IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 May 2023, accepted 5 June 2023, date of publication 9 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284544

RESEARCH ARTICLE

Image Super-Resolution Based on Residual Attention and Multi-Scale Feature Fusion

QIQI KOU[1], JIAMIN ZHAO[2], DEQIANG CHENG[2], (Member, IEEE), ZHEN SU[2], AND XINGGUANG ZHU[2]

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[2]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: Deqiang Cheng (chengdq@cumt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 52204177.

ABSTRACT At present, deep residual network has been widely used in image super-resolution and proved to be able to achieve good reconstruction results. However, the existing super-resolution algorithms based on deep residual network have the problems of indiscriminately learning feature information of different regions and low utilization rate of feature information, which make them difficult to further improve the reconstruction effect. In view of the above problems, a novel super-resolution reconstruction network based on residual attention and multi-scale feature fusion (RAMF) is proposed in this paper. Firstly, a lightweight multi-scale residual module (LMRM) is proposed in the deep feature extraction stage, by which the multi-scale features are extracted and further cross-connected to enrich the information of different receptive fields. Then, to fully improve the utilization rate of feature information, a dense feature fusion structure is designed to fuse the output feature of each LMRM. Finally, a residual spatial attention module (RSAM) is proposed to specifically learn and better retain high-frequency feature information, so as to improve the reconstruction effect. Experimental tests and comparisons are conducted with the current advanced methods on four baseline databases, and the results demonstrate that the proposed RAMF can achieve better reconstruction effect with fewer parameters, low computational complexity, fast processing speed and high objective evaluation index. Especially, the peak signal-to-noise ratio measured on Urban100 data set increases by 0.13dB on average, and the reconstructed image has better visual effect and richer texture detail features.

INDEX TERMS Dense feature fusion, attention mechanism, super-resolution, residual learning.

I. INTRODUCTION

How to reconstruct high-resolution (HR) images with high definition and rich detail information from low-resolution (LR) images is a hot research topic at home and abroad [1]. Since the super-resolution (SR) reconstruction technology has the advantages of easy implementation, low hardware dependence and low cost, it has been widely used in the fields of pedestrian re-identification, image enhancement [2] and texture classification [3], etc.

Since 1960s, large numbers of image super resolution algorithms have been proposed one after another. Harris was the first to study image super-resolution reconstruction and proposed Harri spectrum extrapolation method [4], which laid a foundation for subsequent research on image super-resolution methods. On the basis of Harris's research, Tsai and Huang [5] proposed to obtain HR images through frequency domain transformation of multi-frame LR images. Since then, image super-resolution reconstruction technology has gradually gained the attention of researchers, and has been explored for decades. However, since single image super resolution (SISR) reconstruction is an ill-posed problem, there are always multiple HR images corresponding to the same LR image. Therefore, how to solve this problem has been the focus of scholars' attention. In recent years, with the vigorous development of deep learning, image super resolution reconstruction algorithm based on convolutional neural network [6], [7], [8] has made many achievements.

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

As the existing SR models based on convolutional neural networks mainly focus on designing deepened or widened networks [9], [10] while ignoring the loss of high-frequency feature information of images, scholars have proposed a series of solutions to this problem. For example, by using convolution kernels with different scales of 1×1, 3×3 and 5×5 to extract rich feature information, Zhang and Guo [11] obtained reconstructed images with better subjective visual effects and objective evaluation indexes. By combining multi-scale convolution kernel with residual structure, Lu et al. [12] proposed a multi-scale information aggregation block to extract image features efficiently without increasing the number of parameters. By adding extended convolution branches to the residual block to expand the receptive field, Esmaeilzehi et al. [13] also proposed a multi-scale network, which can obtain superior performance with fewer parameters.

Although the multi-scale convolution kernel can effectively improve the model reconstruction effect, the same calculation method is adopted for the feature information of different positions when learning the feature information collected by the upper layer. As a result, the high-frequency information with high contribution to the reconstruction effect cannot be fully learned, resulting in a waste of computing resources. Motivated by the fact that people usually focus on the parts conducive to brain analysis in images when watching them, Xu et al. [15] proposed the attention mechanism, which can effectively utilize computing resources and improve reconstruction performance. Subsequently, to improve the utilization rate of high-frequency feature information in LR images and solve the problem of network degradation in deep networks, Zhang et al. [14] combined the residual structure with channel attention and proposed the RIR (ResNet in ResNet) residual structure to adaptively scale the feature information of each channel. Thus, the robustness of the model and the utilization rate of image feature information are improved. By integrating the spatial attention and channel attention, Woo et al. [16] propose a convolutional block attention module (CBAM), which can effectively enable the network to focus on learning the image areas with more high-frequency feature information and extract the important features of the images accordingly. To further improve the utilization rate of image high-frequency information and multi-scale features, Lu et al. [17] added long-short jump connections and mixed attention mechanisms into the model, leading to the reconstruction effect of image edge information and texture structure information improved. By combining a multi-scale detail extraction block with a multi-content information channel attention module to enhance image detail, Wang and Zheng [10] proposed a multi-scale detail enhancement network (MS-DEN), which can restore more accurate detail and achieve better reconstruction effect.

However, although the above methods have improved the model performance to a certain extent, most of them still have the following problems: (1) low utilization rate of feature information. The design of feature extraction module is too complicated and there are redundant parameters, which cannot effectively extract the image feature information. (2) Undifferentiated learning of upper feature information. Using the same learning method to learn the image features of different regions, it will not be able to treat and learn the high-frequency feature information pointedly.

In order to solve the above problems, a super-resolution reconstruction network based on residual attention and multi-scale feature fusion (RAMF) is constructed for all kinds of LR images. In the proposed network, the key contributions can be summarized as follows:

1) A lightweight multi-scale residual module (LMRM) is proposed, which can obtain abundant image feature information of different sensitivity fields with fewer parameters.

2) a dense feature fusion structure is designed, which can fully fuse the output feature information of each LMRM and improve the utilization rate of feature information.

3) We develop a residual spatial attention module (RSAM), which can specifically learn high-frequency feature information and reasonably allocate computing resources.

4) For experimental evaluation, the proposed RAMF has the advantages of fewer parameters, low computational complexity, fast processing speed and high objective evaluation index, which can achieve better reconstruction effect.

The rest sections are organized as follows. Section II presents the details of the proposed RAMF. T the experimental results and analysis are shown in Section III, and the conclusions and future work are summarized in Section IV.

II. PROPOSED METHOD

In this section, a novel super-resolution reconstruction algorithm based on residual attention and multi-scale feature fusion (RAMF) is proposed and its overall structure is shown in Fig. 1. As can be seen, the network structure of RAMF consists of four parts, which are shallow feature extraction module, deep feature extraction module, residual spatial attention module and image reconstruction module.

Among them, the shallow feature extraction module contains a 3×3 convolution layer. The deep feature extraction module is composed of two kinds of cross-connected lightweight multi-scale residual modules with different channel numbers, and the number of channels output by convolutional kernel in adjacent lightweight multi-scale residual modules is also different. In addition, the exterior of the module adopts a dense feature fusion structure to fully integrate image features of different depths to reduce the loss of feature information. Subsequently, the residual spatial attention module is proposed and used to further enhance the high frequency feature information and suppress the low frequency feature information, so as to allocate computing resources reasonably. The final image reconstruction module consists of an upper sampling layer and a reconstruction layer, and the principles of the four main parts are described in detail in the following subsections.

