Lightweight Super-Resolution Using Deep Neural Learning

Zhuqing Jiang, Honghui Zhu, Yue Lu, Guodong Ju, and Aidong Men

***Abstract*—There is a gap between recent development of 4K display technologies and the short storage of 4K contents. Super-Resolution (SR) serves as a bridge to harmonize the need and demand. Recently, Convolutional Neural Network (CNN) based networks have demonstrated great property in image SR. However, most existing methods require large model capacity and consume expensive computation for high performance. Besides, most methods keep the upscaling part relatively simple compared with the feature extraction part. For feature fusion, some methods directly concatenate the features of multi-levels, which is suboptimal due to ignoring the importance of different features. In this work, we propose a recursive multi-stage upscaling network (RMUN) with multiple sub-upscaling modules (SUMs) and a discriminative self-ensemble module (SEM). Specifically, we extract local hierarchical features by using a novel feature extraction module (FEM) which is recursive to reduce the number of parameters. Then, we construct multiple sub-upscaling modules to produce various high-resolution features in forward propagation. This strategy enhances the upscaling part and provides multiple error feedback routes. Furthermore, we employ an SEM for global hierarchical feature recalibration, which can selectively emphasize informative features and surpass less useful ones. Extensive quantitative and qualitative evaluations on benchmark datasets show that our proposed method performs comparable with the state-of-the-art methods in terms of the balance of model size and model performance.**

***Index Terms*—Convolutional neural networks, discriminative fusion, self-ensemble, super-resolution.**

I. Introduction

ULTRA-HIGH-DEFINITION (UHD) video is one of the most popular and influential streaming media in broadcasting nowadays. As we witness worldwide, increasing population is being covered by 4K distribution of TV programs and games, and over half of the display devices on sale are 4K resolution.

However, the 4K content is still in short supply due to the limited adoption of UHD video capture and production systems, as well as large amount of low resolution archives. Thus, there is an urgent need for the Super-Resolution (SR) conversion that we reconstruct High-Resolution (HR) frames from abounding Low-Resolution (LR) ones. SR aims to recover details from an LR image to an HR image. It is a notoriously challenging problem since a specific LR input corresponds to a crop of possible HR images and no unique solution exists. To tackle this ill-posed problem, many learning methods have been proposed, such as neighbor embedding methods [1], [2], sparse coding methods [3], [4], [5] and random forest methods [6]. Since Dong *et al.* firstly introduce a Super-Resolution Convolutional Neural Network (SRCNN) [7] to learn a nonlinear LR to HR mapping function, convolutional neural networks (CNNs) based SR has demonstrated outperformance over the traditional methods. Recently, design of neural networks with lightweight model size has attracted much attention for enabling edge computing with limited computation resources.

To improve the performance of SR, some researchers increase network depth [8] or apply recursive layers [9], [10] with appropriate training skills. In addition, various skip-connections have been proposed to enhance the expressiveness of the network structure [11], [12]. Despite achieving notable improvement, these strategies still pose some problems. Firstly, most of them take a bicubic interpolated LR image as input, which expands unnecessary computational cost and memory consumption. Secondly, most methods keep the upscaling part relatively simple compared with the feature extraction part [13]. For instance, EDSR [14] exploits 32 residual blocks for feature extraction, but only utilizes one upscaling module. Thirdly, most methods stack feature extraction modules to improve the performance but do not make full use of the features in shallow layers. These features are wasted halfway as the network depth increases [15]. Some methods combines the multi-stage features equally in an element-wise summation manner or a channel-wise concatenation manner without discrimination, which neglects the different importance among features.

Manuscript received October 11, 2019; revised January 21, 2020; accepted February 3, 2020. Date of publication March 23, 2020; date of current version December 9, 2020. This work was supported in part by MoE-CMCC "Artificial Intelligence" Project under Grant MCM20190701, in part by the Fundamental Research Funds for the Central Universities under Grant 2019PTB-011, and in part by the National Natural Science Foundation of China under Grant 61671077. *(Corresponding author: Honghui Zhu.)*

Zhuqing Jiang is with the School of Information and Communication Engineering, Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jiangzhuqing@bupt.edu.cn).

Honghui Zhu, Yue Lu, and Aidong Men are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhuhonghui510@163.com; manad@bupt.edu.cn).

Guodong Ju is with the R&D Center, GuangDong TUSHoldings TuWei Technology Company Ltd., Guangzhou 511493, China (e-mail: jgd@vip.163.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2020.2977513

