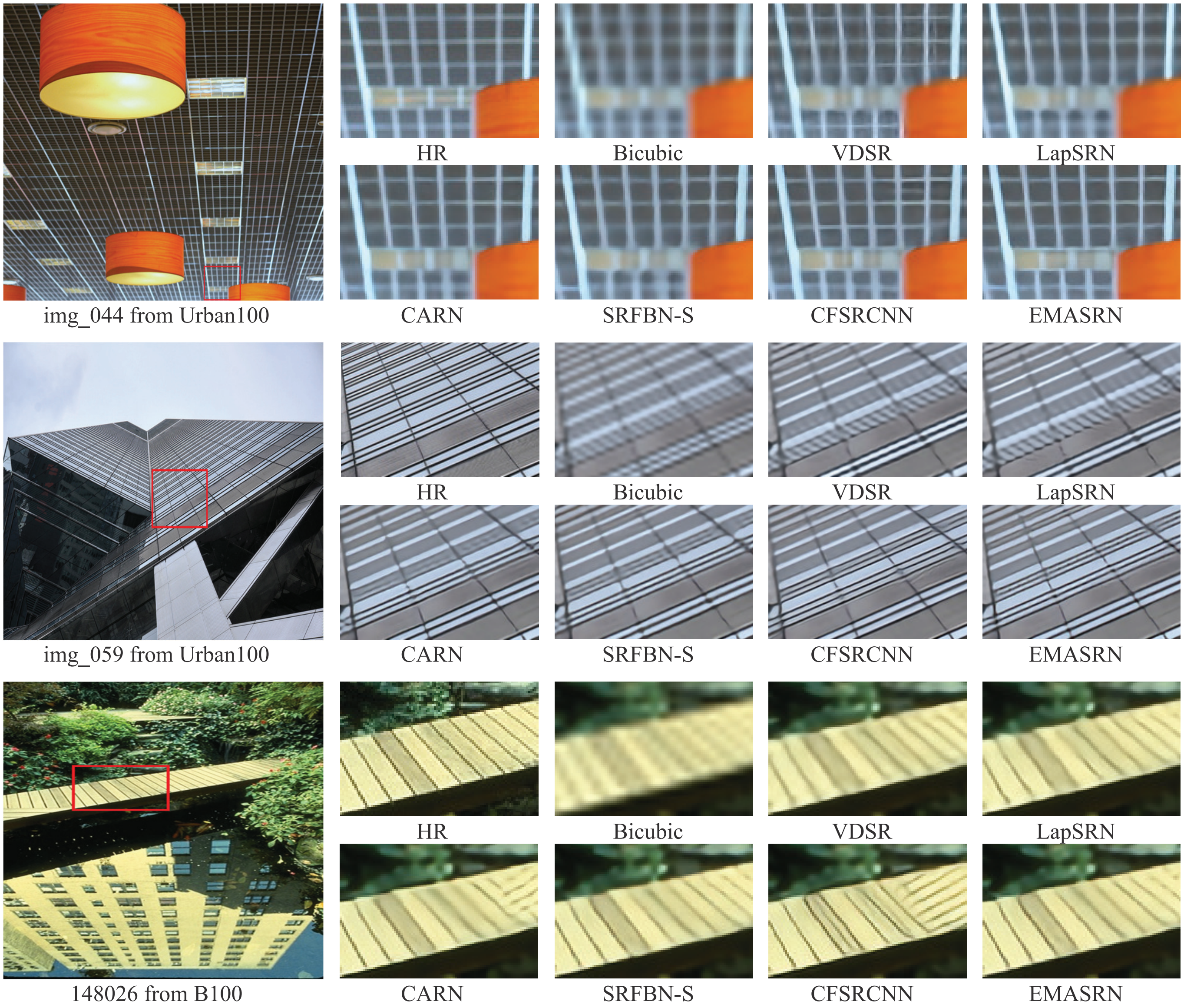

Fig. 7. Visual comparisons of the state-of-the-art methods and our model on different benchmark test datasets for ×4 SR. The key contrast parts in the red rectangle are magnified to display on the right.