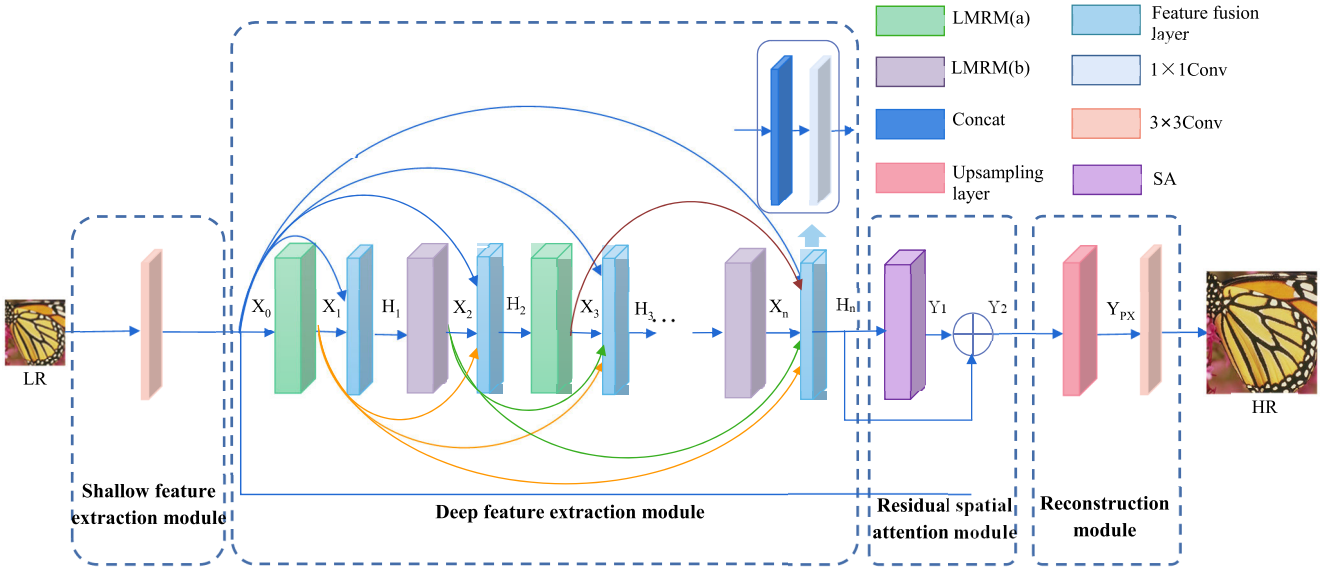

FIGURE 1. The structure diagram of the proposed RAMF network.

A. SHALLOW FEATURE EXTRACTION

The module contains a convolutional layer, and the process of extracting shallow features can be expressed as:

$$X_0 = \omega_{3\times3} * I^{LR} + x_0 \tag{1}$$

where I^{LR} represents the input of the model, $\omega_{3\times3}$ and x_0 represent the weight and bias of the convolutional layer respectively, and X_0 represents the extracted shallow feature, which serves as the input of the deep feature extraction layer.

B. DEEP FEATRUE EXTRACTION

As can be seen from Fig. 1, the deep feature extraction module is composed of n LMRM modules, including $n/2$ LMRM (a) modules and $n/2$ LMRM (b) modules. Besides, LMRM (a) contains multi-scale feature fusion block (a), i.e. MFA(a), and LMRM (b) contains multi-scale feature fusion block (b), i.e. MFA(b), and each MFA uses a short residual line to connect the outside. The detailed structure of the LMRM is presented in Fig. 2. For these n LMRM modules, a dense feature fusion structure is constructed among them, and the input of each LMRM is fused by the output features of all the previous LMRM modules and shallow features to make full use of the image feature information and reduce the number of network parameters. The principle of dense feature fusion structure can be formulated as:

$$H_k = \omega^k_{1\times1} * [X_0, X_1, X_2, \cdots, X_{k-1}, X_k] + b^k \tag{2}$$

where X_k defines the output of the k-th LMRM, X_0 represents the shallow feature extracted by the shallow feature extraction module, H_k represents the output of the k-th feature fusion layer, $[\cdot]$ denotes the splicing operation, $\omega^k_{1\times1}$ and b^k respectively represent the weight and bias of the 1×1 convolutional layer in the kth feature fusion layer.

As can be seen from Fig. 2, each LMRM contains two MFA modules, each of which uses a short residual line to connect to the outside. Wherein, the number of input and output channels of convolutional kernel in LMRM (a) is set to 64, the number of input channels and output channels of convolutional kernel in module LMRM (b) is set to 64 and 128. Let the input feature of the k-th LMRM be H_{k-1}. Then, the features are extracted using convolution kerns with sizes of 3×3 and 5×5 respectively, and then activated by ReLU activation function respectively to obtain features T_k and X_k of different scales.

In order to prevent feature loss and make full use of features with different scales, the multi-scale feature fusion block adopts the feature fusion structure to fuse T_k, X_k and input features H_{k-1} to get fusion feature O_k. Finally, the input feature H_{k-1} and fusion features are added to get X_k. The detailed calculation process of the proposed LMRM can be expressed by the following formulas:

$$T_k = \sigma(\omega^k_{3\times3} * H_{k-1} + b^k_{3\times3}) \tag{3}$$

$$X_k = \sigma(\omega^k_{5\times5} * H_{k-1} + b^k_{5\times5}) \tag{4}$$

$$O_k = \omega^k_{1\times1} * [T_k, X_k, H_{k-1}] + b^k_{1\times1} \tag{5}$$

$$X_k = O_k + H_{k-1} \tag{6}$$

where $\omega^k_{3\times3}$, $\omega^k_{5\times5}$ and $\omega^k_{1\times1}$ represent the weights of convolution layers of different scales in the k-th LMRM, respectively. $b^k_{3\times3}$, $b^k_{5\times5}$ and $b^k_{1\times1}$ define the bias of convolution layers of different scales, respectively. $\sigma(\cdot)$ represents the mapping function of ReLU activation function. $[\cdot]$ denotes the splicing operation, and H_k represents the output of the k-th LMRM.

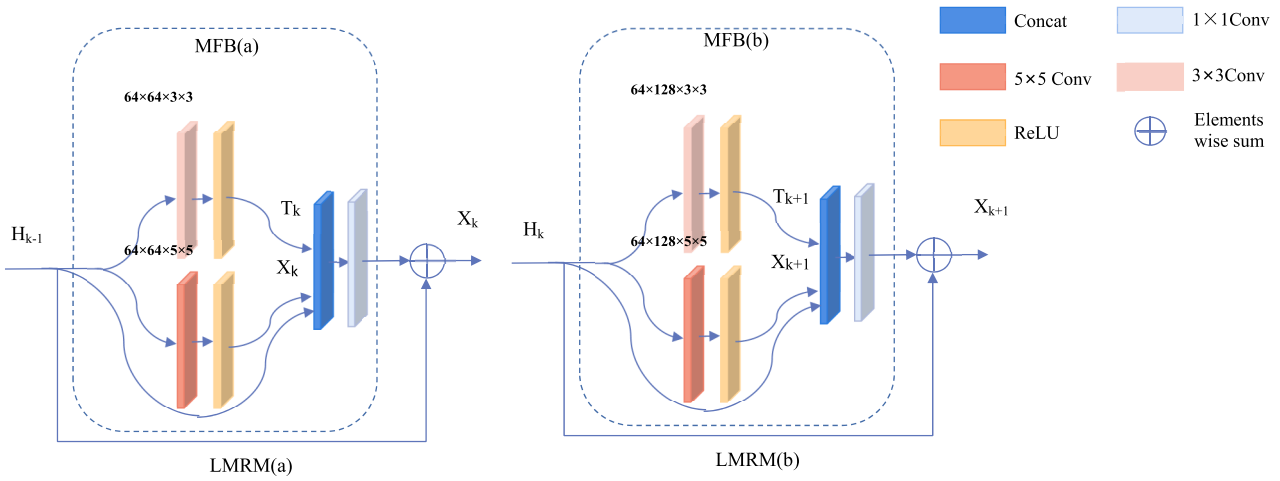

FIGURE 2. The structure diagram of the proposed lightweight multi-scale residual module (LMRM).