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

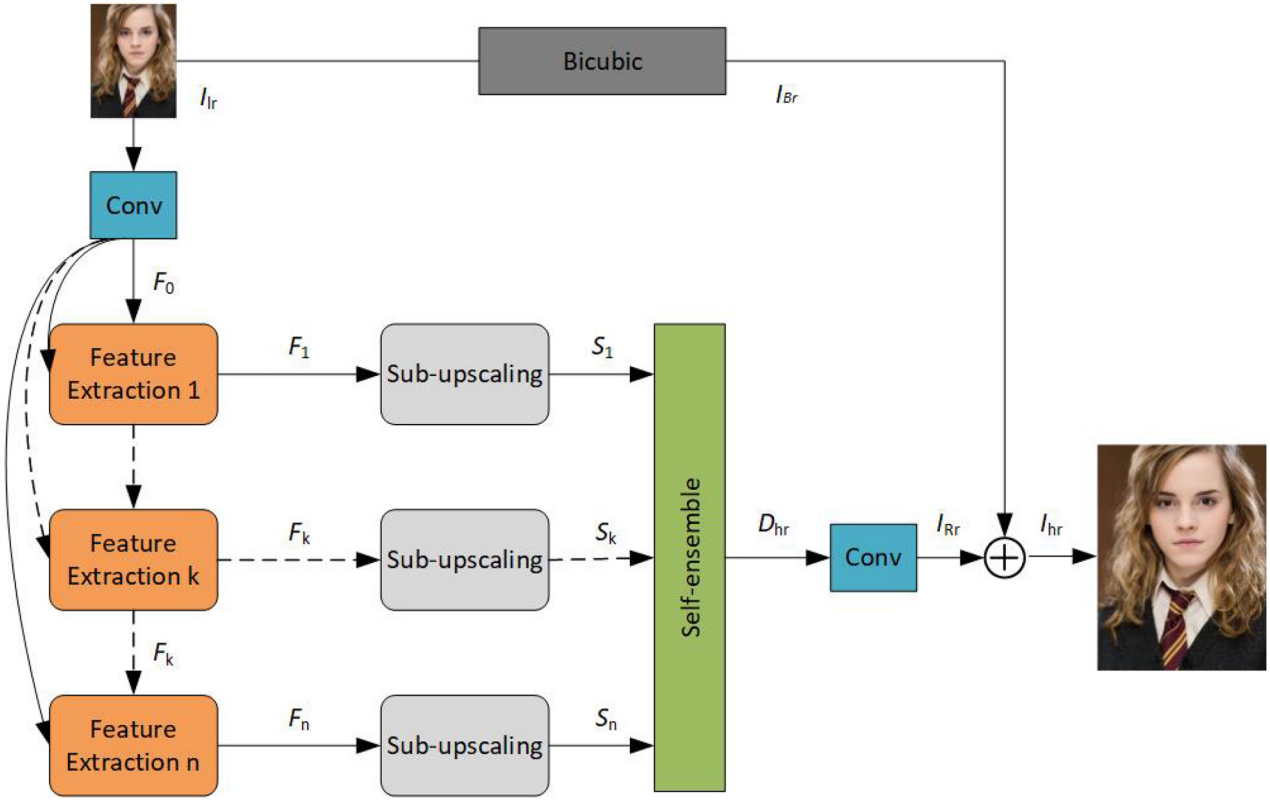

Fig. 1. Architecture of the proposed RMUN(for $\times 2$ SR).

Given the above-mentioned drawbacks, we propose a novel recursive multi-stage upscaling network (RMUN) to obtain local hierarchical features and global hierarchical features. The network structure is illustrated in Figure 1, which involves three modules: feature extraction module (FEM), sub-upscaling module (SUM), and self-ensemble module (SEM). The RMUN starts with a convolution layer that extracts features from an input LR image. Then, several FEMs are stacked to progressively generate multi-stage LR features. Simultaneously, the output LR features of each FEM are sent to a SUM to generate HR features. Finally, we discriminatively fuse the generated HR features by an SEM for the final reconstruction. The main contributions of our work are summarized as follows:

A novel pattern is designed for feature extraction that fuses local features hierarchically to obtain the multi-stage LR features. Specifically, recursive learning is adopted to control the model parameters while increasing the depth.

SUMs are applied to upscale the LR features, outputs of the FEMs, to corresponding HR features. This strategy provides shallow layer features of each stage an opportunity to generate HR features. It also offers multiple error feedback routes through which the final loss can supervise the learning of each sub-upscaling branch.

An SEM is employed to fuse the upscaled HR features discriminatively. This operation makes full use of global hierarchical features. Consequentially, the performance of our model is prompted.

The remainder of this paper is organized as follows. Section II reviews the popular single image SR algorithms in detail. In Section III, we introduce the proposed algorithm. The experiments are conducted in Section IV to verify the capability of our approach, and the conclusions are drawn in Section V.

II. Related Work

SR has been extensively studied in the literature. In this section, we review the recent works of SR. Besides, some works on feature extraction are presented.

A. Image Quality Assessment for Broadcasting

Image quality assessment (IQA) in broadcasting scenario concentrates on users' subjective feelings more than traditional scores. There has been plenty of research on this topic. Shishikui and Sawahata [16] conducted some subjective evaluations to evaluate the psychological effects of viewing UHR images. Multiple regression was utilized to study the sense of realness when the images are super resolved. In the contrast, Nafchi and Cheriet [17] contributed to the no-reference (NR) quality assessment by illustrating that a metric based on Minkowski distance and the entropy are able to predict the quality for the contrast distorted images. Similar to Nafchi, Min *et al.* [18] proposed a novel blind IQA that incorporates multiple pseudo reference images (MPRIs) with the full-reference IQA framework, where the MPRIs are obtained by degrading the already-distorted images in many ways and to certain degrees.

B. Traditional SR Methods

The traditional SISR algorithms are divided into three categories [19]: interpolation-based methods, reconstruction-based methods, and example-based methods. Early approaches apply

interpolation-based methods with sampling theory, such as linear interpolation [20], bicubic interpolation [21], and Lanczos resampling [22]. Those methods are excellent in efficiency but limited in accuracy. Reconstruction-based methods often exploit sophisticated prior knowledge to constrain the possible solution space. Shengyang *et al.* [23] utilized an edge smoothness prior knowledge to approximate the average length of all level lines in an intensity image, which performs well in rebuilding the detailed realistic textures. However, it is time-consuming and vulnerable to degradation when the scaling factor raises.