the pixels in an image. However, for super-resolution tasks, the inputs are usually LR-size images, but what we want to obtain are HR images. If we capture long-range dependencies directly through Non-local, the machine would run out of memory, as shown in Table II. Moreover, if we capture the long-range dependencies on the LR-size images, this process will still limit the model's ability to reconstruct HR images. To test the superiority of the self-attention mechanism on HR-size images, we replaced the original HREMAB in our model with LREMAB, which performs a self-attention mechanism on LR-size images. From the comparative results in Table II, we can see that if we perform the self-attention mechanism on LR-size images, the PSNR value will drop sharply on all of the test sets compared to the HR-size images. To verify that the gain introduced by HREMAB over LREMAB does not come from the increase in parameters, we add convolution layers to LREMAB to make its number of parameters reach 551K. It can be seen from Table II that simply adding convolution layers does not improve the performance, and it can even make the performance worse. From Fig. 6, we draw the same conclusion. These results further prove that capturing the long-range dependence on the HR-size images will help image reconstruction.

5) Effectiveness of T: During the training process, our model uses a feedback network that allows us to flexibly adjust the iteration of the feedback during the test. To study the influence of the number of feedback iterations on the experiment, we conduct corresponding experiments, and the experimental results are shown in Table III. From Table III, we can see that as the number of iterations grows, the performance of the model is constantly improving. Specifically, when setting T=3 and T=4, the performance of the model is not very different. It is worth noting that when T is set to 2, the performance of the model drops dramatically compared to the model in which T is set to 4. Therefore, we can find that the feedback network promotes the performance of our model, especially when the number of iterations is small. In practical applications, we can also flexibly make the model reach a good balance between performance and speed by adjusting T.

D. Comparison with State-of-the-Art Methods

To test the effectiveness of our model, we compare our model with several state-of-the-art methods: SRCNN [13], FSRCNN [53], VDSR [20], DRCN [54], LapSRN [7], DRRN [55], MemNet [56], SelNet [57], IDN [33], SRMDNF [50], SRDenseNet [58], CARN [8], SRFBN-S [9], CFSRCNN [10] and FilterNet [35]. We also adopt the self-ensemble method [44] to further improve our EMASRN, and we denote it as EMASRN+. All of the tests are performed on Set5, Set14, B100, Urban100, Manga109 and DIV2K.

The average peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values of the six datasets are shown in Table IV. Compared with the other methods, our method achieves the best results on all of the datasets with various scaling factors. Even without a self-ensemble, our method still performs favorably against the state-of-the-art results on most datasets. For example, compared with CARN-M, which is similar to the parameters of our method, our EMASRN achieves a notable gain of 0.25dB on PSNR for ×4 upscaling on Set5. Although our method is inferior to CARN on a small part of the dataset, the overall effects of the two are very similar. SRFBN-S has slightly fewer parameters than our model, but its performance is far inferior to our method. When compared with CFSRCNN, our method performs better than it on almost all datasets for scaling factors ×3 and ×4. It is very difficult to increase the performance of the model when the number of parameters is limited, but the number of parameters of CARN and CFSRCNN are almost three times the corresponding number of our method. In other words, our model achieves the same or even better results than the state-of-the-art methods despite having fewer parameters. This result also indicates that our model has more powerful reconstruction capabilities.

We also show the zoomed results of various methods in Fig. 7. We can find that there is a large gap between the image produced by most methods and the HR image. However the image processed by our method is the closest to the original image. For example, for “img_044”, we find that most methods cannot accurately recover the details of the grid in the image. Although SRFBN-S performs slightly better than our method, the reconstruction of the small grid at the bottom of the image is still very blurry. CFSRCNN even produces redundant textures, such as an inexplicable white horizontal line at the top right of the image. Each grid generated by our method is very clear and easy to identify with good visual effects.

