ABSTRACT Recently, deep convolutional neural networks (CNN) have been widely applied in the single image super-resolution (SISR) task and achieved significant progress in reconstruction performance. However, most of the existing CNN-based SR models are impractical to real-world applicants due to numerous parameters and heavy computation. To tackle this issue, we propose a lightweight attended multi-scale residual network (LAMRN) in this work. Specially, we present an attended multi-scale residual block (AMSRB) to extract multi-scale features, where we embed the efficient channel attention block (ECA) to enhance the discrimination of features. Besides, we introduce a double-attention fusion (DAF) block to fuse the low-level and high-level features efficiently. We use spatial attention and channel attention to obtain guidance from the low-level and high-level features, which is used to guide the feature fusion. Extensive experimental results demonstrate that our LAMRN achieves competitive performance against the state-of-the-art methods with similar parameters and computational operations.

INDEX TERMS Attended multi-scale residual block, double-attention fusion, single image super-resolution.

I. INTRODUCTION
Single image super-resolution (SISR) is a classical computer vision task that aims to reconstruct a high-resolution (HR) image from the low-resolution (LR) one. SISR is an ill-posed problem since multiple HR images may degrade to an identical LR image. It has been studied for decades, widely existing in medical imaging [1], remote sensing [2], surveillance, and some scenarios that required high-frequency details. The SR methods are usually divided into three categories: interpolation-based, reconstruction-based, and learning-based.

Recently, numerous CNN-based SR methods [3–8] have achieved remarkable performance than the traditional methods [9], [10]. EDSR [11], RDN [12], RCAN [13] obtain higher reconstruction accuracy by stacking ResBlock [14], Dense Block [15] or embedding attention mechanism [16]. However, those methods generally achieved excellent performance by expanding the width or increasing the depth of the network, which limited their application in the real-world scenario. Consequently, designing a lightweight network while maintaining comparable performance for SISR becomes a rather tough but promising task.

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

CARN [17] constructed a lightweight network by stacking cascading residual blocks, achieving comparable results with fewer parameters and computational operations. Hui et al. [18] proposed a lightweight network with the information distillation mechanism (IDN) for better recovering the high-quality SR images. Based on IDN, IMDN [19] applied channel separation and advanced channel-wise attention mechanism to improve the network representation. Most of them rarely utilized the multi-scale structures to extract the multi-scale features due to numerous parameters and computational operations. However, multi-scale feature representations could boost the performance by providing more information in various vision tasks [20–23].

On the other hand, the deep feature module in SR models generates features of different levels. The low-level features are rich in spatial details, while the high-level features contain plentiful contextual information. But the previous SR works [11], [13], [17], [19] generally used the output features or concatenated these features as the input of the reconstruction module. Most of them do not fully utilize the difference between these features to guide the fusion.

To extract multi-scale features and efficiently fuse the different features produced in the deep feature extraction, we propose a lightweight attended multi-scale residual network.
Y. Yan et al.: LAMRN for SISR

FIGURE 1. The overall architecture of our proposed lightweight attended multi-scale residual network (LAMRN). The attended multi-scale module (AMSM) is proposed for multi-scale feature extraction and DAF donates our double-attention fusion block.

network (LAMRN) for SISR in this work, as shown in Fig.1. Inspired by Res2Net [24], we utilize an attended multi-scale residual block (AMSRB) to extract multi-scale features. Specially, AMSRB constructs hierarchical residual-like connections within one single residual block to represent multi-scale features. To improve the discrimination of features of each scale, we embed the efficient channel attention block [25] with negligible parameters and computation. Besides, we propose a double-attention fusion block (DAF) to fuse the low-level and high-level features efficiently. Similar to [26], our DAF utilizes spatial attention and channel attention to obtain guidance from the low-level and high-level features, respectively. Then the guidance is used to fuse the low-level features and high-level features efficiently. Experiments on the benchmark datasets demonstrate the effectiveness of our model with similar parameters and computation.