One type of example-based methods is self-example based methods which utilize the self-similarity property and extract example merely from the LR image across different scales. Huang *et al.* [24] extended self-similarity based SR to affine transformations. Nevertheless, this kind of methods is invalid for the textural appearance variations in the scene. Another type of example-based algorithms is external-example based methods. These methods focus on analyzing statistical relationships between the LR image and its corresponding HR image. Freeman *et al.* [1] employed Markov Random Field (MRF) with abundant real-world images, which synthesized visually pleasing image textures. Chang *et al.* [2] exploited neighbor embedding to restore HR image patches with the help of similar local geometry between the LR images and the HR images. Despite the effectiveness they have achieved, they are suboptimal because the extracted features and mapping functions are not adaptive.

C. Convolutional Neural Networks Based SR Methods

Different from the traditional SR methods, the convolutional neural network based SR methods rely on training a deep neural network to form the mappings of LR to HR patches. Dong *et al.* [7] applied a three-layers convolutional neural network named SRCNN to SR, achieving remarkable superiority over traditional SR methods at that time. To learn more expressive representations of SR, early methods deepen neural network architectures. VDSR [8] exploited a very deep network to obtain more hierarchical representations, which used a relatively big initial learning rate and residual learning to ease the difficulty of training. DRCN [9] stacked 16 recursive layers to learn the mapping from the bicubic to the residual where a multi-strategy reduces the burden of training.

Since it is hard for a plain architecture to go deeper, various skip-connections are widely used in further deep networks to enhance the expressiveness. Ying *et al.* [10] engaged basic residual units to form a recursive block, and all the blocks shared the same parameters as DRCN. Mao *et al.* [25] introduced encoder-decoder networks and symmetric skip connections into image restoration, showing that those nested skip connections provide fast and improved convergence.

Besides, several methods focus on more effective upsampling strategies. ESPCN [26] extracted features in LR space and proposed an efficient sub-pixel convolution in the final upsampling phase, avoiding overwhelming operations in HR space. LapSRN [27] proposed the Laplacian pyramid structure with Charbonnier loss to gradually estimate residual image, striking a balance between reconstruction quality and running time.

Considering that the combination of the shallow and deep features can enhance the construction quality, MemNet [28] proposed a very deep persistent memory network, tackling the long-term dependency problem. Hui *et al.* [29] combined an enhancement unit with a compression uint into a distillation block to effectively extract the local long and short-path features, which is superior to most prior methods in terms of accuracy and speed. MSRN [15] combines the outputs of each feature extraction blocks for global feature fusion to solve the problem that features disappear in the transmission process.

Even though these methods take advantage of the features in shallow and deep layers, they treat the multi-stage features equally and simply combine these features in an element-wise summation or just concatenate LR features without discrimination. Hence, it is suboptimal for feature integration. In the proposed RMUN, the multiple sub-upscaling strategy is applied for LR features upscaling and the concatenation is performed on the HR features. Furthermore, we utilize an SEM to recalibrate the HR features for better feature fusion.

D. Feature Extraction Modules

Recently, various FEMs have been proposed to extract efficient features of the input image. SRDenseNet [12] employed the DenseNet [30] structure as a FEM, where the feature maps of each layer are propagated into all subsequent layers. It provides an effective way to combine the shallow and deep features. Based on ResNet [31], Lim *et al.* [14] stacked residual blocks as FEMs to build a very wide network EDSR, achieving remarkable progress in SR. Moreover, Zhang *et al.* [32] proposed a residual dense module for local feature extraction, making full use of the hierarchical features. Unfortunately, these models increase the depth of network and are extremely difficult for training.

Different from the RDN [32] that the outputs of the preceding RDB and each layer have direct connections to all subsequent layers, our proposed FEM applies concatenations only in the mid-layer and the last layer, which obtains a good trade-off between the network complexity and performance. The concatenations are designed between local shallow and deep layers, which helps to fully exploit the local hierarchical features. In addition, a 1×1 convolution layer is applied to compress feature maps, which contributes to feature fusion and reduces computation complexity. We will give a more detailed description in Section III.

E. Upscaling Strategies

Although these methods present expressive features, their upscaling part is relatively simple compared with the feature extraction part. The deep learning based SR methods can be categorized into four types as follows.

Pre-upscaling [7], [8], [9], [10]: The input LR images are bicubic interpolated before entering the networks. The performance is limited because this upscaling method is not end-to-end learning and is suboptimal in terms of computational complexity [13].

Post-upscaling [33], [26], [11], [14]: The network is divided into two parts: the feature extraction part and the upscaling part and the upscaling part is located at the end of the networks.

Iterative up and downsampling is proposed by DBPN [34], which generates variants of the HR features using upsampling layers but also projects it back to the LR spaces using downsampling layers. The HR features from all of the upsampling layers are concatenated to reconstruct the HR image.

Multiple sub-upscaling is proposed by our network. We aim at increasing the upscaling rate of LR features in different depths and providing multiple error feedback routes. Hence, the LR features in shallow layers are supervised under the multiple upscaling strategy. The similarity between our RMUN and DBPN is that we both combine the HR features from multi-stages. However, our RMUN does not use downsampling layers, and we apply an SEM to recalibrate the HR features rather than directly concatenation.

III. Method

The technical details of the proposed RMUN is presented in this section. We first illustrate the network structure consisting of three branches and then elaborate on the essences: the feature extraction module (FEM), the sub-upscaling module (SUM), the self-ensemble module (SEM) and the loss function.