As shown in “img_059”, compared with the original image, the lines of most methods are seriously merged. For example, VDSR only separates the bottom line, and the other lines are merged together. Although CFSRCNN separates more lines, the lines are very blurred and even distorted in some areas. Each line generated by our method is more independent and clearer, and can reflect the details of the original image more accurately.

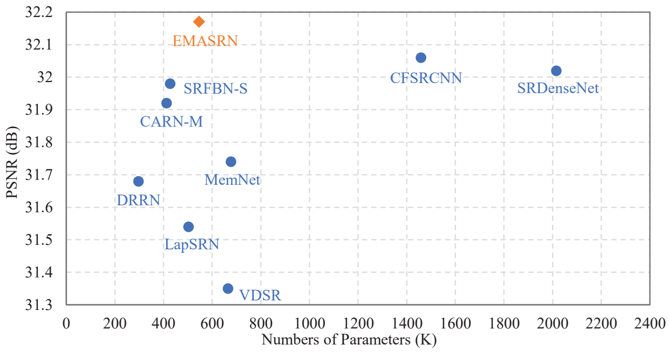

Fig. 8. Trade-off between the performance and the parameters on the Set5 dataset for ×4 SR.

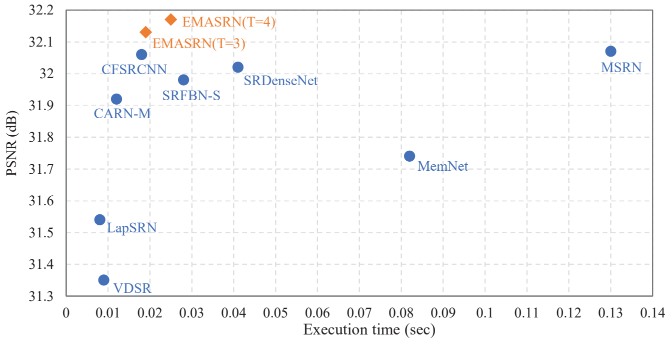

Fig. 9. Trade-off between the performance and the execution time on the Set5 dataset for ×4 SR.

It can be seen from “148026” that compared with the original image, the lines of most methods are unclear. For example, the lines processed by VDSR are obviously jagged. Although CARN can produce slightly sharper lines, it generates many false textures. CFSRCNN produced some lines that did not exist in the original image. Again, we find that each line generated by our method is more independent and clearer, and can reflect the details of the original image more accurately.

E. Parameters and Execution Time

In this section, we study the trade-off of PSNR vs. parameters, and PSNR vs. execution time. In our work, we follow the setting in [33], [35] to use Set5 dataset to evaluate the parameters and execution time of our algorithm. The Set5 dataset contains five images with different resolutions, i.e., 512×512, 288×288, 256×256, 280×280 and 228×344 respectively. This dataset covers a wide range of resolutions to compare the efficiency with state-of-the-art methods objectively. The experimental results are shown in Fig. 8 and Fig. 9. It can be seen from Fig. 8 that CFSRCNN is the closest to the performance of our method, but the number of parameters is approximately three times that of our method. SRFBN-S is close to the number of parameters of our method, but its performance is poor compared to our method. Therefore, compared with these methods, our EMASRN achieves a good balance between the parameters and performance.

Next, we study the trade-off between the execution time and performance of different methods. As shown in Fig. 9, MSRN has a similar performance compared with our method, but it has a long execution time. Although our method requires slightly more time than CFSRCNN, our method achieves better performance, which allows us to disregard a slightly longer execution time. Furthermore, when we set T=3, our method not only achieves the same execution time as CFSRCNN but also has a better performance. This result shows that we can flexibly achieve better performance or less execution time by adjusting the total generation T.

In summary, compared with other methods, our method can greatly reduce the parameters of the model while ensuring performance. Furthermore, our model has a fast execution speed, and we can also obtain a good balance between the performance and execution time by adjusting T.