C. RESIDUAL SPATIAL ATTENTION MODULE (RSAM)

Usually, different positions of each feature map have different feature information. However, most of the feature information is low-frequency feature information whose details or color changes gently and does not need to allocate a lot of computing resources for learning. However, a small amount of important high-frequency details such as edge information and texture features in the image often need to be calculated with emphasis. Indiscriminately calculating high-frequency and low-frequency information will not only cause a serious waste of computing resources, but also can not retain useful high-frequency information well, which reduces the performance of model reconstruction. Therefore, by adding a residual spatial attention module (RSAM) to the deep feature extraction module and using the spatial attention mechanism (SA) to increase the network's attention to high-frequency feature information, the proposed RAMF can carry out targeted learning of high-frequency features with high contribution to model reconstruction performance, effectively solving the problem of undifferentiated learning of feature information from different regions in feature graphs by CNN. The internal structures of the RSAM and its internal SA are shown in Figure 1 and Figure 3 respectively.

Let's start with SA and look at Fig. 3, it firstly calculates the average pooling and maximum pooling of input feature graphs with the size of H × W × C according to the direction of channel axis, and all the sizes of the feature graphs after pooling are H × W × 1. Then, the average pooling values and the maximum pooling values are spliced into a two-channel feature graph, and a 7 × 7 convolution check is used to fuse the two groups of pooling values, leading to a single channel fusion feature graph output. Finally, the attention feature map is obtained after activation by Sigmoid function, and the values of elements in different positions of the feature map represent the different learning weights of corresponding positions of the input feature map.

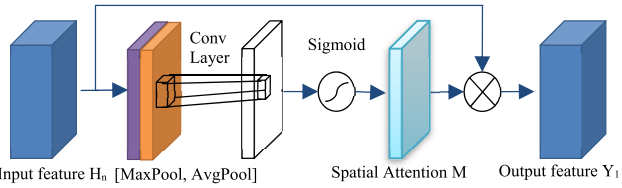

FIGURE 3. Structure diagram of the spatial attention mechanism.

Let's come to RSAM, whose input information is the output feature H_n of the deep feature extraction module. T represents the number of input feature map channels and H_n^t is the feature information of the t-th channel in the feature map. Each channel contains $H \times W$ elements, and $H_n^t(h, w)$ represents the value of element in h-th row and w-th column of t-th channel. The input feature map H_n is used to calculate the average pooling feature map H_n^{Avg} and the maximum pooling feature map H_n^{Max} respectively according to the direction of the channel axis.

The element values at different positions of the two pooling feature maps can be represented as $H_n^{Avg}(h, w)$ and $H_n^{Max}(h, w)$, where h and w represented the rows and columns that the element values located. The detailed pooling process is as follows:

$$H_n^{Max} = \underset{t\in\{1,\cdots,T\}}{Max} H_n^t(h, w) \tag{7}$$

$$H_n^{Avg} = \frac{\sum_{t=1}^{T}(H_n^t(h, w))}{T} \tag{8}$$

The average pooling feature graph H_n^{Avg} and maximum pooling feature graph H_n^{Max} calculated in Equations 7 and 8 are first spliced and then a 7 × 7 convolution kernel is used for fusion, so as to obtain fusion feature A_r. The principle is given by:

$$A_r = W_a^{7\times7} \times [H_n^{Max}, H_n^{Avg}] + b_a \tag{9}$$

where $W_a^{7\times7}$ and b_a represent the weight and bias of 7×7 convolution kernel respectively, $[H_n^{Max}, H_n^{Avg}]$ defines the concatenation process of average pooling feature map H_n^{Avg} and maximum pooling feature map H_n^{Max}, and A_r represents the single-channel position feature map output by the convolutional layer. Finally, sigmoid activation function is used to activate the location feature map A_r to obtain SA feature A_a.

$$A_a = \sigma(A_r) \tag{10}$$

where σ indicates the sigmoid activation function, and SA feature A_a denotes the position feature map of single channel. The element values of different positions represent the different weights of corresponding positions of the input feature map. Then, the output feature map Y_1 can be multiplied by the input feature map H_n.

$$Y_1 = A_a * H_n \tag{11}$$

In this way, different location feature information is learn specifically according to its contribution to model reconstruction performance, which can not only realize the enhancement of high frequency features and suppression of low frequency features, but also make full use of computing resources and effectively improve the model reconstruction performance.

In addition, in order to reduce the loss of feature information and alleviate the problem of network degradation, the RSAM in RAMF is connected by long and short lines. Subsequently, the input features, shallow features and attention-added feature maps of the module are added, respectively. The detailed calculation principle is as follows:

$$Y_2 = H_n + X_0 + Y_1 \tag{12}$$

where Y_1 represents the output feature map with added attention, X_0 defines the shallow image feature extracted by the shallow feature extraction module, and H_n represents the input feature of RSAM.

D. IMAGE RECONSTRUCTION

As can be seen from Fig. 1, the image reconstruction module is mainly composed of the up-sampling layer and the reconstruction layer. Considering that subpixel convolution has the advantages of faster speed and better effect compared with interpolation-based method and deconvolution method, this paper adopts it for image up-sampling. The principle of image reconstruction module can be expressed by the following formula:

$$Y_{CN} = H_{CN}(Y_n) \tag{13}$$

$$Y_{PX} = H_{PX}(Y_{CN}) \tag{14}$$

$$I^{SR} = H_{RC}(Y_{PX}) \tag{15}$$

where Y_n represents the input of image reconstruction module; $H_{CN}(\cdot)$, $H_{PX}(\cdot)$ and $H_{RC}(\cdot)$ denote the mapping function of convolutional layer, pixel recombination layer and reconstruction layer, respectively. Y_{CN} and Y_{CN} represent the output of convolutional layer and pixel recombination layer respectively. I^{SR} defines the reconstructed high-resolution image.

In addition, L1 loss function is adopted to train the network, which can be expressed by the following formula:

$$S_{L1} = \frac{1}{MN}\sum_{x=0}^{M}\sum_{y=0}^{N}|f_{SR}(x,y) - f_{HR}(x,y)| \tag{16}$$

where $f_{SR}(x, y)$ represents the reconstructed image and $f_{HR}(x, y)$ denotes the real image. M and N are the width and height of the image, and S_{L1} defines the calculated L1 loss function.

III. EXPERIMENTAL RESULTS

A. DATASETS AND EXPERIMENTAL SETUP

In this section, to validate the superiority of our RAMF, comprehensively comparative evaluations with the existing methods are carried out on the server based on Ubuntu20.04 system, using the deep learning framework, and configured as pytoch1.8, Cuda11.4, and NVIDIA RTX 3090Ti graphics card. For the datasets, this paper uses DIV2K dataset as the training dataset, and four benchmark datasets, Set5 [18], Set14 [19], BSD100 [20], Urban100 [21] as the test set.