Overall, the main contributions of our method are three-fold:

- We propose a lightweight attended multi-scale residual network for single image super-resolution (LAMRN), which utilizes the attended multi-scale residual block to extract multi-scale information effectively.
- We propose a double-attention fusion block (DAF) to efficiently fuse the low-level and high-level features, further boosting the performance.
- The experimental results on benchmark datasets show that the proposed method achieves competitive performance with the state-of-the-art lightweight SR methods.

The rest of the paper is organized as follows: Section II introduces the related work. Section III elaborates on the structure of the proposed method. Section IV shows our experimental results on benchmark datasets. Finally, Section V draws the conclusions.

II. RELATED WORK

A. CNN-BASED SINGLE IMAGE SUPER-RESOLUTION

As the pioneering work, SRCNN [3] utilized a three-layer convolutional neural network to resolve the interpolated LR images, outperforming the traditional SR methods [9], [10] on both reconstruction accuracy and speed. VDSR [4] increased the network depth to improve performance, which demonstrated the network depth plays an important role on network performance. FSRCNN [27] extracted features from the original LR image and upsampled the features to the desired size by the deconvolutional layer at the tail of the networks, significantly reducing the parameters. Subsequently, some efficient upscale modules [28], [29] were proposed and widely used. EDSR [11] further increased the network depth and won first place in the NTIRE2017 Super-Resolution Challenge [30]. RDN [31] utilized the residual dense block (RDB) to fully reuse the feature maps, which achieved comparable result with EDSR but reduced nearly half parameters. RCAN [13] used channel attention to rescale the feature maps, achieving state-of-the-art performance.

Recently, one research direction of SR is exploring lightweight but competitive architecture, which is more suitable for resource-constrained devices. Ahn et al. [17] proposed a fast, accurate, and lightweight structure (CARN) with cascading residual block. IDN [18] utilized the information distillation mechanism to design their lightweight model. IMDN [19] applied channel separation and advanced channel-wise attention to further improve the network representation. AWSRN [32] made full use of the previous features for reconstruction by the adaptive weighted multi-scale module. Li et al. [33] proposed a flexibly adjustable super-lightweight SR network to improve performance with limited parameters and operations.

B. MULTI-SCALE REPRESENTATIONS

Multi-scale feature representations are of great importance to various vision tasks, such as semantic segmentation [20], object detection [21], and so on. MS-RHDN [23] utilized the atrous-spatial-pyramid-pooling (ASPP) structure to learn the multi-scale features. MRSRN [22] introduced a multi-scale residual network to extract multi-scale features, achieving
competitive performance. Res2Net [24] constructed hierarchical residual-like connections within one single residual block, exploring the multi-scale representation efficiently.

C. ATTENTION MECHANISM

The attention mechanism is proven to be a potential means to enhance the performance of CNN. RCAN [13] incorporated the channel attention mechanism into the residual block to further improve the performance. CBAM [34] utilizes both channel-wise attention and spatial-wise attention to obtain consistent improvements in performance. ECA-Net [25] proposed a local cross-channel interaction strategy without dimensionality reduction to learn channel attention, achieving competitive performance with SENet [16] but fewer parameters.

III. PROPOSED METHOD

This section presents the details of our LAMRN. The overall architecture is first introduced, followed by the detailed descriptions of our proposed attended multi-scale residual block (AMSRB) and double-attention fusion block (DAF).

A. OVERALL ARCHITECTURE

As shown in Fig.1, our model LAMRN mainly consists of four components: shallow feature extraction, deep feature extraction, feature fusion, and reconstruction modules. Let’s donate the input and output images of our LAMRN as $I_{LR}$ and $I_{HR}$, respectively.

Firstly, we use the shallow feature extraction module to extract shallow features $F_{sf}$ from the original LR input:

$$F_{sf} = H_{SF}(I_{LR}).$$  

(1)

where $H_{SF}$ donates the shallow feature extraction module, which contains a $3 \times 3$ convolutional layer.

Then, $F_{sf}$ is passed to the deep feature extraction module, which is stacked by $M$ attended multi-scale modules (AMSM). So we can further obtain several multi-scale features of different levels $F_m$ ($m \in [1, M]$):

$$F_m = H_{AMSM}(H_{AMSM}(\cdots H_{AMSM}(F_{sf}) \cdots)),$$  

(2)

where $H_{AMSM}$ donates the attended multi-scale module. Those features are then passed into the feature fusion module.