A. Network Structure

Our structure employs the Laplacian pyramid framework [27] to progressively predict residual images on the $log_2 S$ pyramid levels, where S is the upscaling factor. For simplicity, Figure 1 only demonstrates the $\times 2$ model as an example. Our model involves three branches: (1) an identity branch, (2) a feature extraction branch, and (3) a multi-path reconstruction branch.

1) Identity Branch: In this branch, the low frequency components of the LR image take a shortcut to the last layer, whereas the rest are restored by the complex treatment of other branches. The input LR image is bicubic interpolated instead of deconvoluted like LapSRN [27]. This operation obtains a pure HR image and obviates training difficulty.

2) Feature Extraction Branch: This branch extracts multi-stage features of the input LR image and refines features of different levels. By stacking FEM modules, we extract local hierarchical features of different stages from shallow to deep.

3) Reconstruction Branch: LR features of each stage are upscaled for corresponding HR features. Then we fuse the HR features and the interpolated original image to generate a final HR image. In this branch, SUM is engaged for hierarchical LR features upscaling and SEM for HR features upscaling and fusion.

B. Feature Extraction Module

FEM is the main module of the feature extraction branch that is adopted to progressively extract features. Figure 2 displays its internal implementation. Recursion framework is employed to extract and fuse local hierarchical features. In the k-th FEM, we denote $\boldsymbol{F}_{k-1}$, $\boldsymbol{F}_k \in \mathbb{R}^{H\times W\times C}$ as the input and output feature maps, where H, W, and C denote the height, width, and channel number of the feature maps. Let g_i denote a function of a convolutional layer: $g_i(F) = \boldsymbol{w}_i * F + \boldsymbol{b}_i$, weight and bias matrices are $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$, for i = 1, 2, 3, 4.

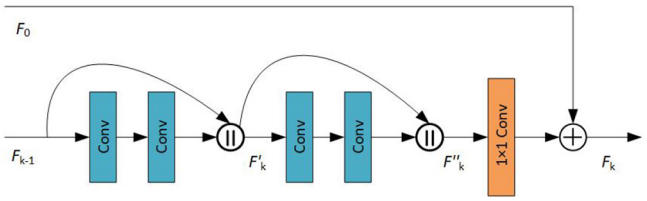

Fig. 2. Feature extraction module (FEM) architecture.

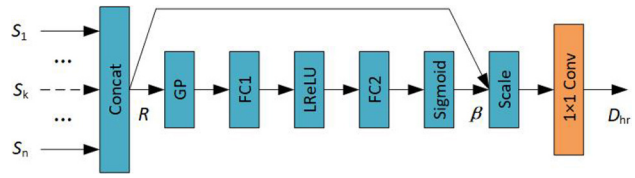

Fig. 3. Self-ensemble module (SEM) architecture.

Firstly, shallow layer feature maps are obtained by two convolutional layers, and then they are concatenated with the input $\boldsymbol{F}_{k-1}$ in channel dimension to generate the $\boldsymbol{F}'_k$:

$$\boldsymbol{F}'_k = g_2 \circ (g_1(\boldsymbol{F}_{k-1})) \quad || \quad \boldsymbol{F}_{k-1}, \tag{1}$$

where the operator $\circ$ denotes a function composition and || denotes concatenation.

Then we repeat (1) to generate deep layer feature maps and incorporate the shallow feature maps by concatenation:

$$\boldsymbol{F}''_k = g_4 \circ \left(g_3\left(\boldsymbol{F}'_k\right)\right) \quad || \quad \boldsymbol{F}'_k, \tag{2}$$

where the parameters of g_1, g_2, g_3, and g_4 are shared in each FEM.

Finally, we utilize a function $h(F)$ of a $1*1$ convolution layer to align the input dimension of each recursion. Besides, The compressed feature maps are added to the source LR feature maps $\boldsymbol{F}_0$ in case of the gradient vanishing problem:

$$\boldsymbol{F}_k = h\left(\boldsymbol{F}''_k\right) + \boldsymbol{F}_0. \tag{3}$$

C. Sub-Upscaling Module

Multiple SUMs are engaged to upscale the output LR features of each FEM, therein allowing F_k to generate the final HR features. This approach enhances the upscaling part and enables the network to preserve the HR features by learning various SUMs while generating deeper features.

We denote $\boldsymbol{S}_k \in \mathbb{R}^{SH\times SW\times C}$ as the output of k-th sub-upscaling module, where S denotes the upscaling factor. The sub-upscaling process is expressed as

$$\boldsymbol{S}_k = U_k(\boldsymbol{F}_k), \quad k = 1, 2, \ldots, n, \tag{4}$$

where $\boldsymbol{U}_k$ denotes the function of k-th upscaling. Each $\boldsymbol{F}_k$ is converted into various HR features that are regarded as the components of the desired HR features. In this work, we employ an LReLU activation function followed by a deconvolution layer.

Note that these multiple upscaling paths also serve as error feedback routes. In the back propagation, the loss is fed back to each FEM through multiple upscaling paths rather than one path, preventing the gradient vanishing. Moreover, it is potent for training deep networks because each FEM is supervised by the losses of both SUM and its subsequent layers.