For parameter setting, the model parameters in this paper are set as: the number of LMRM is 16, in which the number of LMRM(a) and LMRM(b) are both 8, and the two are directly cross-connected. Besides, the input and output channels of 3×3 and 5×5 convolution nuclei in LMRM (a) are both 64, while the input channels and output channels of 3×3 and 5×5 convolution nuclei in LMRM (b) are 64 and 128 respectively. Moreover, dense feature fusion structure is adopted to connect different LMRM. In addition, the initial learning rate is set as 10^{-4}, and the learning rate is halved for every 200 epoch trained. The adaptive momentum estimation (ADAM) is selected as the optimization method in this paper, and its setting parameters are: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$.

B. EVALUATION METRICS

Generally, according to the audience perception effect and visual attributes, the model performance evaluation can be divided into subjective evaluation and objective evaluation. In this paper, these two ways are both used to evaluate the effects of reconstructed SR images.

For the subjective evaluation, it is based on the natural and clear perception effect of human eyes on the SR image, which can directly reflect the visual effect of the reconstructed SR image. However, due to the strong subjectivity of the subjective evaluation, it is not scientific and accurate enough, so it is necessary to adopt objective evaluation indicators to make a more comprehensive judgment on the quality of image reconstruction.

For the objective evaluation, the most widely used image evaluation metric in the field of image super-resolution reconstruction is peak signal-to-noise ratio (PSNR) (unit: dB), and

TABLE 1. Effects of different feature extraction modules and RSAM on model performance.

Model	RSAM	Params	PSNR(dB)	SSIM
MSRB	×	4.69M	32.36	0.9304
	√	4.69M	32.39	0.9305
VLDB	×	5.17M	32.36	0.9301
	√	5.17M	32.43	0.9308
MFRB	×	4.89M	32.41	0.9310
	√	4.89M	32.46	0.9314
LMRM	×	**4.53M**	32.55	0.9322
	√	**4.53M**	**32.57**	**0.9325**

the formula for calculating PSNR can be given by

$$PSNR = 10\log_{10}(\frac{MAX^2}{MSE}) \tag{17}$$

where MAX represents the maximum value of the pixel range in the image, which is 255 in this paper. MSE denotes the mean square error between the real image and reconstructed image. The higher the PSNR measured between two images, the closer the reconstructed image is to the real image.

In this paper, to ensure the comprehensiveness of model evaluation results, the structural similarity (SSIM) are also used as performance evaluation metric. For the SSIM, it measures the similarity of images comprehensively from the three dimensions of brightness, contrast and structure, and the formula for calculating SSIM can be given by

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{18}$$

where x and y correspond to SR image and original HR image respectively. μ_x and μ_y represent gray average values of the two images respectively. σ_{xy} is the covariance of x and y, σ_x and σ_y denote the variances of x and y respectively. c_1 and c_2 are constants. The closer the SSIM value is to 1, the more similar the structure of the SR image and the original HR image is, and the better the model reconstruction effect will be.

C. ABLATION EXPERIMENT

To verify the improvement of the model reconstruction effect of LMRM and RSAM proposed in this paper. Four kinds of network blocks, MSRB [22], VLDB [23], MFRB [24] and LMRM, are connected and unconnected to the RSAM respectively. The connection modes between modules all adopted the dense feature fusion structure. The objective evaluation metrics PSNR, SSIM and params are used for evaluation, and Table 1 shows the 2x reconstruction results of each model on Urban100 data set with more details. The best value is bolded, and the second-best value is underlined.

As can be seen from Table 1, adding RSAM to the network with the same feature extraction module can not only effectively improve the reconstruction effect, but also hardly increase the number of parameters in the network model. Herein, the reason why we say they barely increase the number of parameters while they have the same number of parameters is that we round the argument to the third decimal place, and we should always keep in mind that the parameter is measured in megabits, a relatively small weight and rank unit, So its third decimal place is of little real significance as an indicator of the number of parameters in a model.

For example, after adding RSAM to a VLDB network, its PSNR and SSIM increase by 0.07dB and 0.0007 dB respectively, while its corresponding parameter number hardly increases. Similarly, after adding RSAM to the MSRB, MFRB and LMRM networks, the number of the three model parameters did not increase significantly, while their PSNR increased by 0.03dB, 0.05dB and 0.02dB, respectively, and SSIM increased by 0.0003 on average. This proves that the RSAM can effectively improve the model reconstruction effect and the number of parameters is almost constant.

For the same connection mode, the reconstruction effect of the proposed LMRM is obviously better than that of MSRB VLDB and MFRB. For example, when the RSAM module is added to all four networks, compared with MSRB, VLDB and MFRB, the PSNR of the proposed LMRM proposed is increased by 0.18dB, 0.14dB and 0.11dB, respectively, and the SSIM is increased by 0.0016 on average. Furthermore, the number of parameters decreased by 0.16M, 0.64M and 0.36M, respectively, which proves that our LMRM can effectively utilize image feature information and improve model reconstruction performance.

All in all, through the above experiments, we can draw the following conclusions: both the LMRM and RSAM modules proposed in this paper can effectively improve the model reconstruction effect.

D. EXPERIMENTAL RESULTS AND DISCUSSION

In this subsection, to further verify the validity of our RAMF network, 11 advanced image SR networks (i.e., MSRN [22], DID-D5 [23], MDFN [24], Bicubic [25], SRCNN [26], VDSR [27], DRCN [28], LapSRN [29], IMDN [30], OISR-SK2 [31] and LatticeNet [32]) are tested on the four benchmark datasets and compared with the proposed algorithm in terms of objective evaluation indicators and subjective visual effects. All the networks are tested and compared under the three scaling factors of ×2, ×3 and ×4 to fully verify the effect of the proposed RAMF network.

1) OBJECTIVE EVALUATION

The results of different models under objective evaluation metrics PSNR and SSIM are shown in Table 2.

As can be seen from Table 2, compared with the Bicubic, SRCNN, VDSR, DRCN and LapSRN networks on the 4 benchmark datasets with the ×2 scaling factor, the PSNR of the RAMF proposed in this paper increases by 4.14dB, 1.74dB, 0.92dB, 0.9dB and 1.02dB on average, respectively. Besides, compared with the multi-scale network models

TABLE 2. Comparison of index performance of each network under the four benchmark datasets when scaling factor is 2, 3 and 4, respectively.