The low-level features are rich in spatial information, while the high-level features are full of contextual information. Considering the difference between features of different levels, we use $D$ double-attention fusion (DAF) block(s) to fuse these features efficiently in the feature fusion module. First, the low-level features $F_1$ and high-level features $F_M$ are fed into the first DAF block. Then, the output features of the previous DAF $F_d$ and the features $F_m$ are transferred into the next DAF block:

$$F_d = \begin{cases} H_{DAF}(F_1, F_M), & \text{if } d = 1 \\ H_{DAF}(F_m, F_{d-1}), & \text{if } d \in (1, D) \end{cases}$$  

(3)

where $H_{DAF}$ donates the DAF block.

Next, we add the output features of the final DAF block $F_D$ with the shallow features $F_{sf}$ for feature reuse. Finally, we use the reconstruction module to upscale the feature maps to the desired size and reconstruct the SR images $I_{SR}$:

$$I_{SR} = H_{Rec}(F_D + F_{sf}) = H_{conv}(H_{up}(F_D + F_{sf})).$$  

(4)

where the reconstruction module $H_{Rec}$ consists of an upsampling module $H_{up}$ and a convolutional layer $H_{conv}$. Following [13], [17], [19], our upsampling module contains a convolutional layer to increase the number of the channels of feature maps and a sub-pixel layer to rearrange the feature maps to the desired size. But we use $1 \times 1$ convolutional layer instead of $3 \times 3$ convolutional layer to reduce the model parameters and computation.

Recently, several loss functions have been investigated to train the CNN-based SR models, such as $L_1$, $L_2$, perceptual and adversarial losses. Following the previous works [11], [12], [19], [32], we optimize our model with $L_1$ loss function. Given a training set $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$, which contains $N$ LR input images and the corresponding HR counterparts, the loss function is defined as:

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \|H_{AMRN}(I_{LR}^i) - I_{HR}^i\|_1,$$  

(5)

where $\Theta$ donates the parameter set of our model.

B. ATTENDED MULTI-SCALE MODULE

As shown in Fig.2, our attended multi-scale module (AMSM) mainly consists of $B$ attended multi-scale residual blocks (AMSRB) and $1 \times 1$ convolutional layer.

As is well known, multi-scale representations of CNN [6], [22], [23] could provide more information and boost the performance. However, most multi-scale structures are constructed with several convolutional layers with different kernel sizes, bringing numerous parameters and computation on account of the large kernel sizes.

Inspired by Res2Net [24], we propose an attended multi-scale residual block (AMSRB) to extract multi-scale features, which does not increase the heavy burden of model parameters and computation. As shown in Fig.2, AMSRB contains several convolutional blocks (CB), two $1 \times 1$ convolutional layers, and a skip connection.

Firstly, the input feature maps of the $b$-th AMSRB are fed into a $1 \times 1$ convolutional layer to increase the channel numbers. Then the features are split into $s$ feature map subsets along the channel axis, donated by $X_i (i \in \{1, 2, \ldots, s\})$. Each feature subset $X_i$ has the same spatial size as the input feature maps. Except for $X_1$, each $X_i$ is passed to a convolutional block $H_{CB}$ for feature extraction, which contains a convolutional layer, a ReLU, and an efficient channel attention (ECA) block. We donate the output features of the $i$-th CB as $Y_i$. The feature subset $X_1$ is added with $Y_{i-1}$ and then passed to the next $H_{CB}$ for extracting features with different receptive fields. Following [24], we omit the convolutional block for $X_1$ to reduce parameters and computational operations, which
The output features from BAMS and then compress the channel number with the $1 \times 1$ convolutional layer, as shown below features channel number reduction. So we obtain the fused multi-scale all of them and feed them into a $1 \times 1$ wise multiplication.

where

$H_y = H_{ECA}(\delta(H_x(Y_i)))$, $\delta$ is the function of ReLU, $H_x$ donates the $3 \times 3$ convolutional layer with kernel size $k$ is utilized on $z$. In other words, the weight of $z_i$ is calculated only by considering the interaction between itself and the $k$ neighbors. Finally, the output features $Y_i$ are obtained by rescaling the input features $Y$ with the normalized weights:

where $H_{C1D}$ donates the 1D convolutional layer, $H_{GAP}$ represents global average pooling, $\otimes$ means element-wise multiplication.

