IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 20, 2022, accepted May 17, 2022, date of publication May 20, 2022, date of current version May 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176605

MESR: Multistage Enhancement Network for Image Super-Resolution

DETIAN HUANG AND JIAN CHEN
College of Engineering, Huaqiao University, Quanzhou 362021, China

Corresponding author: Detian Huang (huangdetian@hqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61901183 and Grant 61976098, in part by the Fundamental Research Funds for the Central Universities under Grant ZQN-921, in part by the Natural Science Foundation of Fujian Province under Grant 2019J01010561, in part by the Foundation of Fujian Education Department under Grant JAT170053, and in part by the Science and Technology Bureau of Quanzhou under Grant 2017G046.

ABSTRACT Recently, the deep-learning-based image super-resolution methods have achieved astounding advancement. Whereas most of these methods utilize features from the low-resolution image space exclusively, and ignore the dependency between contextual features simultaneously, resulting in their limited ability to restore details. To this end, a multi-stage enhancement image network for super-resolution (MESR) is proposed. The network consists of two stages, where the first stage is used to generate a coarse reconstructed image, and the second one is to refine the coarse image, which enhances the super-resolution performance. Specifically, in the first stage, to acquire more abundant features, an effective funnel-like multi-scale feature extractor is proposed, incorporating a channel attention mechanism to boost the feature representation capability. Moreover, an adaptive weighted residual feature fusion block is designed to effectively explore and exploit the dependency between contextual features for generating more beneficial features. In the second stage, a refinement block is proposed to additionally strengthen the details of the reconstructed image by exploring the feature information from the high-resolution image space. Experimental results demonstrate that the proposed method achieves superior performance against the state-of-the-art SR methods in terms of both subjective visual quality and objective quantitative metrics.

INDEX TERMS Image super-resolution, multi-stage network, multi-scale feature, feature fusion, image refinement.

I. INTRODUCTION

In recent years, image super-resolution (SR), one of the pivotal techniques in computer vision, is designed to reconstruct a high-resolution (HR) image with affluent details from one or more existing low-resolution (LR) images [1], [2], and it has a wide range of applications in diverse fields, such as medical imaging [3], video surveillance [4], remote sensing images [5], *etc*. However, image SR is essentially an ill-posed inverse problem due to the fact that multiple HR solutions may correspond to the same LR input [6].

Although deep learning has accomplished excellent outcomes in SR tasks, there still exist several issues as follows. (1) Most deep-learning-based SR methods utilize a single-scale convolution kernel to construct wider or deeper feature extraction module, and ignores the potential correlation between multi-scale features, as in [7]–[11], resulting in the obtained features are still relatively homogeneous at the scale level, and not well adapted to SR tasks of various scales [12]. (2) As the network depth continues to increase, some features may gradually disappear during transmission [13]–[15], as well as the training difficulty increases. Although ResNet effectively alleviates the training difficulty and mitigates the issue of feature disappearance [16], it also ignores the validity of features from different levels and the correlation between contextual features [17]–[19], which limits the super-resolution performance to some extent [20]. For most deep-learning-based SR methods, there remains a challenge in reconstructing high-frequency details. On the one hand, the pre-upsampling SR methods can directly extract features from the HR image space, while it may lead to unstable training and its reconstructed images contain artificial artifacts due to the introduction of redundant noise [21]. On the other hand, the post-upsampling SR methods alleviate

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

the problem of artificial artifacts in the reconstructed images caused by the pre-upsampling methods [22]. Nevertheless, a majority of these methods only retrieve the feature information from the LR image space and fail to further exploit the high-frequency features of the reconstructed images obtained by the upsampling module, resulting in the significant image details being ignored [23], [24].

Numerous studies [21], [25]–[28] on image SR have revealed that the potential HR image features cannot be absolutely described by using feature information from the LR image space during super-resolution solely, whereas the feature information of a coarse reconstructed image can facilitate an LR image to reconstruct a better reconstructed image. These methods [21], [25]–[28] employed a coarse-to-fine mechanism to enhance the high-frequency details of the final reconstructed image incrementally through exploiting features of the coarse reconstructed image. Additionally, multi-stage structure networks for image restoration have received increasing attention since they are able to generate enriched and refined high-frequency details, such as [8], [10], [21], [24], [29]–[31]. Such methods simplify the information flows among different stages as well as enhance the communication and connection between the features of different stages by learning and exchanging information in different stages, leading to a significant improvement in the resilience of high-frequency details. Inspired by the literature mentioned above, a Multi-Stage Enhancement Image Network for Super-Resolution (MESR) is proposed, which adopts a multi-stage structure network to fuse the information from the LR and HR image spaces to accomplish coarse-to-fine super-resolution on the input LR image. Moreover, we design a Refinement Block (RB) to rationally exploit the high-frequency features of the coarse reconstructed image. Specifically, we capture the complementary SR features by exploring the available cue information in the HR image space to minimize the inconsistency between the generated reconstructed image and the original HR image, which compensates for the abridgment of feature information in the LR image space.

Since for different viewing angles, and aspect ratios, the same or similar objects exhibit different features in an image at different scales [32]–[35], exploring multi-scale feature is an attractive option to boost the super-resolution performance [32]. Meanwhile, related studies in [7], [36]–[39] have revealed that the attention mechanism that explores the interdependence between features is an effectual access to advance the feature representation ability. Furthermore, for the deeper networks, as a way to boost its performance, increasing the network width is more powerful than increasing the network depth [29], [40], [41]. Consequently, a multi-scale feature extraction block (MSFEB) is proposed to agglutinate multi-level information at different perceptual fields by expanding the network width and incorporating the channel attention mechanism, which enables sharing and exchanging of features at divers scales between different levels, and strengthens the inherent complementarity and connection of features in local regions effectively, so as to better generate deep features with high-frequency details.

As SR task is a low-level vision task, the feature information among input and output images and the intermediate layers of the network ought to be profoundly correlated, which implies that it is an essential way to utilize the feature information from different layers. As far as local features, which are separated from various levels of the network provide differential representative of high-frequency details. Nonetheless, most deep-learning-based SR methods just process the features of the previous module and rarely focus on the features of adjacent or previous modules, resulting in unfortunate dependency between contextual features. As for the global features, the shallow features contain more edge and texture information, and the deep features contain more contextual feature information [42]. Considering that the powerful combination of local and global features provides dramatic improvement in super-resolution performance, we apply the residual learning mechanism [16] for local and global features fusion and proposed an adaptive weighted Residual Feature Fusion Block (RFFB), which not only successfully alleviates the problem of feature disappearance during transfer [11], [17]–[19], but also facilitates the fluent transfer of contextual residual features.

In general, the proposed MESR consists of two stages. Specifically, in the first stage, features learning is performed on the LR image to obtain the feature mapping of the corresponding HR image, and then features of different scales are merged to strengthen the dependency between contextual features, and subsequently, the coarse reconstructed image is generated by the upsampling module. In the second stage, the feature information learned from the HR image space is used to refine the rough reconstructed image, so as to further produce a refined reconstructed image with more high-frequency details.

The primary contributions of this paper are as follows.

- A multi-stage enhancement network for super-resolution (MESR) is proposed to reconstruct as many high-frequency details as possible by two stages from coarse to fine. Extensive experimental results indicate that the proposed MESR outperforms the state-of-the-art methods in terms of both visual effects and quantitative metrics.
- A multi-scale feature extraction block (MSFEB) with channel attention is designed to acquire the feature information by varying perceptual fields effectively.
- An adaptive weighted residual feature fusion block (RFFB) is designed to fuse all the hierarchical features from the LR image space effectively, and the shallow and deep features are fused to attain more beneficial global feature information with a global residual learning mechanism.
- A refinement block (RB) is proposed to minimize the inconsistency between the generated reconstructed image and the original HR image by exploring the high-frequency features from the HR image

space, so as to greatly promote the detail restoration ability.

The organization of this paper is as follows. In Section II, we summarize the related work. And then, the details of the proposed MESR are presented in Section III. In Section IV, experimental results are analyzed and compared with the recently state-of-the-art SR methods. Finally, we conclude the paper in Section V.

II. RELATED WORKS

The SR methods are broadly classified into early traditional SR methods [43]–[46], the sparse-learning-based SR methods [47]–[52], and the deep-learning-based SR methods. In this section, we will principally present the CNN-based SR methods related to this paper, which can be mainly divided into the pre-upsampling SR methods, the post-upsampling SR methods, and the coarse-to-fine SR methods.