VI. Conclusion

In this paper, we propose a lightweight single image super-resolution network with an expectation-maximization attention mechanism. Our EMASRN first extracts the deep features of low-resolution images though a depth projection block. At the same time, a progressive multi-scale feature extraction block is employed to extract the feature information of different scales. Then, we propose an HR-size expectation-maximization attention block that can directly capture the long-range dependencies of HR-size feature maps. Finally, we use the feedback network to feed back the high-level features of each generation to the next-generation's shallow network. Extensive experimental results show that our EMASRN not only has a small number of parameters but can also reconstruct images with higher quality. Comparisons with existing lightweight SISR methods have also demonstrated the state-of-the-art performance of our EMASRN.

In our future work, we will apply the expectation-maximization attention mechanism to improve the performance of stereo image super-resolution and video super-resolution which require more efficient network architecture than the classical SISR applications. In addition, we will apply the proposed network architecture into other low-level computer vision tasks, such as image denoising and image deblurring.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 62076255; in part by the Hunan Provincial Science and Technology Plan Project 2020SK2059; in part by the National Science Foundation of Hunan Province, China, under Grant 2019JJ20025 and Grant 2019JJ40406; in part by the National Social Science Fund of China (No. 20&ZD120).

References

[1] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 206–221.

[2] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Proc. Int. Conf. Articulated Motion Deformable Objects*. Springer, 2016, pp. 175–184.

[3] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, Jan. 2009.

[4] A. Swaminathan, M. Wu, and K. R. Liu, "Digital image forensics via intrinsic fingerprints," *IEEE Trans. Inf. Forensic Secur.*, vol. 3, no. 1, pp. 101–117, Mar. 2008.

[5] M. T. Merino and J. Nunez, "Super-resolution of remotely sensed images with variable-pixel linear reconstruction," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 5, pp. 1446–1457, May. 2007.

[6] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Trans. Med. Imaging*, vol. 2, no. 1, pp. 31–39, Mar. 1983.

[7] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 624–632.

[8] N. Ahn, B. Kang, and K. A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 252–268.

[9] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3867–3876.

[10] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C. W. Lin, "Coarse-to-fine cnn for image super-resolution," *IEEE Trans. Multimedia*, to be published. DOI: 10.1109/TMM.2020.2999182.

[11] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *J. Vis. Commun. Image Represent.*, vol. 4, no. 4, pp. 324–335, May. 1993.

[12] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 996–1011, Jun. 1996.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[14] W. Yang, S. Xia, J. Liu, and Z. Guo, "Reference-guided deep super-resolution via manifold localized external compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1270–1283, May. 2019.

[15] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, and T. S. Huang, "Learning temporal dynamics for video super-resolution: A deep learning approach," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3432–3445, Jul. 2018.

[16] Y. Fan, H. Shi, J. Yu, D. Liu, W. Han, H. Yu, Z. Wang, X. Wang, and T. S. Huang, "Balanced two-stage residual networks for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 161–168.

[17] Y. Jing, X. Liu, Y. Ding, X. Wang, E. Ding, M. Song, and S. Wen, "Dynamic instance normalization for arbitrary style transfer," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4369–4376.

[18] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2503–2516, Jul. 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[20] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1646–1654.

[21] J. Yu, Y. Fan, J. Yang, N. Xu, X. Wang, and T. S. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[22] H. Yu, D. Liu, H. Shi, H. Yu, Z. Wang, X. Wang, B. Cross, M. Bramler, and T. S. Huang, "Computed tomography super-resolution using convolutional neural networks," in *Proc. IEEE Int. Conf. Inf. Process.*, 2017, pp. 3944–3948.

[23] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[24] R. Timofte, E. Agustsson, L. V. Gool, M. H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 114–125.

[25] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 517–532.

[26] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1664–1673.

[27] R. Timofte, S. Gu, J. Wu, and L. V. Gool, "Ntire 2018 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 852–863.

[28] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, May. 2020.

[29] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4180–4189.

[30] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 297–306, Jan. 2020.

[31] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 5690–5699.