Method	Scale	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic [25]	×2	33.66	0.9299	30.24	0.8687	29.56	0.8431	26.88	0.8401
SRCNN [26]	×2	36.66	0.9542	32.43	0.9063	31.36	0.8879	29.50	0.8946
VDSR [27]	×2	37.54	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140
DRCN [28]	×2	37.63	0.9584	33.06	0.9108	31.85	0.8947	30.76	0.9147
LapSRN [29]	×2	37.53	0.9591	33.08	0.9109	31.80	0.8949	30.41	0.9112
MSRN [22]	×2	38.08	0.9605	33.74	0.9170	32.23	**0.9013**	32.22	**0.9326**
IMDN [30]	×2	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283
OISR-SK2 [31]	×2	38.12	0.9609	33.80	0.9193	32.26	0.9006	32.48	0.9317
LatticeNet [32]	×2	38.15	**0.9610**	33.78	0.9193	32.25	0.9005	32.43	0.9302
DID-D5 [23]	×2	38.15	**0.9610**	33.77	0.9190	32.27	0.9006	32.38	0.9305
MDFN [24]	×2	38.14	**0.9610**	33.83	0.9196	32.27	0.9006	32.41	0.9310
RAMF (our)	×2	**38.17**	**0.9610**	**33.87**	**0.9198**	**32.29**	0.9009	**32.57**	0.9325
Bicubic	×3	30.39	0.8682	27.54	0.7736	27.21	0.7384	24.46	0.7344
SRCNN	×3	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989
VDSR	×3	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279
DRCN	×3	33.85	0.9215	29.89	0.8317	28.81	0.7954	27.16	0.8311
LapSRN	×3	33.82	0.9227	29.89	0.8320	28.83	0.7973	27.08	0.8272
MSRN	×3	34.38	0.9262	30.34	0.8395	29.08	0.8041	28.08	0.8554
IMDN	×3	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519
OISR-SK2	×3	34.55	0.9282	30.46	0.8443	29.18	0.8075	28.50	0.8597
LatticeNet	×3	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538
DID-D5	×3	34.55	0.9280	30.49	0.8446	29.19	0.8069	28.39	0.8566
MDFN	×3	34.60	0.9284	30.50	0.8449	29.21	0.8075	28.52	0.8591
RAMF (our)	×3	**34.68**	**0.9288**	**30.57**	**0.8461**	**29.24**	**0.8084**	**28.65**	**0.8619**
Bicubic	×4	28.42	0.8104	26.00	0.7019	25.96	0.6674	23.14	0.6570
SRCNN	×4	30.48	0.8628	27.49	0.7503	26.90	0.7101	24.53	0.7221
VDSR	×4	31.35	0.8830	28.01	0.7680	27.29	0.7251	25.18	0.7543
DRCN	×4	31.56	0.8810	28.15	0.7620	27.24	0.7150	25.15	0.7530
LapSRN	×4	31.54	0.8855	28.19	0.7720	27.32	0.7280	25.21	0.7553
MSRN	×4	32.07	0.8903	28.60	0.7751	27.52	0.7273	26.04	0.7896
IMDN	×4	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838
OISR-SK2	×4	32.32	0.8965	28.72	0.7843	27.66	0.7390	26.37	0.7953
LatticeNet	×4	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873
DID-D5	×4	32.33	0.8968	28.75	0.7852	27.68	0.7386	26.36	0.7933
MDFN	×4	**32.41**	0.8976	28.78	0.7860	27.69	0.7393	26.39	0.7944
RAMF (our)	×4	**32.41**	**0.8979**	**28.80**	**0.7865**	**27.71**	**0.7401**	**26.49**	**0.7978**

MSRN, DID-D5 and MDFN of the same type with the ×4 scaling factor, the PSNR obtained by our RAMF on the four benchmark datasets is also improved by 0.30dB, 0.07dB and 0.04dB on average, respectively.

In addition, it is worth noting that the RAMF proposed in this paper also has a breakthrough in the evaluation index of SSIM which is difficult to improve. For example, compared with the Bicubic, SRCNN, VDSR, DRCN and LapSRN networks on the 4 benchmark datasets with the ×3 scaling factor, the SSIM obtained by our RAMF is also improved by 0.0827, 0.0324, 0.0168, 0.0164 and 0.0165 on average, respectively. Furthermore, compared with the advanced

networks, IMDN, OISR-SK2 and LatticeNet in recent years on the 4 benchmark datasets with the $\times 2$ scaling factor, the SSIM obtained by the proposed RAMF is improved by 0.0020, 0.0004 and 0.0008 on average, respectively. Moreover, compared with the multi-scale network models, MSRN, DID-D5 and MDFN, of the same type with the $\times 4$ scaling factor, the SSIM obtained by our RAMF on the four benchmark datasets can also be improved by 0.0088, 0.0021, 0.0012 on average, respectively.

According to the above results analysis, we can conclude that the image reconstruction effect of RAMF proposed in this paper are obviously superior to the other 11 networks under the three scaling factors. Especially, compared with the DID-D5 network published in ICPR 2021 and MDFN network published in 2022, the PSNR on the Urban100 test set with rich image texture information can be improved by 0.19dB and 0.13dB on average, and SSIM increased by 0.0039 and 0.0027 on average.

2) SUBJECTIVE EVALUATION

In terms of subjective visual effects, to more intuitively observe and discover the superiority of the proposed RAMF, partial reconstruction results of different networks under the Urban100 data set with rich image details are shown in Figure 4-6. Since the larger the scaling factor, the higher the requirement on network model performance, and the more difficult it is to reconstruct. Therefore, to fully prove the effectiveness of the proposed RAMF network, the reconstruction results of different networks with the $\times 4$ scaling factor are selected in this paper for display.

As can be seen from Fig. 4, the clarity of glass window images reconstructed by different networks is quite different. For example, the reconstructed images by the Bicubic and SRCNN are very fuzzy with unclear lines and poor visual effects. Although the reconstruction effect has been improved to a certain extent by the IMDN and DID-D5 networks, and the line images are basically clear, the vertical lines are still overlapped and the edges are blurred. For the MSRN and MDFN, the line outline of the image reconstructed by the two networks is clear, but the lines at the intersection are blurred and have a ringing effect. However, compared with the other networks, RAMF has the clearest image texture details, and basically does not have ringing phenomenon, achieving the best reconstruction effect. Similarly, almost the same reconstruction effect can also be observed in Fig. 5.

Seen from Fig. 6, we can find that except for RAMF, the zebra crossing images reconstructed by the other six methods all have the problem of line direction confusion. Especially, the image reconstructed by Bicubic, SRCNN, IMDN and DID-D5 could hardly see the lines in the right direction, resulting in serious blurring. What's more, although some lines in the correct direction can be seen in the image reconstructed by MSRN and MDFN, serious distortion still exists and ringing effect is obvious. Compared with the other 6 networks, the RAMF network proposed in this paper has the best reconstruction effect, especially to solve the problem of line orientation confusion, and the reconstructed image texture details are very rich. To sum up, the above experimental results demonstrate that the proposed RAMF using RSAM and LMRM can achieve outstanding reconstruction effect both in terms of objective evaluation indicators and subjective visual effects.

TABLE 3. Comparison of complexity and performance indicators of different models.

Method tested	Params/M	Flops/G	PSNR/dB	SSIM
MSRN	6.08	107.27	26.04	0.7896
OISR-SK2	5.51	117.41	26.37	0.7953
DID-D5	5.21	93.03	26.36	0.7933
MDFN	4.89	87.76	26.39	0.7944
RAMF	**4.53**	**81.95**	**26.49**	**0.7978**

E. PARAMETER ANALYSIS

To set the parameters of the proposed RAMF network reasonably, this subsection further analyzes the effects of the number Q of LMRM in the RAMF and the number of convolutional layer output channels N in the LMRM on the model performance, in which the N of LMRM (a) is 64 and that of LMRM (b) is 128. Two sets of experiments are designed to test the influence of Q and N on the model performance, and the test results are shown in Fig. 7. In addition, the NH in the Fig. 7 indicates the cross connection of LMRM (a) and LMRM (b).