A. PRE-UPSAMPLING SR METHODS

In the earlier days, deep-learning-based SR methods mostly employed pre-upsampling operation, such as SRCNN [25], VDSR [13], DRCN [53], DRRN [18], MemNet [54], *etc*. These methods first perform the upsampling operation to enlarge the LR image to an intermediate image with the desired resolution by the bicubic interpolation [44], and then feed the obtained intermediate image into the network. Since the network only needs to refine the coarse intermediate image directly by the pre-defined conventional method, which fundamentally lessens the learning difficulty. However, the pre-upsampling method not only provokes some side effects (*e.g.*, amplifies noise, introduces artificial feature), but also appoint incremental costs in time and space due to learning high-dimensional features [22], [55].

B. POST-UPSAMPLING SR METHODS

To reduce the computational complexity while improving super-resolution performance, related methods, such as FSRCNN [22], IDN [56], EDSR [15], CARN [17], RCAN [37], SAN [38], PAN [7], *etc.*, replace the pre-upsampling operation with end-to-end learnable post-upsampling module. For the post-upsampling SR methods, the time and space costs are greatly reduced because the non-linear feature mapping occurs only in the low-dimensional features space and the upsampling module is located at the very end of the network. Therefore, the post-upsampling SR method has become one of the most dominant frameworks in the super-resolution field. Notwithstanding, since these methods center around separating features in the LR image space, deeper or more extensive complex networks are frequently expected to acquire adequate information for supervising fine HR images. In addition, it is difficult to train large scaling factor with a separate upsampling module at the end of the network.

C. COARSE-TO-FINE SR METHODS

Considering that the feature information from the LR image space is limited, and exploring the features of the coarsely reconstructed image can facilitate an LR image to reconstruct a better reconstructed image, the coarse-to-fine SR method has received more and more attention. To achieve more noteworthy refinement of the reconstructed images, Lai *et al.* [35] proposed a deep laplacian pyramid network for super-resolution (LapSRN). For each pyramid level, LapSRN incorporates the feature maps of the coarse reconstructed images for progressive prediction the high-frequency details of the final reconstructed images. To ease the training difficulty of the large scaling factor, Wang *et al.* [57] exploited a fully progressive approach for super-resolution (ProSR), which utilizes a progressive upsampling method to refine the coarse reconstructed images by layer-by-layer and scale-by-scale feature extraction.

In view of the limited feature information in the LR image space and abundant complementary feature information contained in the HR image space, Haris *et al.* [26] proposed a deep back-projection network (DBPN), which not only utilizes the upsampling module to generate the coarse reconstructed images, but also uses the downsampling module to map them to the LR image space, enabling the network to continuously refine the details of the final reconstructed images with the HR feature maps. And then, Wen *et al.* [28] proposed a unique learning mechanism by CNN to learn larger filter kernels firstly to generate the coarse reconstructed images, and then learn smaller filter kernels to generate the finer reconstructed images. To further exploit the potential of the coarse reconstructed image, Li *et al.* [27] proposed a super-resolution based on feedback network (SRFBN), which takes the coarse reconstructed image generated by the feedback module together with the original LR image as its new input and re-inputs it into the feedback module, thus leveraging the high-frequency features of the coarse reconstructed image to compensate for the insufficient spatial feature information of the LR image. With several feedback rounds, SRFBN efficiently strengthens the high-frequency details of the reconstructed images. After that, Tian *et al.* [21] proposed a coarse-to-fine convolutional neural network for super-resolution (CFSRCNN), which uses a coarse-to-fine mechanism by cascading LR and HR features to address the performance degradation caused by unstable training.

Motivated by the mentioned methods, we firstly employ well-designed multi-scale feature extraction block and adopt the residual learning mechanism to capture long-distance dependent and more beneficial features in the LR image space in an effective way. And subsequently, the refinement block is utilized to extract detailed features from the coarse HR image, so that feature information from the LR and HR image spaces can be effectively integrated to further produce enriched high-frequency details.

III. PROPOSED NETWORK

There exist some problems for most CNN-based SR methods, such as extractor of features at a unitary scale, lack of the dependency between contextual features, and only use of feature information from the LR image space while

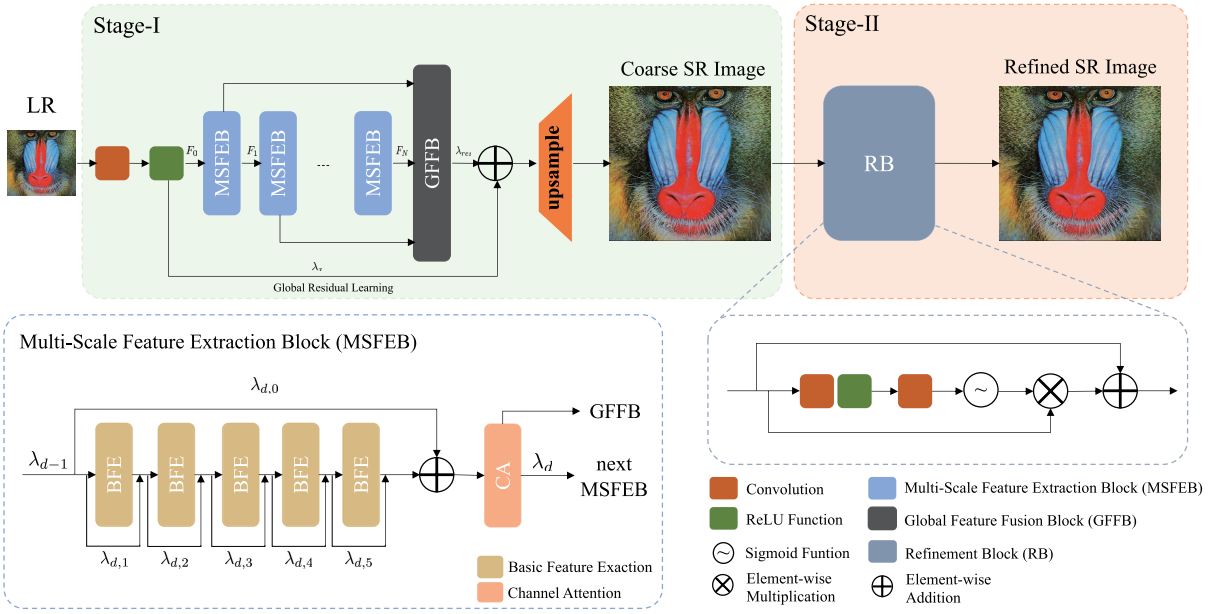


FIGURE 1. Network architecture of the proposed MESR. The proposed network is composed of two stages, gradually reconstructing LR images from coarse to fine.

abortive use of feature information from the HR image space. To solve the above issues, we propose a multi-stage enhancement network for super-resolution (MESR) from coarse to fine to reconstruct HR images with as abounding details as possible.

In this section, the network architecture of the proposed MESR will be presented first, followed by a detailed description of the multi-scale feature extraction block with channel attention, the residual features fusion block and the refinement block, and finally, the loss function used will be introduced.

A. NETWORK FRAMEWORK

The network architecture of MESR is depicted in Fig.1. A coarse reconstructed image is first generated for the input LR image in the first stage, and then the coarse image is further refined in the second stage to generate a final reconstructed image with abundant details. In particular, in the first stage, we adopt Multi-Scale Feature Extraction Block (MSFEB) to acquire features from different scales, and then utilize adaptive weighted Residual Feature Fusion Block (RFFB) to obtain a more beneficial feature representation, and subsequently go through the upsampling module to produce a coarse reconstructed image. In the second stage, the Refinement Block (RB) is investigated to optimize the obtained coarse reconstructed image. Specifically, feature extraction is first performed on the generated HR image space, and then the obtained features are further rearranged to highlight the important features, and eventually, the features extracted in this stage are superimposed with the features of the coarse reconstructed image generated in the previous stage to refine the final reconstructed image.

In the first stage, a 3×3 convolution kernel is used toperform feature extraction on the input LR image I_{LR}, and the shallow features representation F_0 is obtained as follows:

$$F_0 = \sigma\left(H_{ext}\left(I_{LR}\right)\right). \tag{1}$$

where, $H_{ext}(\cdot)$ denotes the shallow features extraction function and σ denotes the ReLU activation function. Then, the shallow features F_0 are then transferred to MSFEB for further obtaining the deep features.