D. Self-Ensemble Module

This module fuses the HR features obtained by SUMs. Huang *et al.* [30] observed that adding all features in the same feature space may disturb the information flow. Thus we denote the input as the concatenation of them:

$$\boldsymbol{R} = \boldsymbol{S}_1 \quad || \quad \boldsymbol{S}_2 \quad || \cdots || \quad \boldsymbol{S}_n, \tag{5}$$

where $\boldsymbol{R} \in \mathbb{R}^{SH \times SW \times nC}$ refers to the feature maps of integration.

Moreover, we apply an one-dimension vector $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k, \ldots, \beta_{nC}]$ to denote the weight for each channel, where β_k represents the calibration coefficient of k-th channel, which can selectively emphasize informative features and surpass less useful ones. Since it is supervised for the learning of $\boldsymbol{\beta}$, we introduce a corp of the SENet [35] for dimension compression. It is realized by a series connection of non-linearity operations. Following the concatenation, we prepare a global average pooling layer to extract the global information vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k, \ldots, \alpha_{nC}]$ across spatial dimensions $SH \times SW$:

$$\alpha_k = \frac{1}{SH \times SW} \sum_{i=1}^{SH} \sum_{j=1}^{SW} \boldsymbol{r}_{k(i,j)}, \tag{6}$$

where $\boldsymbol{r}_k \in \mathbb{R}^{SH \times SW}$ refers to k-th channel of $\mathbf{R}$, and $\boldsymbol{r}_k(i, j)$ is the pixel value of point (i, j).

Then the composition of two fully connected layers (denoted as $FC_1(\cdot)$ and $FC_2(\cdot)$) and two non-linear functions(denoted as δ and σ) is fulfilled to generate $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \sigma(FC_2(\delta(FC_1(\boldsymbol{\alpha})))), \tag{7}$$

where δ and σ refer to the functions of LReLU and sigmoid activation, respectively.

Finally, the integration features R are re-weighted by $\boldsymbol{\beta}$ and we further apply a 1×1 convolution layer (denoted as $S(\cdot)$) to compress it, which has been adopted in many SR methods [12], [29], [36]. This procedure is expressed as

$$\boldsymbol{D}_{hr} = S(\boldsymbol{R} \odot \boldsymbol{\beta}), \tag{8}$$

where $\boldsymbol{D}_{hr} \in \mathbb{R}^{SH \times SW \times nC}$ represents the outputs of DFM and $\odot$ refers to channel-wise multiplication between the channel of the feature maps $\boldsymbol{r}_k$ and the calibration coefficient β_k.

Let I_{lr} and I_{hr} be the input LR image and the output HR image, where $\boldsymbol{I}_{lr} \in \mathbb{R}^{H \times W}$ and $\boldsymbol{I}_{hr} \in \mathbb{R}^{SH \times SW}$. The global residual image (denoted as I_{Rr}) which is obtained by compressing feature dimensions of the D_{hr} from C to 1. On the identity branch, the blurry HR image $\boldsymbol{I}_{Br} \in \mathbb{R}^{SH \times SW}$ is generated by bicucic interpolation. Finally, we estimate the HR image I_{hr} via an element-wise summation:

$$I_{hr} = P(D_{hr}) + B(I_{lr}) = I_{Rr} + I_{Br}, \tag{9}$$

where $P(\cdot)$ compressed the feature dimensions from C to 1 and $B(\cdot)$ denotes the function of bicubic interpolation.

E. Loss Function

Mean square error (MSE) loss favors a high PSNR in super-resolution. Early methods applied MSE loss function as an optimization objective. The expression is as follows:

$$l_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left\| I_i - \hat{I}_i \right\|_2^2. \tag{10}$$

While as reported in [37] that MAE loss could guide an NN to reach a better local minimum with faster convergence, it is believed to be more robust against MSE loss. So we adopt MAE loss in this paper, it is formulated as follows:

$$l_{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left\| I_i - \hat{I}_i \right\|_1, \tag{11}$$

where $\hat{I}_i$ denotes the predicted HR image and I_i means the corresponding ground-truth image. N refers to the patch size.

IV. Experiments

In this section, we first brief the training and testing datasets used in our method. Then the details of our training setup are presented. In addition, we evaluate the manifestation of our method on several standard benchmark datasets. Finally, we discuss the effect of module components of our network respectively.

A. Datasets

For a fair comparison with the popular CNN-based methods, we select the training datasets from Lai *et al.* [27] for training. Note that the training datasets include 291 images, where 91 images are from Yang *et al.* (T91) [38], and the other 200 images are from Berkeley Segmentation Dataset (BSD200) [39]. Testing is executed on four widely applied benchmark datasets: Set5 [40], Set14 [41], BSD100 [39], and Urban100 [24], which involve 5, 14, 100, and 100 images, respectively. The Set5, Set14, and BSD100 datasets compromise different natural scenes, and the Urban100 dataset consists of many challenging images with details in different frequency bands.

B. Implementation Details

First, input images are preprocessed by randomly scaling, rotating, and flipping as [27]. Besides, all the RGB images are converted into YCbCr color space, and only the Y-channel is retained for training. The underlying reason is that human vision is more sensitive to changes in brightness than the chromatic aberration.

We employ 6 FEMs which share the same parameters in this work. The kernel size of each convolution layer (except the 1×1 one) is set to be 3×3 with the stride and padding of 1. Every convolution or deconvolution layer is followed by a leaky rectified linear units (LReLU) with a negative slope of 0.2 for non-linear mapping.