After extracting the multi-scale features $Y_i$, we concatenate all of them and feed them into a $1 \times 1$ convolutional layer for channel number reduction. So we obtain the fused multi-scale features $F_{AMSRB}^b$ from the $b$-th AMSRB:

To fully use the intermediate features, we concatenate all the output features from $B$ AMSRBs at the end of the $m$-th AMSM and then compress the channel number with the $1 \times 1$ convolutional layer, as shown below:

$F_m = H_{conv}(Concat([F_{AMSRB}^1, \ldots, F_{AMSRB}^b]), (9)$

where Concat donates the concatenation operation, and $H_{conv}$ is the $1 \times 1$ convolutional layer.

**C. DOUBLE-ATTENTION FUSION BLOCK**

Generally speaking, the spatial expression ability of the network decreases while the semantic expression ability increases as the network get deeper. The previous works only used the output features of the deep feature extraction [11], [13], [35] or simply concatenated features at different levels and then passed to a $1 \times 1$ convolutional layer [19], [22]. Inspired by [26], we use a double-attention fusion block (DAF) to fuse the features from different levels efficiently.

As shown in Fig.4, our DAF is composed of two attention blocks: channel attention (CA) block and spatial attention (SA) block. Taking the first DAF as an example, the low-level features $F_1$ are pass through the SA block to obtain the spatial attention $Z_{SA} \in \mathbb{R}^{H \times W \times 1}$. Specially, we first use two FC layers to generate the spatial statistics along the channel axis and then use the sigmoid function to normalize the statistics. While the high-level features $F_M$ are fed into the CA block to produce the channel attention $Z_{CA} \in \mathbb{R}^{1 \times 1 \times C}$. Firstly, a global average pooling is used for generating channel statistics. Then, two FC layers are adopted for capturing the cross-channel interaction and a sigmoid function for normalization.

The low-level features contain more spatial information, while the high-level features involve rich semantic context.
We multiply $Z_{SA}$ with the high-level features, refining the boundaries of the high-level features. $Z_{CA}$ is multiplied by the low-level features, providing contextual information for the low-level features. Finally, the rescaled low-level and high-level features are fused by an element-wise addition. Hence, the fused features pay attention to both spatial details and semantic context, further improving the network performance.

$$Z_{SA} = \sigma(H_{\text{conv}}(\delta(H_{\text{conv}}(F_1)))),$$

$$Z_{CA} = \sigma(H_{\text{conv}}(\delta(H_{\text{GAP}}(F_M)))),$$

$$F_{1f} = H_{\text{DAF}}(F_1, F_M) = F_1 \otimes Z_{CA} + F_M \otimes Z_{SA}. \quad (10)$$

IV. EXPERIMENTS

In this section, we will first introduce the datasets and evaluation metrics. Then, we illustrate the detail of the experimental settings. Finally, we present the results and ablation study to demonstrate the effectiveness of our proposed method.

A. DATASETS AND EVALUATION METRICS

DIV2K dataset is published in NTIRE 2017 Challenge on single image super-resolution [30], which contains 800 training, 100 evaluation, and 100 testing 2K images. Following [13], [17], [18], [32], we use the training set of DIV2K to train our models. The low-resolution images are obtained by bicubic downsampling from the high-resolution images. The commonly used evaluation metrics PSNR and SSIM are utilized for quantitative comparisons with other SR methods. The visual results are also provided for more intuitive comparisons with other methods.