As can be seen from Fig. 7(a), when the number of convolutional layer output channels N in LMRM remains unchanged, the more LMRM, the more model parameters, the better model performance, indicating that increasing the number of LMRM can improve the model reconstruction effect, but also lead to an increase in the number of model parameters. It can be observed from Figure 7 (b) that compared with the model only used LMRM (a) to extract image features, the PNSR of the model used LMRM (a) and LMRM (b) cross-connected is higher and much more stable. What's more, compared with the model only used LMRM (b) to extract image features, the model used LMRM (a) and LMRM (b) cross-connected have fewer parameters and almost no reduction in reconstruction performance.

According to the experiments in this section, increasing the number of LMRM in the RAMF and the number of convolutional layer output channels in the LMRM can improve the model performance to some extent, while the number of model parameters will also increase, leading to the reconstruction efficiency will also be affected.

Therefore, to balance the number of model parameters and reconstruction effect, and obtain higher reconstruction efficiency, the number of LMRM in RAMF is finally set to 16, and LMRM (a) and LMRM (b), whose output channels of the convolutional layer are set to 64 and 128 respectively, are cross-connected.

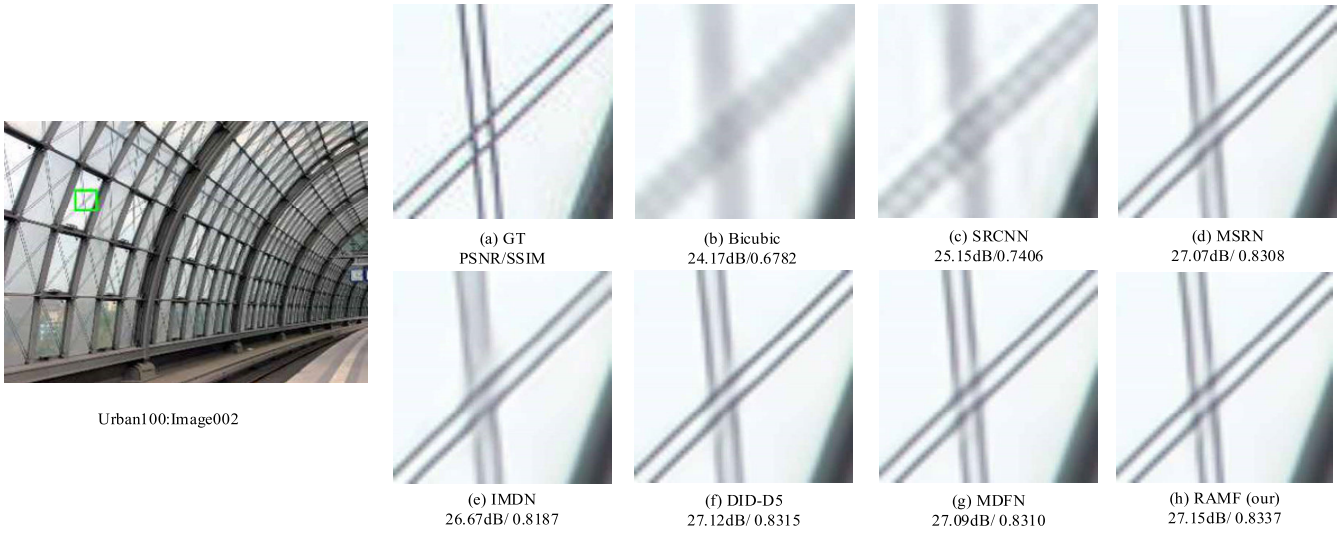

FIGURE 4. Reconstruction results of Image002 in Urban100.

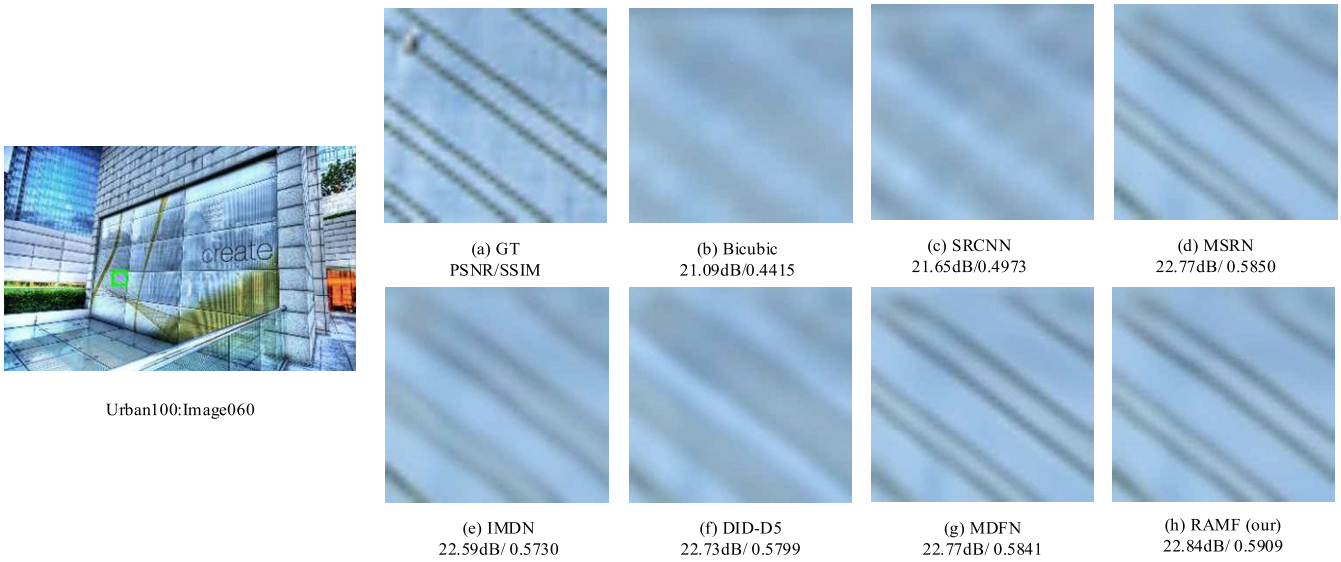

FIGURE 5. Reconstruction results of Image060 in Urban100.

F. COMPLEXITY ANALYSIS

Generally, the complexity of the network will affect the computing speed and reconstruction performance of the model. Although the reconstruction performance can be improved to a certain extent by deepening the depth of the network, the complexity of the model and the training time will also increase, resulting in the decline of the reconstruction efficiency. To balance the number of model parameters and the reconstruction effect, the lightweight multi-scale residual module LMRM and residual spatial attention module RSAM are proposed in this paper. Then, the RAMF network is constructed, which can improve the model performance while reducing the number of model parameters.

To verify the effectiveness of the proposed RAMF, the four networks, DID-D5, OISR-SK2, MSRN and MDFN, with better performance among the 11 comparison algorithms in this paper are selected for comparison experiment in terms of model complexity and reconstruction performance respectively. The Urban100 ($\times 4$) is used for test dataset, and Table 3 shows the complexity information of different networks, as well as the reconstruction results in PSNR and SSIM. The best value is bolded, and the second-best value is underlined. The best results are shown in bold and the sub-optimal values are underlined.