As training details, we utilize a batch size of 40 and crop the size of HR patches to 128×128. The initialization of the convolution filters is similar to the method of He *et al.* [42]. The

TABLE I
BENCHMARK RESULTS. AVERAGE PSNR/SSIM FOR SCALE FACTOR ×2, ×3, ×4, AND ×8 ON DATASETS SET5, SET14, BSD100, AND URBAN100. RED COLOR INDICATES THE BEST PERFORMANCE AND BLUE COLOR REFERS THE SECOND BEST

Dataset	Scale	Bicubic	VDSR	LapSRN	DRRN	MemNet	IDN	RMUN (Ours)
Set5	×2	33.66/0.9299	37.53/0.9590	37.52/0.9591	37.74/0.9591	37.78/0.9597	37.83/0.9600	37.77/0.9600
	×3	30.39/0.8682	33.66/0.9213	33.81/0.9220	34.00/0.9244	34.09/0.9248	34.11/0.9253	34.12/0.9251
	×4	28.42/0.8104	31.35/0.8830	31.54/0.8850	31.68/0.8888	31.74/0.8893	31.82/0.8903	31.84/0.8901
	×8	24.40/0.6580	25.93/0.7240	26.15/0.7380	26.18/0.7380	26.16/0.7414	-/-	26.27/0.7476
Set14	×2	30.24/0.8688	33.05/0.9130	33.08/0.9130	33.23/0.7136	33.28/0.7142	33.30/0.9148	33.21/0.9143
	×3	27.55/0.7742	29.77/0.8314	29.79/0.8325	29.96/0.8349	30.00/0.8350	29.99/0.8354	30.00/0.8360
	×4	26.00/0.7027	28.02/0.7680	28.19/0.7720	28.21/0.7721	28.26/0.7723	28.25/0.7730	28.32/0.7750
	×8	23.10/0.5660	24.26/0.6140	24.35/0.6200	24.42/0.6220	24.38/0.6199	-/-	24.58/0.6296
BSD100	×2	29.56/0.8431	31.90/0.8960	31.08/0.8950	32.05/0.8973	32.08/0.8978		32.02/0.8979
	×3	27.21/0.7385	28.82/0.7976	28.82/0.7980	28.95/0.8004	28.96/0.8001		28.94/0.8016
	×4	25.96/0.6675	27.29/0.7260	27.32/0.7270	27.38/0.7284	27.40/0.7281		27.44/0.7314
	×8	23.67/0.5480	24.49/0.5830	24.54/0.5860	24.59/0.5870	24.58/0.5842		24.66/0.5918
Urban100	×2	26.88/0.8403	30.77/0.9140	30.41/0.9101	31.23/0.9188	31.31/0.9195	31.27/0.9196	31.10/0.9181
	×3	24.46/0.7349	27.14/0.8279	27.07/0.8275	27.53/0.8378	27.56/0.8376	27.42/0.8359	28.11/0.8359
	×4	23.14/0.6577	25.18/0.7540	25.21/0.7560	25.44/0.7638	25.50/0.7630	25.41/0.7632	25.50/0.7663
	×8	20.74/0.5160	21.70/0.5710	21.81/0.5810	21.88/0.5830	21.89/0.5825	-/-	22.07/0.5978

Fig. 4. SR results of "barbara" (Set14) with a scale factor ×4. The line is straightened and clear in our results and the DRRN, whereas other methods behave a curved trend.

initial learning rate is set to be 10^{-5} and then decreased by half every 150 epochs. Also, the optimization is using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and the weight decay of 10^{-4}. The proposed method is implemented in MATLAB based on MatConvNet [43], and runs at an NVIDIA GTX 1080TI GPU.

C. Comparisons With State-of-the-Art Methods

We provide quantitative and qualitative comparisons with 6 state-of-the-art SR algorithms, including Bicubic [44], VDSR [8], LapSRN [27], DRRN [10], MemNet [28], and IDN [29]. The public codes for VDSR, LapSRN, DRRN, and MemNet are employed as the benchmarks. The peak signal-to-noise ratio (PSNR) and the structure similarity (SSIM) are adopted as evaluation metrics. Both PSNR and SSIM are the most common and widely applied image evaluation indicators.

Table I exhibits quantitative comparisons of scale factor ×2, ×3, ×4, and ×8 over the four benchmark datasets. Results demonstrate that our method achieves the best performance and surpasses the prior methods by a considerable margin for ×4 and ×8 on all testing datasets. As for ×2 SR, we obtain relatively poor performance, which is probably owing to the pyramid structure is more expressive for larger scale factors. Comprehensive analysis indicates that our model with the enhanced sub-upscaling strategy and the discriminative SEM can handle large scale SR tasks better.

We present the visual comparisons with state-of-the-art SR methods in Figure 4, 5, and 6 for the qualitative analysis. These images embody rich high frequency information, therefore they are challenging for SR. In the image 'barbara' displayed in Figure 4, all these methods fail to recover the stripe trend of the read box correctly. It is possibly due to the severe loss of high frequency information in the downsampling process which can be inferred by the result of bicubic interpolation. For visual perception, only our RMUN and the DRRN recover roughly the outline of several stacked books. In other examples of Figure 5 and 6, our method achieves the optimal results for both the PSNR and the SSIM. Besides, our method gains clear

Fig. 5. SR results of "8023" (BSD100) with a scale factor ×4. Our result shows unobstructed separation between stripes while in other methods, stripes are vague.

Fig. 6. SR results of "img027" (Urban100) with scale factor ×4. Lines of the building are sharp while building edges are blurry in other methods.

contours without serious artifacts while other methods explicit different degrees of fake information.