Assuming that the MSFEB module corresponds to a function $H_{MSFEB}(\cdot)$, the depth feature extraction module can be formed by stacking N MSFEBs,

$$F_N = H^N_{MSFEB}\left(H^{N-1}_{MSFEB}\left(\ldots H^1_{MSFEB}\left(F_0\right)\right)\right) \tag{2}$$

For purpose of acquiring a more attractive feature representation, the proposed Global Feature Fusion Block (GFFB) is acclimated to fuse the features extracted from each MSFEB module at various levels. The fused features can be expressed as:

$$F_{FB} = H_{GFFB}\left(F_1, \cdots, F_N\right) \tag{3}$$

where, F_i indicates the features extracted by the i-th MSFEB module, $i = 1, 2, \cdots, N$; F_{FB} indicates the feature map obtained by fusing the features from different layers F_i using the composite function $H_{GFFB}(\cdot)$.

Furthermore, the Global Residual Learning (GRL) module is acclimatized to integrate shallow and deep features as well as use a sub-pixel convolution structure [55] is used to upsample the fused feature maps to obtain a coarse reconstructed image I_{SR1},

$$I_{SR1} = U\left(F_{FB} \oplus F_0\right) \tag{4}$$

where, $U(\cdot)$ denotes an upsampling operation, and $\oplus$ denotes a sum-by-element operation.

Finally, the RB module is designed to explore the detail features of the HR image space corresponding to I_{SR1} for refining the final reconstructed image I_{SR2},

$$I_{SR2} = R(I_{SR1}) \tag{5}$$

where, $R(\cdot)$ denotes the operation corresponding to the RB module.

B. MULTI-SCALE FEATURE EXTRACTION BLOCK

Given that most depth models only process the features acquired from the previous module and pay less attention to the correlation between the features obtained from adjacent modules, resulting in the lack of the dependency between contextual features within the image region. Inspired by ResNet [16], we designed a multi-scale feature extraction block (MSFEB) with adaptive weighted contextual residual features association structure is designed, as shown in the bottom-left of Fig.1. Specifically, each MSFEB contains five Basic Feature Exaction (BFE) modules and a Channel Attention (CA) module, where each BFE module and the corresponding skip connection (with weight value λ) together form a sub-residual block, thus effectively enhancing the dependency between contextual features. At the same time, to alleviate the absence of feature richness caused by single-scale convolution kernel, a funnel-like multi-scale feature extraction unit is proposed to effectively acquire feature information from different receptive field scales, which will be introduced in detail in Section III-B2.

Assuming that F_{d-1} is the input of the d-th MSFEB, the output $F_{d,1}$ of the first sub-residual block of the d-th MSFEB, the output $F_{d,B}$ of the B-th sub-residual block of the d-th MSFEB, and the output F_d of the d-th MSFEB can be expressed by the following expressions, respectively:

$$F_{d,1} = \lambda_{d,1} F_{d-1} \oplus H_{BFE}(F_{d-1}) \tag{6}$$

$$F_{d,B} = \lambda_{d,B} F_{d,B-1} \oplus H_{BFE}\left(F_{d,B-1}\right) \tag{7}$$

$$F_d = H_{CA}\left(\lambda_{d,5} F_{d,5} \oplus \lambda_{d,0} F_{d-1}\right) \tag{8}$$

where, $H_{BFE}(\cdot)$ and $H_{CA}(\cdot)$ denote the operations corresponding to the BFE and CA modules, respectively.

1) CHANNEL ATTENTION-BASED FEATURE EXTRACTOR

To fully advance the usage of extracted features, the channel attention mechanism is incorporated into the proposed MSFEB module. Channel attention [58] assigns different concerns to different channel features, so as to highlight the channel information related to the important features and suppress the invalid channel information [37]. In contrast to the mainstream channel attention mechanisms that employ dimensional reduction to abstract the correlation between channel features, an adaptive channel attention module [59] is adopted, which effectively avoids the adverse effect of dimensional reduction operation to disrupt the direct correlation between channels and its weights. The literature [59] points out that the trends of frequency signals are the same for different images in the same convolution layer, and exhibit strong local periodicity, therefore, the CA module used in this paper only computes the correlations for the k nearest neighboring channels.

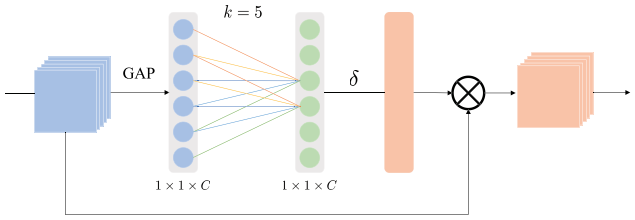


FIGURE 2. Channel attention module.

As shown in Fig.2, the CA module is composed of a global average pooling (GAP) layer, the nearest neighbor fully connected module, and a channel feature representation layer. Where, the nearest neighbor fully connected module only connects the k nearest channels to exploit the relationship between channels. Thus, the weight of the i-th channel y of the features y_i can be acquired as follows:

$$w_i = \delta\left(\sum_{j=1}^{k} \alpha_i^j y_i^j\right), \quad y_i^j \in \Omega_i^k \tag{9}$$

where, δ represents the Sigmoid activation function, α_i^j represents the parameter of the convolution kernel corresponding to y_i, and Ω_i^k represents the set of k adjacent channels of y_i.

Since the adaptive channel attention is intended to capture cross-channel information fittingly, it is incredibly necessary to determine the range of channel interaction k. From the previous analysis, it can see that k should be a certain mapping relationship to C. While the linear mapping relationship has certain limitations, and the number of channels in the SR model is normally a multiple of 2 to further raise the flexibility of the SR model, the mapping relationship can be represented as an exponential function with a base 2:

$$C = \Theta(k) = 2^{(\rho * k - b)} \tag{10}$$

where, ρ is the coefficient value and b is the offset value.

Consequently, according to the given C, k can be calculated,

$$k = \theta(C) = \left|\frac{\log_2(C)}{\rho} + \frac{b}{\rho}\right|_{odd} \tag{11}$$

where, $|t|_{odd}$ denotes the nearest odd number of t.

2) MULTI-SCALE CONVOLUTION UNIT

To further promote the feature representation ability, a common method is to construct a very deep network with tremendous convolution layers to explore more abundant features. Although such a method is able to boost the network performance to an assertive extent, it additionally brings added problems. On the one hand, the extending of the network

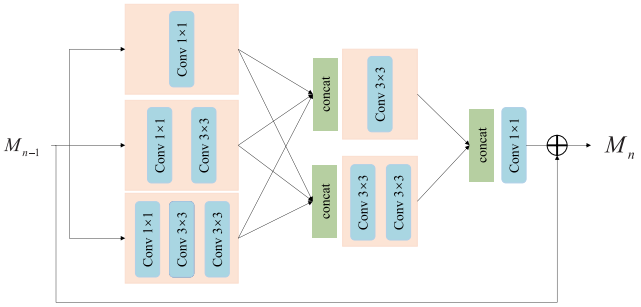


FIGURE 3. Funnel-like multi-scale feature extraction unit.

depth is accompanied by a continuous expansion in the number of parameters, which may prompt over-fitting in the case of inadequate training data. On the other hand, a network with an incredible number of layers implies higher computational complexity and is prone to introduce gradient exploding or vanishing, which makes the network difficult to optimize. Fortunately, the GoogLeNet [32] solves these problems well, as the critical thought of the Inception is to build a dense block structure. It utilizes multi-scale convolution kernels to acquire the feature map and then combines the output of several branches to the subsequent layer. The structure accelerates both the qualities and the resilience to scale without increasing the network depth of the network, thus improving the capability of feature extraction capability.

Although multi-scale features are widely used in low-level vision tasks such as image restoration and image super-resolution [60]–[68], most existing multi-scale feature extraction methods mainly suffer from the following issues. (1) Some methods perform a single scale convolution kernel on feature maps at different scales to achieve feature extraction, such as [60], [61]. However, the spatial and texture features in LR images are usually irregular and complex, and yet most of these methods rely on the feature information at specific image scales (*e.g.*, feature maps at specific scales such as $\times 2$, $\times 3$, *etc.*) in the spatial dimension, and cannot fully capture high-frequency detail information on images at the same scale. (2) Although some methods apply multiple-scale convolution kernels for feature extraction on the same scale feature map, such methods tend to adopt larger convolution kernels, such as 5×5, 7×7, and even 9×9 [62]–[64]. These methods effectively cover the features at different scales, while they entail a larger number of parameters and computational costs. (3) Some methods [65]–[68] implement feature extraction that uses smaller convolution kernels, but ignores the correlation between feature information internally, which limits the features representation capability.