As can be seen from Table 3, compared with the MSRN network with larger parameters, the PSNR and SSIM of the proposed RAMF are significantly improved by 0.45dB and 0.0082, respectively. Furthermore, its parameter number is reduced by 25.49%, while Flops is also significantly decreased. Compared with the DID-D5 and OISR-SK2 networks, the PSNR and SSIM of the proposed RAMF network not only increase by 0.13dB and 0.0035 on average, but

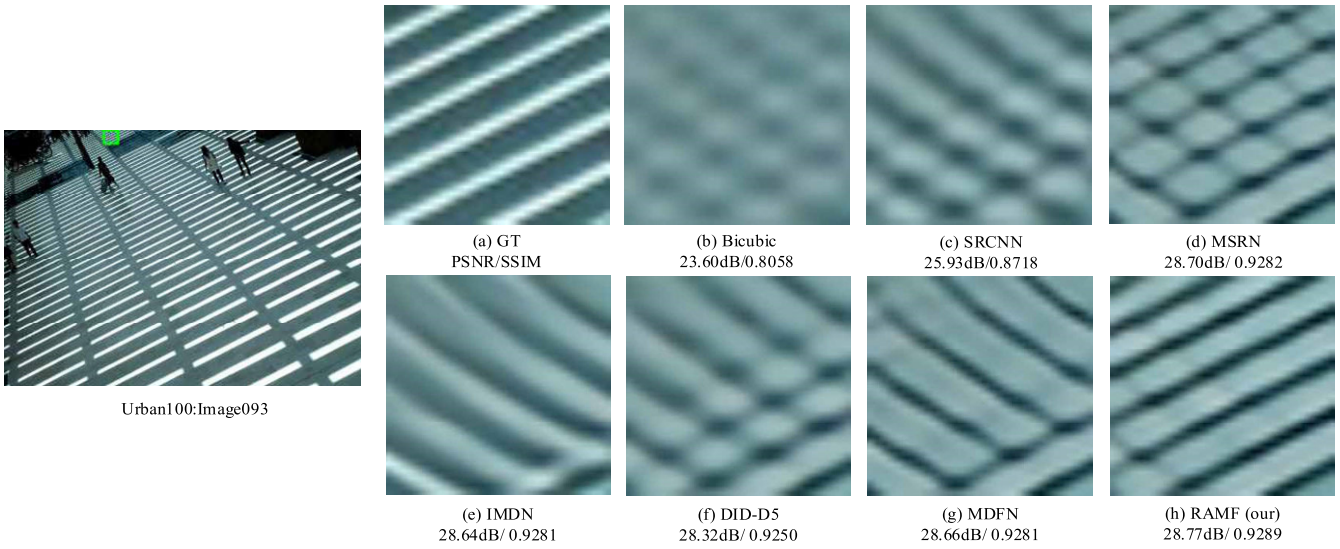

FIGURE 6. Reconstruction results of Image093 in Urban100.

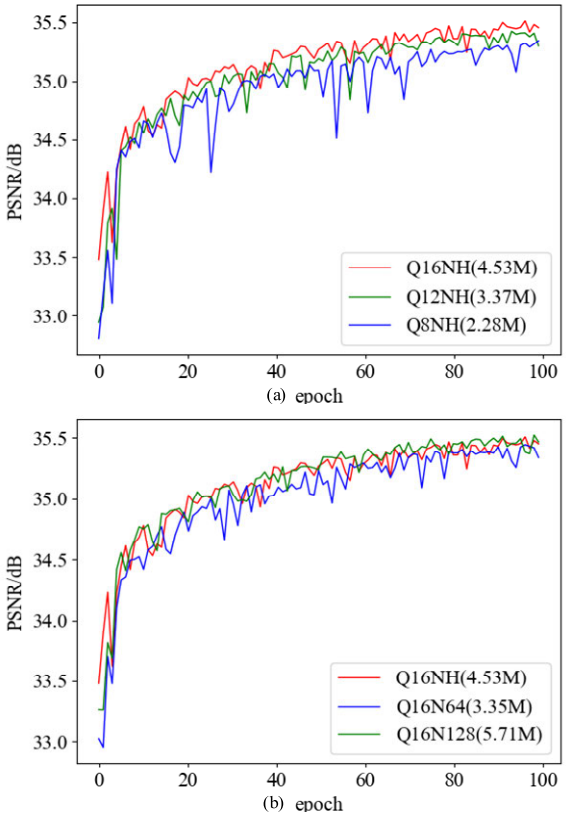

FIGURE 7. PSNR comparison of the RAMF with different Q and N.

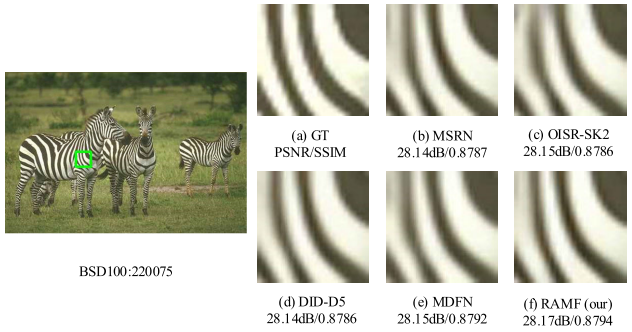

FIGURE 8. Reconstruction effect of 220075 in BSD100.

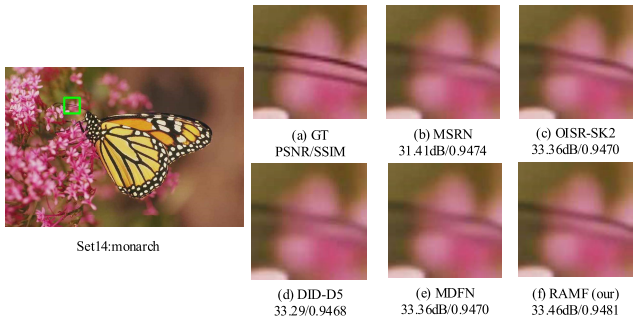

FIGURE 9. Reconstruction effects of monarch in Set14.

also the number of parameters and Flops decreased significantly. For the MDFN network, its number of parameters and the amount of computation are 7.9% and 7.1% higher than the proposed RAMF, but its PSNR and SSIM are 0.1dB and 0.0034 lower than our RAMF, respectively. The above experiments fully prove that the proposed RAMF can not only achieve good reconstruction effect, but also has fewer parameters, low computational complexity, fast processing speed and high objective evaluation index.

To observe the specific reconstruction effects of the five networks more directly in subjective visual evaluation, Fig. 8 and Fig. 9 show their $\times 4$ reconstruction results on 220075 in test set BSD100 and monarch in test set Set14, respectively.

As can be seen in Fig. 8, OISR-SK2 and MDFN show ringing effect at the black stripes on the zebra body, and the

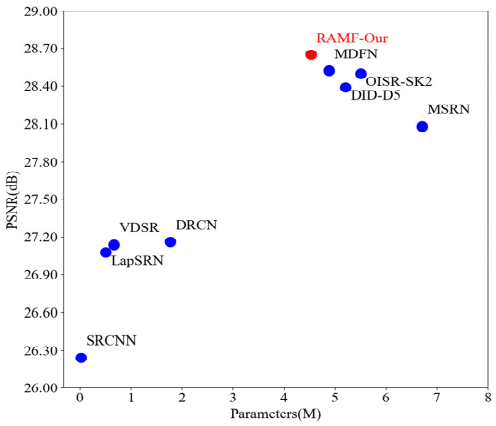

FIGURE 10. Comparison of PSNR and parameter number of different models on Urban100 (×3).

lines are obviously blurred, making it impossible to see the outline of the lines. Besides, the zebra stripes reconstructed by MSRN and DID-D5 are relatively clear, but the fringe of the stripes is still blurred. Compared with the other four algorithms, the image fringes reconstructed by the proposed RAMF network are the clearest, and it can effectively reduce the edge blur phenomenon. In Fig. 9, compared with MSRN, OISR-SK2, DID-D5 and MDFN, the butterfly image reconstructed by our RAMF has a clearer outline at the antennae and is more clearly distinguishable from the pink background, which indicates that our network can reconstruct a more realistic image. The experimental results of the above subjective and objective indicators show that our RAMF can not only effectively improve the image reconstruction effect, but also reduce the number of model parameters, improve the utilization rate of feature information, and realize more efficient image super-resolution reconstruction.