D. Discussion and Analysis

1) Model Parameters: We present the trade-off between the reconstruction performance and the number of network parameters of CNN-based SR methods in Figure 7. For the sake of low memory consumption, we stack multiple compact but effective FEMs, which are recursive to reduce parameters. Our RMUN outperform SRCNN, FSRCNN, DRRN, and VDSR. Moreover, our RMUN performs better than LapSRN with 19% fewer parameters on 4× upscaling. As is shown in Figure 7 that MSRN and D-DBPN outperform our RMUN. However, our RMUN has about 89% and 93% fewer parameters than MSRN and D-DBPN, respectively. This evidence demonstrates that our network obtains a good trade-off between performance and the number of parameters. This lightweight network is applicable for edge computing with limited computation resources.

2) Visualization of the Feature Maps: The function of the SUM will be illustrated in this part. For better visualizing the intermediary of the proposed model, we consider an operation T that can transform a $3D$ tensor M to a flatted $2D$ tensor defined over the spatial dimensions. It is formulated as follows:

$$T : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{H\times W}, \tag{12}$$

where H, W, and C denote the height, width, and channel dimensions of the feature maps. Specifically, we take the mean of the feature maps over channel dimensions to visualize the outputs of each SUM, which is described by

$$T_{mean}(M) = \frac{1}{C}\sum_{i=1}^{C} M_i, \tag{13}$$

where M refers to a $3D$ tensor and $M_i = M(i, :, :)$. We can induce that the average feature map is an approximate representation of the whole feature maps. Besides, we calculate the corresponding average weights which are obtained by the SEM. These average weights are marked below subfigures. As illustrated in Figure 8, average feature maps gradually increase the pixel values of edge texture from the Figure 8 (1) to (3). It indicates that the FEMs in shallow layers mainly restore the local texture details of the image. From Figure 8 (4) to (6), we see that the pixel values are higher than those of the first three subpictures, and they have less edge detail information. It infers that the FEMs in deep layers mainly reserve the global brightness information rather than texture details. According

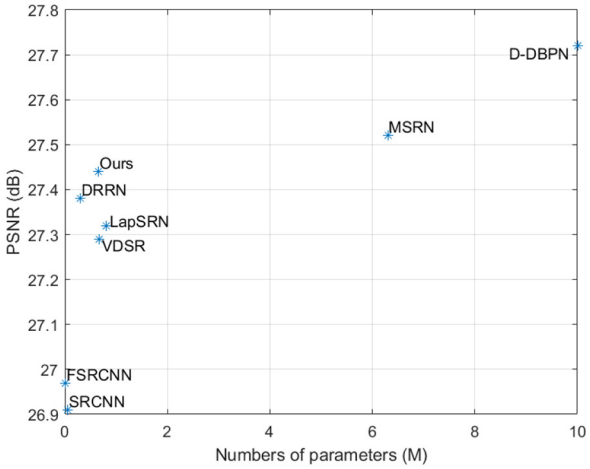

Fig. 7. Performance vs number of parameters. The results are evaluated with BSD100 dataset for 4× upscaling.

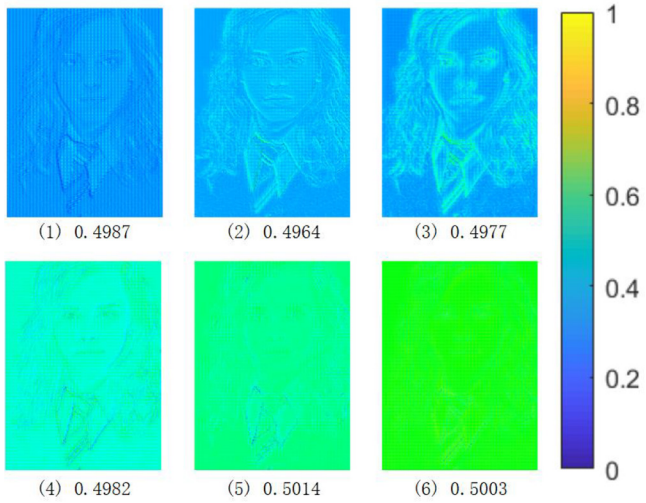

Fig. 8. The average feature maps of SUMs.

TABLE II
ABLATION EXPERIMENTS OF RMUN ON SET14

SEM	SUM	×4 SR	×8 SR
		28.17	24.45
✓		28.24	24.51
	✓	28.19	24.49
✓	✓	**28.25**	**24.55**

to the distribution of the average weights, the values in the last three average maps are higher than those in the first three average maps. We infer that the feature maps in deep layers include more beneficial information for HR image generation. Furthermore, the difference of values between shallow and deep layers is small, which means that both the local texture information and global brightness information are essential for HR reconstruction.

3) Ablation Study: As discussed in Section III, our RMUN contains two main components including SUM and SEM. We conduct ablation studies to verify the contributions of the two modules in the pipeline. We remove the SUMs and SEM as the baseline. From Table II we can validate that each component contributes to the final reconstruction. It is worth noted that the SUM only has one deconvolution layer, which is relatively simple compared to the feature extraction module. Besides, the SEM has about 0.05M parameters, which is 7% of the network parameters. Hence, we believe that the performance gain is mainly brought by the effectiveness of the two components rather than the complexity. If we adopt SEM on the basis of SUM, the network will obtain more gain. This confirms our supposition that direct integration of multi-stage features without discrimination may not enhance performance. Furthermore, with the growth of scale factors, SUM can provide multiple upscaling operations which are especially beneficial for large scale enlargement. Overall, the ablation study verifies the contributions of the SUM and SEM.