Inspired by the Inception module, a funnel-like multi-scale feature extraction unit is designed for the SR task, which can be used to adaptively extract abounding features at different scales by introducing convolution kernels with different receptive field scales. Moreover, skip connections are used between different scales to effectively enhance the inherent complementarity and connection of features in local regions. Specifically, convolution kernels of different sizes are integrated into the network firstly, such as 1×1, 3×3, and 5×5. Nevertheless, considering that larger-scale convolution kernels will cause larger computational costs and increase the time complexity, smaller-scale convolution kernels are adapted instead of large-scale ones to minimize the number of parameters [33], [69]. In the network design, only convolution kernels of 1×1 and $3 \times 3$are used, while convolution kernels of 5×5 are replaced by two cascaded 3×3 convolution kernels[50]. As seen in Fig.3, a multi-scale feature extraction structure with several levels is designed, which contains three levels and several branches in each level. In the first level, there are three branches with convolution kernel scales of 1×1, 1×1 and 3×3, 1×1 and a set of cascaded 3×3, which correspond to different receptive fields. In the subsequent levels, the multi-scale features are fused from the previous level, so that the parameter information from different convolution kernels can be shared to obtain more representative feature information. Given M_{n-1} be the input of the module, and the output features M_n of the module can be expressed as follows.

$$f_{1,1} = \sigma\left(w_{1\times1}^{1,1} \otimes M_{n-1} + b^1\right) \tag{12}$$

$$f_{1,2} = \sigma\left(w_{3\times3}^{1,2} \otimes \sigma\left(w_{1\times1}^{1,1} \otimes M_{n-1}\right) + b^1\right) \tag{13}$$

$$f_{1,3} = \sigma\left(w_{3\times3}^{1,3} \otimes \sigma\left(w_{3\times3}^{1,2} \otimes \sigma\left(w_{1\times1}^{1} \otimes M_{n-1}\right)\right) + b^1\right) \tag{14}$$

$$f_{2,1} = \sigma\left(w_{3\times3}^{2,1} \otimes \left[f_{1,1}, f_{1,2}, f_{1,3}\right] + b^2\right) \tag{15}$$

$$f_{2,2} = \sigma\left(w_{3\times3}^{2,2} \otimes \sigma\left(w_{3\times3}^{2,1} \otimes \left[f_{1,1}, f_{1,2}, f_{1,3}\right]\right) + b^2\right) \tag{16}$$

$$f_3 = w_{1\times1}^{3,1} \otimes \left[f_{2,1}, f_{2,2}\right] + b^3 \tag{17}$$

$$M_n = \beta M_{n-1} \oplus f_3 \tag{18}$$

where, σ denotes the ReLU activation function, $\otimes$ denotes the element multiplication operation, and $\oplus$ denotes the element addition operation. W and b denote the weight and bias values of the corresponding convolution layers, respectively, where the number in the superscript of W denotes the j-th convolution module in the i-th layer, and the subscript of W denotes the size of the corresponding convolution kernel, and the number in the superscript of b denotes the number of layers in which it is located. $\left[f_{1,1}, f_{1,2}, f_{1,3}\right]$ and $\left[f_{2,1}, f_{2,2}\right]$ in (15),(16),(17) denote the cascade operation, which means that the corresponding features are cascaded according to the channel dimension to derive a more abundant feature description.

It is worth noting that in a way to make the extracted features representative, the residual structure is introduced, as shown in (18), by assigning weights β to the input features and then fusing them with the features extracted by the multi-scale structure to acquire the output features of the module.

C. RESIDUAL FEATURE FUSION BLOCK

With the acquisition of features at different scales, an adaptive weighted Residual Feature Fusion Block (RFFB) is further proposed to exploit the dependency between contextual features to generate more beneficial features. RFFB is composed of a Global Feature Fusion Block (GFFB) and a Global Residual Learning (GRL) module.

1) GLOBAL FEATURE FUSION BLOCK

The conventional residual module, which usually consists of a series of feature extraction modules, fuses the original features with the final extracted residual features through long-skip connections and transmits the fused features to the subsequent modules. Such a design only collects complex features, but does not fully utilize the original features obtained from each feature extraction module, and it is, more importantly, the dependency between contextual features cannot be exploited, limiting the feature representation performance of the network.

The GFFB module is designed to fuse as well as possible the features from all MSFEB modules. However, if all the features are simply superimposed together, there will be too numerous redundant features and the number of parameters in the network will increase drastically, making the network even more difficult to train. To address the above problem, the features obtained from each MSFEB module are adaptively fused in the proposed network. As shown in Fig.1, the features obtained from each MSFEB are transferred to the GFFB structure through the long-skip connection structure, making full use of the correlation of each contextual feature $[F_1, F_2, \cdots, F_N]$, which not only improves the features utilization, but also makes the contextual feature transfer more convenient. Meanwhile, to ease the training difficulty of the network, inspired by MemNet[39], a 1×1 convolution layer is introduced after the features fusion module for controlling the number of output features to reduce the dimensional of the features.

2) GLOBAL RESIDUAL LEARNING

Since shallow features contain a large number of edge texture information, and deep features contain rich semantic information, both of them play a key role in improving the quality of reconstructed images for SR tasks. Therefore, to effectively integrate the shallow features with the deep features, an effective global residual learning mechanism is employed to maximize the utilization of the features.

In order to further improve the performance of residual features, Jung *et al.* [70] proposed a weighted Residual Unit (wRU), which allowed the generation of different weights of the residual unit for different features of the input through a weight Squeeze-and-Excitation (wSE) structure. Although this approach improves the performance of the residual units, it brings additional parameters and the costs of calculating weights to the wSE structure. Therefore, inspired by wRU, an adaptive weighted residual unit is used that adaptively learns the weights of residual features. Specifically, the Weight Normalization (WN) operation is used for adaptive updating of weights. Since WN is a parameterization operation to reparameterize the weight vector, the weights are scaled by separating the direction and length of the weight vector. Not only does the WN avoid the effect of the mini-batch, but also effectively reduces the effect of noisy data generated by the mini-batch during the gradient propagation process. More importantly, the WN does not cause additional space and parameters for saving the variance and mean of the data required for the normalization operation, which makes it particularly suitable for the task of updating the weights in the proposed residual modules. As shown in Fig.1, all λ are learnable parameters, whose initial values are set to 1, and the weights are updated adaptively through continuous iterative learning in the proposed MESR.

D. REFINEMENT BLOCK

Currently, most of the SR methods usually add an upsampling module at the end of the network to enlarge the size of the feature map to obtain the reconstructed image directly. These SR methods merely extract and map the features in the LR image space, without exploring the HR image space information corresponding to the coarse reconstructed images generated by upsampling. However, with the limited information in the LR image space, it is incomplete to learn the coarse reconstructed image features by extracting features directly from the LR image space to portray the potential HR image features [21]. Simultaneously, the existing upsampling module will possibly cause training instability and performance degradation in the process of scaling [21]. Consequently, we need to assemble a learning mechanism that effectively explores the features of HR image space, extracts the feature information of it in a valid way, and further refines the high-frequency details of the reconstructed images to minimize the inconsistency between the generated reconstructed images and the original HR image.

The literature of [21], [25]–[28] have revealed that the rational use of feature information of the coarse reconstructed images helps to refine the details of the final reconstructed image. Therefore, a Refinement Block (RB) for image detail optimization is applied, which optimizes the coarse reconstructed image into a more detailed reconstructed image by extracting the high-frequency features of the coarse image to compensate for the lost local detail information of the LR image. It is able in capturing the commutual SR features and abbreviation the information loss caused by the upsampling structure.