[32] H. Wu, Z. Zou, J. Gui, W. J. Zeng, J. Ye, J. Zhang, H. Liu, and Z. Wei, "Multi-grained attention networks for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 512–522, Apr. 2021.

[33] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 723–731.

[34] C. Xie, W. Zeng, and X. Lu, "Fast single-image super-resolution via deep network with component learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3473–3486, Nov. 2019.

[35] F. Li, H. Bai, and Y. Zhao, "Filternet: Adaptive information filtering network for accurate and fast image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1511–1523, Mar. 2020.

[36] J. Yin, J. Shen, C. Guan, D. Zhou, and R. Yang, "Lidar-based online 3d video object detection with graph-based message passing and spatiotemporal transformer attention," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 11 495–11 504.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7132–7141.

[38] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Aˆ 2-nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7794–7803.

[41] T. Dai, J. Cai, Y. Zhang, S. T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 11 065–11 074.

[42] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9167–9176.

[43] A. R. Zamir, T. L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese, "Feedback networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1308–1317.

[44] R. Timofte, R. Rothe, and L. V. Gool, "Seven ways to improve example-based single image super resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1865–1873.

[45] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8934–8943.

[46] H. Yu, X. Chen, H. Shi, T. Chen, T. S. Huang, and S. Sun, "Motion pyramid networks for accurate and efficient cardiac motion estimation," in *Proc. Med. Image Comput. Comput. Assisted Intervention*, 2020, pp. 436–446.

[47] Y. Jing, Y. Liu, Y. Yang, Z. Feng, Y. Yu, D. Tao, and M. Song, "Stroke controllable fast style transfer with adaptive receptive fields," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 238–254.

[48] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 801–818.

[49] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1874–1883.

[50] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3262–3271.

[51] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2472–2481.

[52] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 126–135.

[53] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2016, pp. 391–407.

[54] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, June 2016, pp. 1637–1645.

[55] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3147–3155.

[56] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4539–4547.

[57] J. S. Choi and M. Kim, "A deep convolutional neural network with selection units for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 154–160.

[58] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4799–4807.

Xiangyuan Zhu received the B.E. degree in automation from Xiangtan University, Xiangtan, China, in 2017, and the M.E. degree in Control Science and Engineering from Central South University, Changsha, China, in 2020. He is currently pursuing the Ph.D. degree with the school of computer science and engineering with the Central South University. His research interests include low-level vision and deep learning.

Kehua Guo received the Ph.D. degree in Computer Science and Technology from Nanjing University of Science and Technology, Nanjing, China, in 2008. He is currently a professor with Central South University, Changsha, China. He has served as a Guest Editor, Workshop Chair, Publicity Chair, Technical Program Committee Member, and Reviewer of international journals/conference proceedings. His research interests include social computing, ubiquitous computing, big data and image retrieval.

Sheng Ren received the M.S degree in Software Engineering in 2010 from Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in computer science and engineering. His research interests include medical big data, image, and video super-resolution.

Bin Hu received a master's degree in Control Engineering from Central South University in 2019. He is currently a Ph.D. student in the School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include video analysis and understanding, graph convolution networks and abnormal behavior recognition.

Ming Hu received the B.S. degree from Anhui University of Science and Technology in 2019. She is currently a graduate student in the School of Computer Science and Engineering, Central South University, China. Her research interests include image processing and big data.

Hui Fang received the B.S. degree from the University of Science and Technology, Beijing, China, in 2000 and the Ph.D. degree from the University of Bradford, U.K., in 2006. He is currently with the Computer Science Department, Loughborough University (LU). Before joined LU, he has carried out research in several world-leading universities, such as University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualization, visual analytics, and artificial intelligence. Recently, he was awarded several grants as PI and co-PI, including Innovate UK funded "An agent based modelling solution for reliable decision making in crisis and market turmoil in consumer retail", EPSRC funded "RAMP VIS: Making Visual Analytics an Integral Part of the Technological Infrastructure for Combating COVID-19", and NIHR funded "Computer vision to automatically monitor urine output". During his career, he has published more than 60 journal and conference papers.