4) Running Time and Potential Feasibility in Broadcasting: Our technique could be valid in the broadcasting given the test of running time that the real-time processing time is 0.058 seconds. Thus it would be potentially feasible in broadcasting and real-time scenario after it was revised slightly according to a specific application.

V. CONCLUSION

In this work, we demonstrate that most existing deep learning methods with cascaded topology fail to incorporate the shallow and deep features. Besides, most methods generate the final HR image only using the last layer, which does not make full use of the information of the shallow and deep layers. We propose an RMUN with discriminative self-ensemble for boosting SR. The proposed network adopts recursive FEMs to extract the shallow local texture information and the deep global brightness information. Then, we provide the LR features obtained from every FEM with an opportunity to estimate the HR features by using SUMs. Finally, an SEM is applied to fuse those features discriminatively for the final reconstruction. The comprehensive experiments have illustrated that our proposed RMUN network achieves competitive performance both in quantitative and qualitative comparisons. Particularly, RMUN outperforms state-of-the-art methods for large factors significantly.

REFERENCES

[1] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.
[2] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, p. 1.
[3] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.
[4] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Comput. Vis. (ACCV)*, vol. 9006, Mar. 2015, pp. 111–126.
[5] W. Yang, Y. Tian, Z. Fei, Q. Liao, C. Hai, and C. Zheng, "Consistent coding scheme for single-image super-resolution via independent dictionaries," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 313–325, Mar. 2016.
[6] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3791–3799.
[7] 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. 2016.

[8] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[9] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.
[10] T. Ying, Y. Jian, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2790–2798.
[11] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[12] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.
[13] J.-H. Kim and J.-S. Lee, "Deep residual network with enhanced upscaling module for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, pp. 800–808.
[14] 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 (CVPRW)*, Jul. 2017, pp. 1132–1140.
[15] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 527–542.
[16] Y. Shishikui and Y. Sawahata, "Effects of viewing ultra-high-resolution images with practical viewing distances on familiar impressions," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 498–507, Jun. 2018.
[17] H. Z. Nafchi and M. Cheriet, "Efficient no-reference quality assessment and classification model for contrast distorted images," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 518–523, Jun. 2018.
[18] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.
[19] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.
[20] A. Bücken and J. Rossmann, "An efficient, fractal-based, bi-linear algorithm for the interpolation of inhomogeneously distributed geo-data," in *Proc. RSPSOC Conf.*, Sep. 2011, pp. 13–15.
[21] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.
[22] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.
[23] D. Shengyang, H. Mei, X. Wei, W. Ying, G. Yihong, and A. K. Katsaggelos, "SoftCuts: A soft edge smoothness prior for color image super-resolution," *IEEE Signal Process. Soc.*, vol. 18, no. 5, p. 969, May 2009.
[24] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[25] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," Jun. 2016. [Online]. Available: arXiv:1606.08921.
[26] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[27] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-S. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5835–5843.
[28] Y. Tai, J. Yang, X. Liu, and C. Xu, "MEMNET: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4549–4557.
[29] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 723–731.
[30] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[33] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *CoRR*, vol. abs/1608.00367, pp. 391–407, Sep. 2016.
[34] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," *CoRR*, vol. abs/1803.02735, pp. 1664–1673, Jun. 2018.
[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access.
[36] J. Xu, Y. Chae, B. Stenger, and A. Datta, "Dense BYNET: Residual dense network for image super resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 71–75.
[37] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[38] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[39] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[40] M. Bevilacqua, A. Roumy, C. Guillemot, and M. A. Morel, "Neighbor embedding based single-image super-resolution using semi-nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1289–1292.
[41] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*, 2012, pp. 711–730.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[43] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[44] C. D. Boor, "Bicubic spline interpolation," *J. Math. Phys.*, vol. 41, no. 3, pp. 212–218, 1962.

Zhuqing Jiang received the B.S. degree from Beijing Forestry University in 2008 and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, in 2014, where he is currently pursuing the Ph.D. degree. Since 2014, he has been a Lecturer in communication and information engineering with the BUPT. He has published several papers in journals and international conference. His research interests include satellite communications and multimedia signal processing.

Honghui Zhu received the B.S. degree from the Beijing University of Posts and Telecommunications in 2018, where she is currently pursuing the master's degree. Her research interests include image processing and image super-resolution.

Yue Lu received the B.S. degree from the Beijing University of Posts and Telecommunications in 2017, where he is currently pursuing the master's degree. He has published several papers in international conference. His research interests include image processing and image super-resolution.

Guodong Ju is the Manager of GuangDong TUSHoldings TuWei Technology Company Ltd., Guangzhou, China. He is dedicated to researching the industrial resolution of video encoding and decoding.

Aidong Men received the B.S., M.S., and Ph.D. degrees from the Department of Radio Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, in 1994, where he was an Associate Professor with the Department of Radio Engineering from 1994 to 2000. Since 2000, he has been a Professor with the Telecom Engineering College, BUPT. He has published over 100 papers in journals and international conference. His research interests include multimedia communication, digital TV, and images and speech signal processing, and transmission. He is currently a fellow of the Chinese Institute of Electronics and the China Institute of Communications. He is also an invited fellow of the Science and Technology Committee of State Administration of Radio, Film and Television.