The RB module proposed can be represented as:

$$I_{SR2} = I_{SR1} \otimes \delta\left(H_{1\times1}\left(\sigma\left(H_{3\times3}\left(I_{SR1}\right)\right)\right)\right) + I_{SR1} \quad (19)$$

where, $\otimes$ represents the element multiplication operation, $H_{1\times1}$ and $H_{3\times3}$ represent the 1×1 and 3×3 convolution operations, respectively, δ represents the Sigmoid activation function, σ represents the ReLU activation function, and I_{SR1} and I_{SR2} represent the coarse and refined reconstructed

images, respectively. Although the proposed RB module is relatively simple, it is essential for reconstructing images containing more detailed information. The necessity of this structure will be analyzed in Section IV-C2.

E. LOSS FUNCTION

For the proposed MESR model, a new loss function is designed to further strengthen the super-resolution effect by carrying out loss calculations on the reconstructed images generated in two stages.

Most of the current SR methods use Mean Square Error (MSE) as the loss function, while it produces smooth reconstructed images and has insufficient supervision on high-frequency information such as contours of the image [15]. Compared with MSE, the L_1 loss can supervise the high-frequency information to a certain extent, but it may have unpredictable effects in the super-resolution process as it is not derivable at the zero point. Considering that Charbonnier Loss [35] can not only effectively tackle the problem of non-derivability at the zero point, but also accelerate the convergence speed of the model, the Charbonnier Loss is applied to construct the loss function in the proposed work, and its expression is as follows.

$$\mathcal{L} = \omega_1 \mathcal{L}_C \left(I_{SR1}, I_{GT}\right) + \omega_2 \mathcal{L}_C \left(I_{SR2}, I_{GT}\right) \tag{20}$$

where, ω_1 and ω_2 are two constants, $\mathcal{L}_C$ denote Charbonnier Loss, and the Charbonnier loss as follows:

$$\mathcal{L}_C = \sqrt{\|I_{SRi} - I_{GT}\|^2 + \zeta^2} \tag{21}$$

where, I_{SRi} denotes the reconstructed image of the i-th stage involved in the reconstruction, I_{GT} denotes the original high-resolution image, and ζ is a small constant, usually set to 10^{-3}.

IV. EXPERIMENTAL AND ANALYSIS

A. DATASETS AND METRICS

To ensure the objectivity of the experiments, the standard dataset DIV2K [71] is selected as the training dataset, which contains 800 2K training images and 200 validation images. Meanwhile, four common benchmark datasets, including Set5 [72], Set14 [73], BSD100 [74], and Urban100 [52], are employed as test sets and tested on three scaling factors (*i.e.*, ×2, ×3, ×4). For evaluation, the reconstructed images obtained were first transformed from RGB space to YCbCr space, and then the metrics were evaluated on the Y channel, including Peak Signal to Noise Ratio (PSNR) [75] and Structural Similarity (SSIM) [76],

$$PSNR = 10 \cdot \log_{10} \frac{(255)^2 \cdot QP}{\|x - \tilde{x}\|^2} \tag{22}$$

$$SSIM = \frac{(2\mu_{\tilde{x}}\mu_x + C_1)\,(2\sigma_{\tilde{x}x} + C_2)}{(\mu_{\tilde{x}}^2 + \mu_x^2 + C_1)\,(\sigma_{\tilde{x}}^2 + \sigma_x^2 + C_2)} \tag{23}$$

where, $\tilde{x}$ denotes the reconstructed image, x denotes the original high-resolution image, Q and P denotes the number of rows and columns of x, respectively. $\mu_{\tilde{x}}$, μ_x and $\sigma_{\tilde{x}}^2$, σ_x^2 are the corresponding mean and variance of $\tilde{x}$ and x, respectively, $\sigma_{\tilde{x}x}$ denotes the covariance, and C_1 and C_2 are two constants.

TABLE 1. The effectiveness of different modules. The results evaluate for a scaling factor of 4 on Set14.

Baseline	MSFEB	RFFB	RB	PSNR(dB)
✓				28.18
	✓			28.37
	✓	✓		28.61
	✓	✓	✓	28.72

B. PARAMETER SETTINGS

During training, Bicubic downsampling is performed on 800 HR images in the DIV2K to obtain the corresponding LR images. In the experiments, random horizontal flipping and rotation are incorporated to enhance the generality of the training images. The batch size is set to 16 and randomly cropped the LR image into 32 × 32 patches as input. The training is optimized by the Adam [77] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. Moreover, the learning rate is initialized as $lr = 10^{-4}$, which is reduced every 200 epochs. In practice, $N = 24$ and $k = 5$. The proposed MESR has been executed on the PyTorch [78] framework and trained to utilize an NVIDIA TITAN RTX 24GB GPU.

C. ABLATION EXPERIMENT

The proposed MESR model involves three major modules, including Multi-Scale Feature Extraction Block (MSFEB), Residual Feature Fusion Block (RFFB), and Refinement Block (RB). In order to verify the effectiveness of these three modules, extensive experiments are conducted on different models.

1) EFFECTIVENESS EVALUATION OF MSFEB, RFFB AND RB MODULES

To verify the impact of the proposed MSFEB, RFFB, and RB modules on the super-resolution performance, the corresponding evaluation experiments are designed. First, the model only uses a single-scale convolution kernel of 3 × 3 as the feature extraction module, and the network structure without the RFFB and RB modules as the baseline for reference. After that, adding the MSFEB, RFFB, and RB modules to the baseline model gradually, and the results are shown in Table 1. In these experiments, the test benchmark is selected as Set14.

Table 1 lists the PSNR metric of the reconstructed images obtained by different models on Set14 with a scaling factor of 4. From Table 1, it can be seen that the PSNR is 28.18 dB when a single-scale convolution kernel is used for feature extraction, whereas the PSNR is improved by 0.19 dB when a multi-scale feature extraction module is used to replace the single-scale convolution kernel. In contrast, the performance improvement of the MSFEB module is limited, while the super-resolution performance is significantly improved when the RFFB module is added to the model. This is due to the fact

that the RFFB module is able to effectively exploit the dependency between contextual features, which improves the feature representation ability and facilitates the information flow, thus reconstructing more edge and texture details. Then, the RB module is further added to the model to compose the final MESR model, and the PSNR of the reconstructed image is advanced again. On the one hand, the RB module is able to explore effective high-frequency feature information from the HR image space and optimize the coarse reconstructed image into a refined reconstructed image. On the other hand, the MESR model is able to simultaneously utilize feature information from both LR and HR image spaces, which greatly enhances the overall super-resolution performance. Consequently, the joint MSFEB, RFFB, and RB modules play an active effect in promoting the super-resolution performance.

2) NECESSITY EVALUATION OF RB MODULE

To further evaluate the necessity of the proposed RB module, this section will analyze the reconstructed images in terms of both quantitative metrics and visual quality. Table 2 illustrates the comparison of quantitative metrics of different models in different benchmarks. Furthermore, Fig.4 depicts the related subjective visual images to show the superiority of the RB module more intuitively. To confirm the effectiveness of the RB module further, a variant of RB† with RB module is designed, which setting ω_1 of the loss function to 0 and the setting ω_2 to 1. It can be noted that the super-resolution performance is negatively affected when ω_1 set to 0. Due to the lack of feature information from the LR image space, the training process becomes difficult and the super-resolution performance is inhibited. This also indicates that the proposed multi-scale feature extraction module with channel attention mechanism and multiple residual learning mechanism can not only extract enriched contextual features, but also enable the communication between contextual features to be more fluent; in addition, the proposed RB module is able to further refine the coarse reconstructed images into the final reconstructed images with abundant details by exploiting the feature information from both LR and HR image spaces.

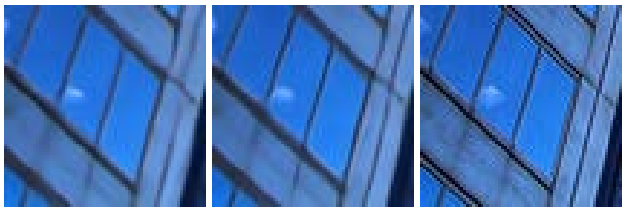
From the subjective visual quality, comparing the coarse reconstructed image (Fig.4a) obtained in the first stage and the refined reconstructed image (Fig.4b) generated in the second stage, it can be noticed that the reconstructed image optimized by the RB module is more prominent in terms of the high-frequency details. Specifically, the edges within the upper-right region in Fig.4(b) are sharper and clearer, and compared with the ground truth (GT) image (Fig.4c), both the structure and details are well preserved. This is owing to the fact that the RB module complements the missing information in the LR image and effectively captures the complementary SR features, thus effectively upgrading the subjective visual quality.