B. IMPLEMENTATION SETTINGS

We randomly crop the LR image into patches with a size of $48 \times 48$ as input for each mini-batch training. The corresponding patch size of HR images is $48r \times 48r$, where $r$ is the scale factor. The batch size is set as 16. Data augmentation is performed on the training set, such as random rotations of $90^\circ$, $180^\circ$, $270^\circ$, and horizontal flip. For the setting of hyper-parameters, we set the number of AMSM as $M = 3$, the number of AMSRB as $B = 3$, the number of DAF block as $D = 2$, and the kernel size of the 1D convolutional layer in ECA as $k = 3$. The kernel size of convolutional layers in CB of AMSRB is $3 \times 3$, while the others are set as $1 \times 1$. Our models are trained by the ADAM optimizer [36] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is set to $2 \times 10^{-4}$ and then decreases to half every $2 \times 10^5$ iterations of back-propagation. We first train a model for $4 \times$ and then use it as the pre-trained model to train models of other scale factors. We implement our models using PyTorch [37] framework with an NVIDIA TITAN Xp GPU.

C. MODEL ANALYSIS

The channel number of the intermediate feature maps is an important factor for calculating the model parameters and computational operations. Following [17], [32], we use Multi-Adds to represent the number of computational operations, which is calculated on a SR image with a size of $1280 \times 720$. To obtain models with different complexity, we train them with different channel numbers: 16, 32, 64, which are donated as LAMRN-16, LAMRN-32, and LAMRN, respectively. For scale $4 \times$, the total parameters range from 89K to 1407K, and Multi-Adds ranges from 5.6G to 85G.

Fig.5 depicts the trade-off between parameter/Multi-Adds and performance for scale $4 \times$. Our LAMRN obtains the higher PSNR on Set5 with fewer parameters and computation than AWRSN [32]. Compared with SRCNN [3] and FSRCNN [27], the smallest model (LAMRN-16) achieves better performance with similar size and computation. The middle model (LAMRN-32) obtains a better trade-off than SRMDNF [31] and IDN [18]. These results show that our
models can achieve a better trade-off between the performance and model size.

We also use a different number ($B$) of AMSRB in AMSM to change the depth of LAMRN. The performance of different models and the corresponding parameters and Multi-Adds for scale ×4 are shown in Table 1. When $B$ increases by 1, the parameter and Multi-Adds increase 444K and 25G, respectively. The performance improves by about 0.1 dB when $B$ is less than 4, but the increment becomes less for the model with $B = 5$ and $B = 6$. Note that the model with $B = 6$ achieves comparable results with EDSR [11] (32.46 dB/43M) with only 2.7M parameters, demonstrating that our LAMRN has strong learning ability with considerably fewer parameters and computational operations. To compare with the state-of-the-art lightweight SR models, we use $B = 3$ as the final setting in this work.

**TABLE 1. Results of LAMRN with different number of AMSRB ($B$) in AMSM for ×4 scale factor on Set5.**

| $B$ | Params. | Multi-Adds | PSNR |
|-----|---------|------------|------|
| 1   | 519K    | 34G        | 31.94|
| 2   | 963K    | 59.2G      | 32.14|
| 3   | 1407K   | 85G        | 32.27|
| 4   | 1851K   | 110.4G     | 32.35|
| 5   | 2295K   | 135.9G     | 32.40|
| 6   | 2739K   | 161.4G     | 32.43|

**D. COMPARISON WITH STATE-OF-THE-ART METHODS**

We compare the proposed LAMRN with the state-of-the-art lightweight SR methods for ×2, ×3, ×4, ×8 scales, including SRCNN [3], [38], FSRCNN [27], VDSR [4], DRCN [6], LapSRN [5], EDSR-baseline [11], SRMDNF [31], IDN [18], CARN [17], AWSRN [32] and MSRN [22]. Following [3], [17], [22], [38], [39], PSNR and SSIM are calculated on the Y channel of the YCrCb color space.

Table 2 shows the quantitative results on five benchmark datasets for different SR algorithms. We also present the parameters and Multi-Adds of these models for more intuitive contrast. As shown in Table 2, our LAMRN-16 outperforms some CNN-based SR methods with a few parameters and computational operations. And the model LAMRN-32, which has fewer parameters and operations than AWSRN-S [32] and IDN [18], achieves better results on ×2, ×3, and ×4. Compared with the state-of-the-art SR methods, our LAMRN achieves competitive performance with fewer parameters and Multi-Adds. As the magnification becomes larger, the degraded LR images lose more high-frequency information, which is hard to recover high-quality SR images. Our LAMRN model achieves better results with limited model size and operations owning to the effectiveness of the proposed attended multi-scale residual block and double-attention fusion block.