To further prove the superiority of RAMF proposed in this paper, Fig. 10 shows the comparison of the PSNR and parameter number obtained by different networks on Urban100 (×3). As can be seen from Fig. 10, although the parameter number of our RAMF is higher than that of SRCNN, VDSR, DRCN and LapSRN, the PSNR of the reconstructed image is also much higher than them. Compared with MSRN, DID-D5, OISR-SK2 and MDFN, our RAMF achieves the highest PSNR while the number of parameters is less than theirs, which fully demonstrates the superiority of the proposed network.

IV. CONCLUSION

In this paper, an image super-resolution reconstruction network based on residual attention and multi-scale feature fusion is proposed. By constructing 16 lightweight multi-scale residual modules, abundant image feature information of different sensitivity fields with fewer parameters are obtained. Then, each LMRM is connected by a dense feature fusion structure, which reduces feature loss and improves information utilization. Finally, a residual spatial attention module is developed and used to specifically learn high-frequency feature information and reasonably allocate computing resources. Comprehensively experimental results on four databases demonstrate that the proposed network can achieve remarkable reconstruction effect, high objective evaluation index and fast processing speed while enjoying a fewer network parameters and lower computational complexity. In future work, channel attention and its combination with spatial attention used in this paper will be explored to further improve the effect of image super-resolution reconstruction.

REFERENCES

[1] D. Cheng, L. Chen, C. Lv, L. Guo, and Q. Kou, "Light-guided and cross-fusion U-Net for anti-illumination image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8436–8449, Dec. 2022.
[2] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 4922–4936, 2022.
[3] Q. Kou, D. Cheng, H. Zhuang, and R. Gao, "Cross-complementary local binary pattern for robust texture classification," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 129–133, Jan. 2019.
[4] J. L. Harris, "Diffraction and resolving power," *J. Opt. Soc. Amer.*, vol. 54, no. 7, pp. 931–933, 1964.
[5] R. Y. Tsai and T. S. Huang, "Multi-frame image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, no. 2, pp. 317–339, 1984.
[6] X. Kong, X. Liu, J. Gu, Y. Qiao, and C. Dong, "Reflash dropout in image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5992–6002.
[7] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 288–297.
[8] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5689–5698.
[9] L. Chen, L. Guo, D. Cheng, and Q. Kou, "Structure-preserving and color-restoring up-sampling for single low-light image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1889–1902, Apr. 2022.
[10] S. Wang and J. Zheng, "Multi-scale detail enhancement network for image super-resolution," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 161–167.
[11] B. Zhang and C. Gao, "Deep residual network for image super-resolution reconstruction," in *Proc. 12th Int. Conf. CYBER Technol. Autom., Control, Intell. Syst. (CYBER)*, Jul. 2022, pp. 620–623.
[12] T. Lu, Y. Wang, J. Wang, W. Liu, and Y. Zhang, "Single image super-resolution via multi-scale information polymerization network," *IEEE Signal Process. Lett.*, vol. 28, pp. 1305–1309, 2021.
[13] A. Esmaeilzehi, M. O. Ahmad, and M. N. S. Swamy, "PHMNet: A deep super resolution network using parallel and hierarchical multi-scale residual blocks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.
[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2018, pp. 286–301.
[15] K. Xu, J. L. Ba, and R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jul. 2015, pp. 2048–2057.
[16] S. Woo, "CBAM: Convolutional block attention Modul," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.
[17] Z. Lu and C. Liu, "Multiscale feature reuse mixed attention network for image reconstruction," *J. Image Graph.*, vol. 26, no. 11, pp. 2645–2658, 2021.
[18] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2012, p. 135.
[19] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. 7th Int. Cof. Curves Surf.*, Avignon, France, Jun. 2010, pp. 711–730.

[20] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vision. (ICCV)*, Jul. 2001, pp. 416–423.
[21] J. 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.
[22] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 527–542.
[23] L. Li, H. Feng, B. Zheng, L. Ma, and J. Tian, "DID: A nested dense in dense structure with variable local dense blocks for super-resolution image reconstruction," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2582–2589.
[24] D. Cheng, "Multi-scale dense feature fusion network for image super-resolution," *Opt. Precis. Eng.*, vol. 30, no. 20, pp. 2489–2500, 2022.
[25] 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.
[26] 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.
[27] 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.
[28] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1637–1645.
[29] W. Lai, J. Huang, N. Ahuja, and M. 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.
[30] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.
[31] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "ODE-inspired network design for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1732–1741.
[32] X. Luo, "LatticeNet: Towards lightweight image super-resolution with lattice block," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 272–289.

QIQI KOU received the B.S. and M.S. degrees from the Anhui University of Science and Technology, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Information and Control Engineering, China University of Mining and Technology, in 2019. He is currently a Lecturer with the School of Computer Science and Technology, China University of Mining and Technology. His research interests include image processing and pattern recognition.

JIAMIN ZHAO received the B.S. degree from the School of Information and Electrical Engineering, Xuzhou University of Technology, in 2019. She is currently pursuing the M.S. degree with the School of Information and Control Engineering, China University of Mining and Technology. Her research interests include image super-resolution, computer vision, and pattern recognition.

DEQIANG CHENG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and information engineering from the China University of Mining and Technology.

He led the establishment of the Intelligent Detection and Pattern Recognition Research Center, China University of Mining and Technology, and the Intelligent Mine Research Center, Artificial Intelligence Research Institute, China University of Mining and Technology. He is currently a Professor with the China University of Mining and Technology. He is also the National Coal Industry Education Advanced Worker, the Jiangsu Province, the National Coal Youth May Fourth Medal Winner, "333 High-Level Talent Training Project" Young and Middle-Aged Academic Technology Leader, the Province "Six Talent Peak" High-Level Talent Selection and Training Object, and the Jiangsu Province Excellent Educator. He is also the Leader of the Excellent Teaching Team of "Qinglan Project" in Jiangsu colleges and universities and the Leader of the Provincial Excellent Grassroots Teaching Organization in Jiangsu colleges and universities. He has published more than 100 scientific research articles in well-known journals at home and abroad. His research interests include machine learning, video coding, image processing, and pattern recognition.

ZHEN SU received the B.S. degree from the School of Information Science and Technology, Qingdao University of Technology, in 2005, and the M.S. degree from the School of Information Science and Engineering, Ocean University of China, in 2008. He is currently pursuing the Ph.D. degree with the School of Information and Control Engineering, China University of Mining and Technology. His research interests include super-resolution reconstruction, image enhancement, and machine learning.

XINGGUANG ZHU received the B.S. degree from the School of Electronic Information, Jiangsu University of Science and Technology, in 2020. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, China University of Mining and Technology. His research interests include machine learning and deep learning-based vision tasks, such as image restoration, super-resolution, and saliency detection.