D. COMPARISONS WITH STATE-OF-THE-ART METHODS

To further validate the effectiveness of the proposed MESR model, a comparison is performed with thirteen CNN-based SR methods, including SRCNN [25], VDSR [13], LapSRN [35], DRCN [53], IDN [56], DASR [79], DPSR [80], PAN [7], LAPAR-A [81], DeFiAN [9], MSRN [19], DRSAN-48m [82] and A2F-L [83]. The best average PSNR and SSIM of the reconstructed images obtained by the above SR methods using diverse scaling factors (*i.e.*, $\times 2$, $\times 3$, $\times 4$) on the four standard benchmarks of Set5, Set14, BSD100, and Urban100 are listed in Table 3, where the optimal and sub-optimal values are marked in **bold** and underlined, respectively.

TABLE 2. The necessity of RB. The results evaluate for a scaling factor of 4.

MESR	Set5	Set14	BSD100	Urban100
w/o RB	32.26	28.61	27.57	26.18
w/ RB†	32.22	28.65	27.51	26.10
w/ RB	**32.39**	**28.72**	**27.64**	**26.29**



(a) w/o RB (b) w/ RB (c) Ground Truth (GT)

FIGURE 4. Visual comparisons of the necessity of RB on 'Img_012' from Urban100 with a scaling factor of 4.

It can be assured from Table 3 that the mean PSNR and SSIM of the reconstructed images obtained from the proposed MESR performs optimally in most of the datasets and sub-optimally in some datasets only. This indicates that our method achieves a competitive performance in terms of overall super-resolution performance. Owing to the multi-stage learning mechanism adopted in this paper, MESR successfully integrates the feature information from LR and HR image spaces, and combines the multi-scale feature extraction module and the multiple residual learning mechanism to make the communication between the obtained contextual features more fluent.

1) VISUAL QUALITY

To visually compare the reconstructed images obtained by different SR methods from the perspective of subjective visual effect, Fig.5 to Fig.8 illustrate the reconstructed images of different methods in the same region with a scaling factor of 4, and the corresponding original HR image is given as a reference. From Fig.5 to Fig.8, it can be seen that most of comparison methods failed to accurately reconstruct edge and texture details, and even generated severe artifacts, while MESR has the ability to produce reconstructed images with abundant edge and texture details.

Fig.5 presents the visual comparison of various methods on the ''Barbara'' image in Set5. As can be seen in Fig.5, the reconstructed images obtained by SRCNN, VDSR, LapSRN,

TABLE 3. Qualitative results of different SR models on standard benchmarks.

Methods	Scale Factor	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN	2	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
VDSR		37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140
LapSRN		37.52	0.9590	33.08	0.9130	31.80	0.8950	30.41	0.9100
DRCN		37.63	0.9588	33.04	0.9118	31.85	0.8942	30.75	0.9133
IDN		37.83	0.9600	33.30	0.9148	32.08	0.8985	31.27	0.9196
DASR		37.87	0.9599	33.34	0.9160	32.03	0.8986	31.49	0.9227
DPSR		37.77	0.9591	33.48	0.9164	32.12	0.8984	31.87	0.9256
PAN		38.00	0.9605	33.59	0.9181	32.18	0.8997	32.01	0.9273
LAPAR-A		38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283
DeFiAN		38.03	0.9605	33.62	0.9181	32.20	0.8999	32.20	0.9286
MSRN		38.08	0.9607	33.70	0.9186	32.23	0.9002	32.29	0.9303
DRSAN-48m		**38.14**	0.9611	33.75	0.9188	32.25	**0.9010**	**32.46**	**0.9317**
A2F-L		38.09	0.9607	33.78	0.9192	32.23	0.9002	**32.46**	0.9313
MESR		**38.14**	**0.9613**	**33.84**	**0.9200**	**32.27**	0.9004	**32.46**	0.9307
SRCNN	3	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989
VDSR		33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279
LapSRN		33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280
DRCN		33.82	0.9226	29.76	0.8311	28.80	0.7963	27.15	0.8276
IDN		34.11	0.9253	29.99	0.8354	28.95	0.8013	27.42	0.8359
DASR		34.11	0.9254	30.13	0.8408	28.96	0.8015	27.65	0.8450
DPSR		34.32	0.9259	30.25	0.8410	29.08	0.8044	28.07	0.8504
PAN		34.40	0.9271	30.36	0.8423	29.11	0.8050	28.11	0.8511
LAPAR-A		34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523
DeFiAN		34.42	0.9273	30.34	0.8410	29.10	0.8053	28.20	0.8528
MSRN		34.46	0.9278	30.41	0.8437	29.15	0.8064	28.33	0.8561
DRSAN-48m		**34.59**	0.9286	30.42	0.8443	29.18	**0.8079**	28.52	0.8593
A2F-L		34.54	0.9283	30.41	0.8436	29.14	0.8062	28.40	0.8574
MESR		34.54	**0.9289**	**30.51**	**0.8452**	**29.21**	0.8077	**28.57**	**0.8600**
SRCNN	4	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221
VDSR		31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524
LapSRN		31.53	0.8850	28.19	0.7720	27.32	0.7280	25.21	0.7560
DRCN		31.53	0.8854	28.02	0.7670	27.23	0.7233	25.14	0.7510
IDN		31.82	0.8903	28.41	0.7730	27.41	0.7297	25.41	0.7632
DASR		31.99	0.8923	28.50	0.7799	27.51	0.7346	25.82	0.7742
DPSR		32.19	0.8945	28.65	0.7829	27.58	0.7354	26.15	0.7864
PAN		32.13	0.8948	28.61	0.7822	27.59	0.7363	26.11	0.7854
LAPAR-A		32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871
DeFiAN		32.16	0.8942	28.62	0.7810	27.57	0.7363	26.10	0.7862
MSRN		32.26	0.8960	28.63	0.7836	27.61	0.7380	26.22	0.7911
DRSAN-48m		32.34	0.8960	28.65	0.7841	27.63	**0.7390**	**26.33**	0.7936
A2F-L		32.32	0.8964	28.67	0.7839	27.62	0.7379	26.32	0.7931
MESR		**32.39**	**0.8980**	**28.72**	**0.7855**	**27.64**	0.7385	26.29	**0.7943**

and DRCN are relatively blurred. Although IDN, PAN, DPSR and DeFiAN restore more high-frequency information, their reconstructed images contain severe artifacts. As compared with the previous methods, MSRN, A2F-L, LAPAR-A and DASR achieve superior super-resolution results due to restoring more abundant high-frequency information. Nevertheless, compared with the methods mentioned above, our MESR has better performance in both contour preservation and detail restoration, which is reflected by the sharper overall contour of the reconstructed image and more delicate edge and texture details.

Fig.6 depicts the visual comparison of various methods on the "Zebra" image in Set14. It can be seen from Fig.6 that the reconstructed image obtained by the proposed MESR preserves the texture details on the zebra well, while the reconstructed images generated by SRCNN, VDSR, LapSRN, DRCN, DPSR, PAN and DeFiAN are blurred because they are unable to restore the texture details effectively. Even though IDN, DASR, LAPAR-A and MSRN can roughly restore the texture details, they also generate a small amount of artifacts. Compared with A2F-L, the proposed MESR is more prominent in detail fidelity, which is mainly related to the fact that the proposed MESR effectively fuses the extracted features from each layer and exploit the correlation of channel features with the channel attention mechanism.

Fig.7 illustrates the visual comparison of various methods on the "210088" image in BSD100. As can be seen from Fig.7, though the reconstructed image of LAPAR-A outperforms that of SRCNN, VDSR, LapSRN, DRCN, IDN, DASR and DPSR in terms of edge detail restoration, it suffers from a large number of artifacts. In contrast to PAN and

TABLE 4. Comparison of the number of parameters, Multi-Adds and PSNR obtained by different methods on Set14 with a scaling factor of 2.

Methods	SRCNN	VDSR	LapSRN	DRCN	IDN	DASR	DPSR
Param.	57K	666K	813K	1774K	579K	5969K	3162K
Multi-Adds	52.7G	612.6G	29.9G	17974.3G	-	-	-
PSNR	32.45	33.03	33.08	33.04	33.30	33.34	33.48
Methods	PAN	LAPAR-A	DeFiAN	MSRN	DRSAN-48m	A2F-L	MESR
Param.	261K	548K	1027K	5930K	377K	1363K	4766K
Multi-Adds	70.5G	171G	-	429.3G	274.6G	306.1G	237.2G
PSNR	33.59	33.62	33.62	33.83	33.75	33.78	33.84

FIGURE 5. Visual comparison of super-resolution results of 'Barbara' (Set5) obtained by different SR algorithms with scaling factor ×4.