Fig.6 and Fig.7 present the visual comparisons for each scale. We select some detail patches from different SR images for better comparisons. For image ‘img_004’, our LAMRN generates more correct textures, while other methods generate distorted textures. For images ‘DollGUN’ and ‘BokuHaSitatakaKun’, our LAMRN-16 achieves comparable performance with IDN [18], and LAMRN produces clearer textures than other lightweight SR methods. Our LAMRN generates clearer textures than other models for image ‘PlatinumJungle’. And for images ‘img_093’ and ‘zebra’, the proposed method produces more correct textures than others, which look closer to the HR patches. These results demonstrate the effectiveness of our LAMRN.

**E. ABLATION STUDY**

We conduct several ablation studies to explore the effectiveness of our proposed method.

To evaluate the effectiveness of the efficient channel attention mechanism in AMSRB, we conduct three ablation studies. Specially, we remove the ECA block in each CB of AMSRB (no ECA), use only one ECA block after fusing multi-scale features from different CB (1ECA), and substitute it with the channel attention mechanism (CA), respectively.
TABLE 2. Quantitative results with BI degradation model. Best and second best results are highlighted and underlined respectively.

| Method | Scale | Param. | Multi-Adds | Set5 | Set14 | B100 | Urban100 | Manga109 |
|--------|-------|--------|------------|------|-------|------|----------|----------|
|        |       | PSNR   | SSIM       | PSNR | SSIM  | PSNR | SSIM     | PSNR     | SSIM     |
| SRCNN [3] | ×2    | 57K    | 52.7G      | 36.66 | 0.9542 | 32.45 | 0.9067  | 31.36    | 0.8879   | 29.50    | 0.8946  | 35.60 | 0.9663 |
| PSRN [27] | ×2    | 12K    | 6.6G       | 37.00 | 0.9588 | 32.63 | 0.9058  | 31.53    | 0.9072   | 29.88    | 0.9020  | 36.17 | 0.9710 |
| VDSR [4] | ×4    | 665K   | 61.2G      | 37.55 | 0.9587 | 33.03 | 0.9124  | 31.90    | 0.8960   | 30.76    | 0.9140  | 37.22 | 0.9750 |
| DRRN [6] | ×4    | 1774K  | 1977.4G    | 37.63 | 0.9588 | 33.04 | 0.9118  | 31.85    | 0.8942   | 30.75    | 0.9133  | 37.63 | 0.9723 |
| LapSRN [5] | ×8   | 813K   | 29.9G      | 37.52 | 0.9590 | 33.08 | 0.9130  | 31.08    | 0.8950   | 30.41    | 0.9100  | 37.27 | 0.9740 |
| EDSR-baseline [11] | ×8 | 1370K  | 316.9G    | 37.59 | 0.9592 | 33.15 | 0.9132  | 31.26    | 0.8952   | 30.61    | 0.9132  | 37.54 | 0.9759 |
| SRCNN [31] | ×32 | 347T    | 37.7G      | 37.99 | 0.9601 | 33.32 | 0.9159  | 32.08    | 0.8953   | 31.33    | 0.9204  | 38.07 | 0.9761 |
| SDRMF [18] | ×32  | 590K   | 174.1G     | 37.83 | 0.9600 | 33.30 | 0.9148  | 32.08    | 0.8985   | 31.27    | 0.9196  | 38.01 | 0.9749 |
| IDN [17] | ×32   | 599K   | 222.8G     | 37.76 | 0.9590 | 33.52 | 0.9166  | 32.09    | 0.8978   | 31.92    | 0.9256  | 36.36 | 0.9765 |
| CARN [17] | ×32   | 397K   | 91.2G      | 37.75 | 0.9590 | 33.51 | 0.9161  | 32.00    | 0.8964   | 31.79    | 0.9200  | 37.90 | 0.9752 |
| AWRSM-N [32] | ×32 | 1063K  | 246.1G     | 38.04 | 0.9605 | 33.66 | 0.9181  | 32.60    | 0.8995   | 32.02    | 0.9223  | 38.66 | 0.9762 |
| AWRSM-N [32] | ×32 | 1397K  | 320.5G     | 37.81 | 0.9608 | 33.78 | 0.9189  | 32.26    | 0.9006   | 32.49    | 0.9316  | 38.87 | 0.9775 |
| MSRN [22] | ×32  | 5930K  | 1365.4G    | 38.08 | 0.9607 | 33.70 | 0.9186  | 32.23    | 0.9002   | 32.29    | 0.9303  | 38.69 | 0.9772 |