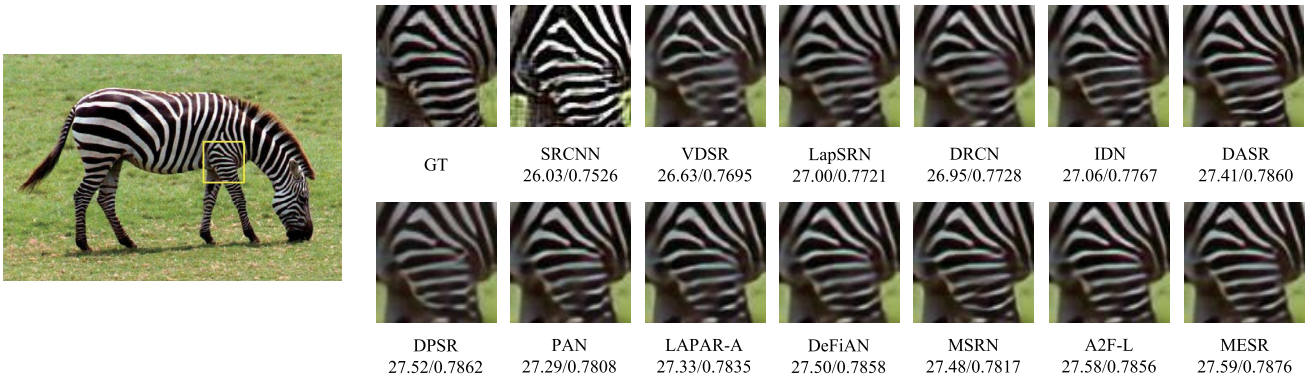


FIGURE 6. Visual comparison of super-resolution results of 'Zebra' (Set14) obtained by different SR algorithms with scaling factor ×4.

DeFiAN, the reconstructed images generated by MSRN and A2F-L contain fewer artifacts. Even though both MSRN and A2F-L effectively suppresses the artifacts, they are still slightly weaker than the proposed MESR in terms of edge detail restoration.

Fig.8 presents the visual comparison of various methods on the "Img_012" image in Urban100. From Fig.8, it can be seen that the reconstructed images obtained by SRCNN, VDSR, LapSRN, DRCN, IDN and DPSR suffers from severe blurring. While PAN, DASR, LAPAR-A, and DeFiAN are able to restore the main image contour, they fail to retrieve further details of the image. Compared with A2F-L and MSRN, the proposed MESR can not only restore more high-frequency details, but also reconstruct sharper reconstructed images. It is mainly due to the fact that the proposed MESR further refines the details of the reconstructed image by using the RB module.

E. MODEL ANALYSIS

To comprehensively measure the performance of different methods, Table 4 illustrates the comparison of the number of parameters, Multi-Adds, and PSNR metric of different methods on the Set14 with a scaling factor of 2. As can be apparent in Table 4, our MESR achieves a better trade-off in the super-resolution performance with the number of parameters and Multi-Adds compared to other comparison methods. Specifically, compared with MSRN and DASR, the proposed

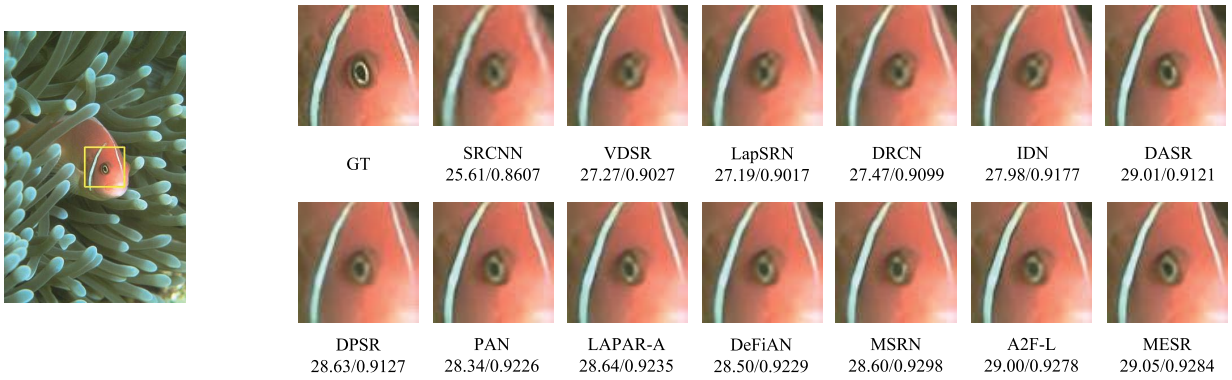


FIGURE 7. Visual comparison of super-resolution results of '210088' (BSD100) obtained by different SR algorithms with scaling factor ×4.

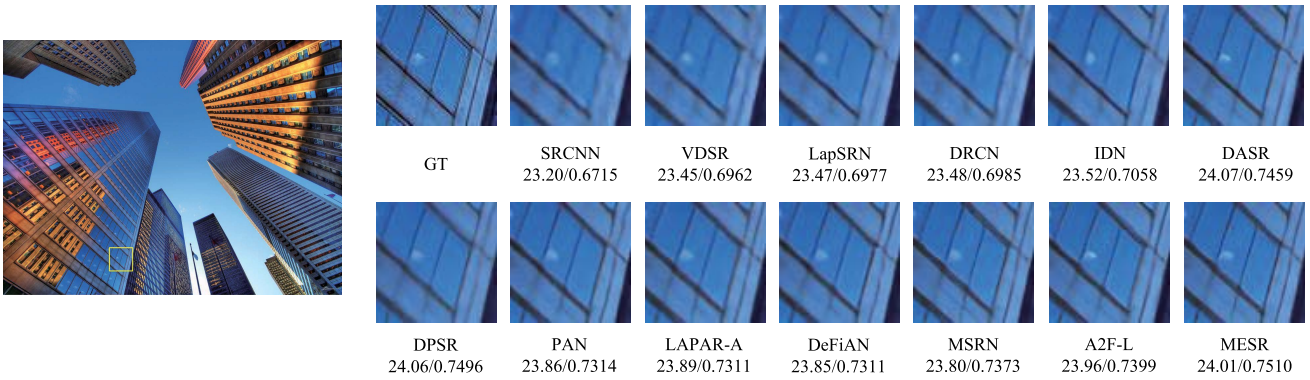


FIGURE 8. Visual comparison of super-resolution results of 'Img_012' (Urban100) obtained by different SR algorithms with scaling factor ×4.

MESR obtains a higher PSNR with nearly 1/6 less number of parameters.

V. CONCLUSION

To reconstruct as many high-frequency details as possible, a multi-stage enhancement network for image super-resolution (MESR) is proposed in this paper. The network consists of two stages, from coarse to fine, to further upgrade the quality of the reconstructed images by fully exploring and utilizing the feature information of LR and HR images. Firstly, the proposed multi-scale feature extraction module, combined with the channel attention mechanism, effectively extracts abounding features at different scales, enabling the network to better adapt to the scale variation in the super-resolution process. Then, the adaptive weighted residual feature fusion block is designed to explore the dependency between contextual features. Finally, a refinement block is constructed to exploit the features in the HR image space to strengthen the details of the reconstructed images. The experimental results validate that both visual quality and evaluation metrics of our MESR have achieved superior performance in comparison with state-of-the-art methods. However, the proposed MESR model still has a large number of parameters. Therefore, in our future work, it is necessary to focus on the compression of the MESR model to further reduce the computational complexity while ensuring the super-resolution performance.

REFERENCES

[1] Y. Zhou, Y. Zhang, X. Xie, and S.-Y. Kung, "Image super-resolution based on dense convolutional auto-encoder blocks," *Neurocomputing*, vol. 423, pp. 98–109, Jan. 2021.

[2] Y. Yan, W. Ren, X. Hu, K. Li, H. Shen, and X. Cao, "SRGAT: Single image super-resolution with graph attention network," *IEEE Trans. Image Process.*, vol. 30, pp. 4905–4918, 2021.

[3] Z. Chen, X. Guo, P. Y. M. Woo, and Y. Yuan, "Super-resolution enhanced medical image diagnosis with sample affinity interaction," *IEEE Trans. Med. Imag.*, vol. 40, no. 5, pp. 1377–1389, May 2021.

[4] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Articulated Motion and Deformable Objects*, F. J. Perales and J. Kittler, Eds. Cham, Switzerland: Springer, 2016, pp. 175–184.

[5] H. Huan, P. Li, N. Zou, C. Wang, Y. Xie, Y. Xie, and D. Xu, "End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network," *Remote Sens.*, vol. 13, no. 4, p. 666, Feb. 2021.

[6] Y. Fu, J. Chen, T. Zhang, and Y. Lin, "Residual scale attention network for arbitrary scale image super-resolution," *Neurocomputing*, vol. 427, pp. 201–211, Feb. 2021.

[7] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Computer Vision—ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham, Switzerland: Springer, 2020, pp. 56–72.

[8] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "HINet: Half instance normalization network for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 182–192.

[9] Y. Huang, J. Li, X. Gao, Y. Hu, and W. Lu, "Interpretable detail-fidelity attention network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 2325–2339, 2021.

[10] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.

[11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[12] H. Wang, X. Hu, X. Zhao, and Y. Zhang, "Wide weighted attention multi-scale network for accurate MR image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 962–975, Mar. 2022.

[13] 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.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "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.

[15] 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.

[16] 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.

[17] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 256–272.

[18] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[19] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 527–542.

[20] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2356–2365.

[21] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2021.

[22] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 391–407.

[23] Y. Lu, Z. Jiang, G. Ju, L. Shen, and A. Men, "Recursive multi-stage upscaling network with discriminative fusion for super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 574–579.

[24] F. Li, H. Bai, and Y. Zhao, "Detail-preserving image super-resolution via recursively dilated residual network," *Neurocomputing*, vol. 358, pp. 285–293, Sep. 2019.

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

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

[27] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3862–3871.

[28] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng, "Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 994–1006, Feb. 2018.

[29] Y. Hu, X. Gao, J. Li, Y. Huang, and H. Wang, "Single image super-resolution with multi-scale information cross-fusion network," *Signal Process.*, vol. 179, Feb. 2021, Art. no. 107831.

[30] Y. Zhang, Y. Wu, and L. Chen, "MSFSR: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2120–2129.

[31] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[35] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[36] Y. Lu, Y. Zhou, Z. Jiang, X. Guo, and Z. Yang, "Channel attention and multi-level features fusion for single image super-resolution," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.

[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 294–310.

[38] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

[39] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 191–207.

[40] S. Zagoruyko and N. Komodakis, "DiracNets: Training very deep neural networks without skip-connections," 2017, *arXiv:1706.00388*.

[41] L. Zhao, M. Li, D. Meng, X. Li, Z. Zhang, Y. Zhuang, Z. Tu, and J. Wang, "Deep convolutional neural networks with merge-and-run mappings," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 3170–3176.

[42] Y. Yan, X. Xu, W. Chen, and X. Peng, "Lightweight attended multi-scale residual network for single image super-resolution," *IEEE Access*, vol. 9, pp. 52202–52212, 2021.

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

[44] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using Delaunay triangulation," *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1427–1441, Dec. 2002.

[45] M. K. Ozkan, A. M. Tekalp, and M. I. Sezan, "POCS-based restoration of space-varying blurred images," *IEEE Trans. Image Process.*, vol. 3, no. 4, pp. 450–454, Jul. 1994.

[46] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[47] K. Xu, X. Wang, X. Yang, S. He, Q. Zhang, B. Yin, X. Wei, and R. W. H. Lau, "Efficient image super-resolution integration," *Vis. Comput.*, vol. 34, nos. 6–8, pp. 1065–1076, Jun. 2018.

[48] 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.

[49] K. S. Ni, S. Kumar, N. Vasconcelos, and T. Q. Nguyen, "Single image superresolution based on support vector regression," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process.*, vol. 2, May 2006, p. 2.

[50] 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./Jul. 2004, p. 1.

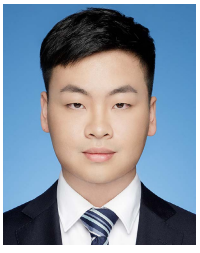
[51] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[52] 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.

[53] 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.

[54] 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.
[55] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[56] 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.
[57] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1–10.
[58] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.
[59] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[60] G. Jiang, Z. Lu, X. Tu, Y. Guan, and Q. Wang, "Image super-resolution using multi-scale space feature and deformable convolutional network," *IEEE Access*, vol. 9, pp. 74614–74621, 2021.
[61] A. Esmaeilzehi, M. O. Ahmad, and M. N. S. Swamy, "MGHCNET: A deep multi-scale granular and holistic channel feature generation network for image super resolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[62] X. Li, F. Xu, X. Lyu, Y. Tong, Z. Chen, S. Li, and D. Liu, "A remote-sensing image pan-sharpening method based on multi-scale channel attention residual network," *IEEE Access*, vol. 8, pp. 27163–27177, 2020.
[63] Y. Sun, Y. Zhang, S. Liu, W. Lu, and X. Li, "Image super-resolution using supervised multi-scale feature extraction network," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 1995–2008, 2021.
[64] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "MDCN: Multi-scale dense cross network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2547–2561, Jul. 2021.
[65] A. Liu, S. Li, and S. Chen, "A progressive network based on residual multi-scale aggregation for image super-resolution," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.
[66] N. Wang, W. Wang, W. Hu, A. Fenster, and S. Li, "Thanka mural inpainting based on multi-scale adaptive partial convolution and stroke-like mask," *IEEE Trans. Image Process.*, vol. 30, pp. 3720–3733, 2021.
[67] W. Shi, F. Jiang, and D. Zhao, "Single image super-resolution with dilated convolution based multi-scale information learning inception module," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 977–981.
[68] S. Gao and X. Zhuang, "Multi-scale deep neural networks for real image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2006–2013.
[69] W. Ma, Y. Wu, Z. Wang, and G. Wang, "MDCN: Multi-scale, deep inception convolutional neural networks for efficient object detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2510–2515.
[70] H. Jung, R. Lee, S.-H. Lee, and W. Hwang, "Active weighted mapping-based residual convolutional neural network for image classification," *Multimedia Tools Appl.*, vol. 80, no. 24, pp. 33139–33153, Oct. 2021.
[71] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.
[72] 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.*, 2012, pp. 135.1–135.10.
[73] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Germany: Springer, 2012, pp. 711–730.
[74] 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. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Feb. 2001, pp. 416–423.
[75] Y. Wang, J. Li, Y. Lu, Y. Fu, and Q. Jiang, "Image quality evaluation based on image weighted separating block peak signal to noise ratio," in *Proc. Int. Conf. Neural Netw. Signal Process.*, vol. 2, Apr. 2003, pp. 994–997.
[76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[78] N. Ketkar and J. Moolayil, *Automatic Differentiation in Deep Learning*. Berkeley, CA, USA: Apress, 2021, pp. 133–145.
[79] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10576–10585.
[80] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.
[81] W. Li, K. Zhou, L. Qi, N. Jiang, J. Lu, and J. Jia, "LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20343–20355.
[82] K. Park, J. W. Soh, and N. I. Cho, "Dynamic residual self-attention network for lightweight single image super-resolution," *IEEE Trans. Multimedia*, early access, Dec. 9, 2021, doi: 10.1109/TMM.2021.3134172.
[83] X. Wang, Q. Wang, Y. Zhao, J. Yan, L. Fan, and L. Chen, "Lightweight single-image super-resolution network with attentive auxiliary feature learning," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–17.

DETIAN HUANG received the B.S. degree in electronic information engineering from Xiamen University, Xiamen, China, in 2008, and the Ph.D. degree in circuits and systems from the University of Chinese Academy of Sciences, Beijing, in 2013. He is currently an Assistant Professor with the College of Engineering, Huaqiao University, Quanzhou, China. His research interests include computer vision, image restoration, target tracking, and machine learning.



JIAN CHEN is currently pursuing the M.S. degree with the College of Engineering, Huaqiao University, Quanzhou, China. His research interests include computer vision, image super-resolution, and deep learning.