| TABLE 3. | Ablation study of different channel mechanisms in AMSRB for ×4 scale factor on Urban100. Each model is trained for 200 epochs.

| Model | no ECA | 1 ECA | CA | LAMRN |
|-------|--------|-------|----|-------|
| Params. | 1407K | 1407K | 1412K | 1407K |
| Multi-Adds | 85G | 85G | 85G | 85G |
| PSNR | 26.03 | 26.16 | 26.08 | 26.23 |

| TABLE 4. | Ablation study of different fusion in the feature fusion module for ×4 scale factor on Urban100. Each model is trained for 200 epochs.

As shown in Table 3 and the left figure of Fig. 8, embedding an ECA block in AMSRB improves the model without ECA by 0.13 dB on Urban100, which keeps the same model size and operations. While embedding ECA in each CB of AMSRB further brings performance improvement, donating the efficient channel attention is effective in our model. What’s more, our LAMRN outperforms the model LAMRN-CA by 0.15 dB with fewer parameters, demonstrating efficient performance.

52208 VOLUME 9, 2021
channel attention is better than traditional channel attention with fewer parameters in our model.

We also explore the effectiveness of our proposed double-attention fusion block. We use the final feature maps to reconstruct SR image (w/o Fus), concatenate features from all AMSMs and then pass to a $1 \times 1$ convolutional layer for channel compression ($1 \times 1$ Fus), remove channel attention (w/o CA), and remove spatial attention (w/o SA), respectively. As shown in Table 4 and the right figure in Fig.8, our model outperforms the model w/o Fus, $1 \times 1$ Fus, w/o CA, and w/o SA by 0.31 dB, 0.20 dB, 0.07 dB, and 0.17 dB on Urban100, respectively. Note that our model has fewer parameters and operations than the model with $1 \times 1$ fusion, demonstrating the effectiveness of our DAF block.

To explain that our DAF focuses on both spatial information and contextual information more intuitively, we present...
FIGURE 7. Visual comparison for $\times 8$ SR on Manga109 and Set14. The best result is highlighted.

FIGURE 8. The training convergence curves of ablation investigations, where each model was trained for 200 epochs. The left one depicts the performance of different attention mechanisms in the AMSRB on Set5 for $\times 4$ scale factor. The right one shows the effect of different fusions in the fusion block on Set5 for $\times 4$ scale factor.

FIGURE 9. Feature visualization of the inputs and output features of the first DAF block for $\times 4$ scale factor.

the feature visualization of the input and output feature maps of the first DAF block. As depicted in Fig.9, the low-level input feature maps contain more spatial information, such as the edge information, while the high-level input feature maps emphasize contextual information. The output features contain both spatial information and semantic context, demonstrating that our DAF block fuses the features from different levels efficiently.
V. CONCLUSION

We propose a lightweight attended multi-scale residual network for single image super-resolution in this paper. Specifically, we use an attended multi-scale residual block as the basic block to extract multi-scale features, where we further embed the efficient channel attention block to improve the network representations. In such a way, the discriminative multi-scale features are extracted at a granular level efficiently. What’s more, we utilize the double-attention module to fuse the features at different levels from the deep feature extraction module efficiently. Extensive experimental results demonstrate the effectiveness of our model with similar parameters and computation.

REFERENCES

[1] A. S. Chaudhari, Z. Fang, F. Kogan, J. Wood, K. J. Stevens, E. K. Gibbons, J. H. Lee, G. E. Gold, and B. A. Hargreaves, “Super-resolution musculoskeletal MRI using deep learning,” Magn. Reson. Med., vol. 80, no. 5, pp. 2139–2154, Nov. 2018.
[2] J. M. Haut, R. Fernandez-Beltran, M. E. Paolletti, J. Plaza, A. Plaza, and F. Pla, “A new deep generative network for unsupervised remote sensing single-image super-resolution,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 11, pp. 6792–6810, Nov. 2018.
[3] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. 13th Eur. Conf. Comput. Vis. Berlin, Germany: Springer, Sep. 2014, pp. 184–199.
[4] 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.
[5] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep Laplacian pyramids for fast and accurate super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 624–632.
[6] 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.
[7] 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. 3147–3155.
[8] 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. 4539–4547.
[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[10] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in Proc. Asian Conf. Comput. Vis. (ACCV). Berlin, Germany: Springer, 2014, pp. 111–126.
[11] 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. 136–144.
[12] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2472–2481.
[13] 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.
[14] 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.
[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[16] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[17] N. Ahn, B. Kang, and K.-A. Sohn, “Fast, accurate, and lightweight super-resolution with cascading residual network,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 252–268.
[18] 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.
[19] 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.
[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.
[21] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.
[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), Sep. 2018, pp. 517–532.
[23] C. Liu, X. Sun, C. Chen, P. L. Rosin, Y. Yan, L. Jin, and X. Peng, “Multi-scale residual hierarchical dense networks for single image super-resolution,” IEEE Access, vol. 7, pp. 60572–60583, 2019.
[24] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2Net: A new multi-scale backbone architecture,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 652–662, Feb. 2021.
[25] 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. 11534–11542.
[26] M. Liu and H. Yin, “Feature pyramid encoding network for real-time semantic segmentation,” 2019, arXiv:1909.08599. [Online]. Available: http://arxiv.org/abs/1909.08599.
[27] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proc. Eur. Conf. Comput. Vis. (ECCV). Berlin, Germany: Springer, Oct. 2016, pp. 391–407.
[28] 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.
[29] J.-H. Kim and J.-S. Lee, “Deep residual network with enhanced upscaling module for super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 800–808.
[30] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, “NTIRE 2017 challenge on single image super-resolution: Methods and results,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 114–125.
[31] K. Zhang, W. Zuo, and L. Zhang, “Learning a single convolutional super-resolution network for multiple degradations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3262–3271.
[32] C. Wang, Z. Li, and J. Shi, “Lightweight image super-resolution with adaptive weighted learning network,” 2019, arXiv:1904.02358. [Online]. Available: http://arxiv.org/abs/1904.02358.
[33] B. Li, B. Wang, J. Liu, Z. Qi, and Y. Shi, “S-LWSR: Super lightweight super-resolution network,” IEEE Trans. Image Process., vol. 29, pp. 8368–8380, 2020.
[34] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 3–19.
[35] 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. 4681–4690.
[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.
[37] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8024–8035.

[38] 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.

[39] K. Zhang, W. Zuo, S. Gu, and L. Zhang, “Learning deep CNN denoiser prior for image restoration,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3929–3938.

YITONG YAN received the B.S. degree in software engineering from the South China University of Technology, in 2018, where she is currently pursuing the master’s degree in software engineering. Her current research interests include single image super-resolution and deep learning.

XUE XU received the B.S. degree in software engineering from the South China University of Technology, China, in 2018, where she is currently pursuing the M.S. degree in software engineering. Her current research interests include unsupervised single image super-resolution and deep learning.

WENHUI CHEN received the B.S. degree in software engineering from the South China University of Technology, in 2019. She is currently pursuing the M.S. degree in software engineering. Her current research interests include single image super-resolution, generative adversarial networks, and video super-resolution.

XINYI PENG is currently a Full Professor with the School of Software Engineering, South China University of Technology, China. In recent years, he has presided over and participated in more than 20 projects, with the project funds of over 9 million. Among them, nine projects above the provincial level have passed the acceptance inspection. His research interests include artificial intelligence and data mining.

* * *