Multi-Resolution Space-Attended Residual Dense Network for Single Image Super-Resolution

JIAYV QIN\textsuperscript{1}, XIANFANG SUN\textsuperscript{2}, YITONG YAN\textsuperscript{1}, LONGCUN JIN\textsuperscript{1}, (Member, IEEE), AND XINYI PENG\textsuperscript{1}

\textsuperscript{1}School of Software Engineering, South China University of Technology, Guangzhou 510006, China
\textsuperscript{2}School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.

Corresponding author: Longcun Jin (lcjin@scut.edu.cn)

This work was supported in part by the Pearl River Technology Nova Project under Grant 201710010020, in part by the Fundamental Research Funds for the Central Universities under Grant 2019MS086, and in part by the National Science Foundation of China under Grant 61300135.

ABSTRACT With the help of deep convolutional neural networks, a vast majority of single image super-resolution (SISR) methods have been developed, and achieved promising performance. However, these methods suffer from over-smoothness in textured regions due to utilizing a single-resolution network to reconstruct both the low-frequency and high-frequency information simultaneously. To overcome this problem, we propose a Multi-resolution space-Attended Residual Dense Network (MARDN) to separate low-frequency and high-frequency information for reconstructing high-quality super-resolved images. Specifically, we start from a low-resolution sub-network, and add low-to-high resolution sub-networks step by step in several stages. These sub-networks with different depth and resolution are utilized to produce feature maps of different frequencies in parallel. For instance, the high-resolution sub-network with fewer stages is applied to local high-frequency textured information extraction, while the low-resolution one with more stages is devoted to generating global low-frequency information. Furthermore, the fusion block with channel-wise sub-network attention is proposed for adaptively fusing the feature maps from different sub-networks instead of applying concatenation and $1 \times 1$ convolution. A series of ablation investigations and model analyses validate the effectiveness and efficiency of our MARDN. Extensive experiments on benchmark datasets demonstrate the superiority of the proposed MARDN against the state-of-the-art methods. Our super-resolution results and the source code can be downloaded from https://github.com/Periter/MARDN.

INDEX TERMS Channel-wise sub-network attention, convolutional neural networks, multi-resolution sub-networks, single image super-resolution.

I. INTRODUCTION

Single image super-resolution (SISR) aims to estimate a high-resolution counterpart from a single low-resolution image and has been researched for decades in both industry and academia. So far, SISR has a myriad of applications in various fields, such as remote sensing [1], medical imaging [2], and surveillance imaging [3]. However, SISR is a notoriously challenging ill-posed problem because a particular low-resolution image may correspond to a crop of plausible high-resolution counterparts. To resolve this undetermined problem, plenty of SISR methods e.g., [4]–[17] have been proposed, among which the Convolutional Neural Network (CNN) based ones achieve outstanding performance, benefiting from their powerful representational capacity.

Recently, several CNN-based methods explore different network architectures for SISR and have made great progress in reconstruction accuracy. EDSR [11] stacks large numbers of residual blocks and removes batch normalization layers. RDN [13] uses residual dense blocks to extract hierarchical feature maps. RCAN [8] further exploits channel attention mechanism and residual in residual blocks. CNF [18] constructs different depth parallel context-wise network fusion framework. However, these methods suffer from varying degrees of over-smoothness in textured regions due to utilizing a single-resolution network to reconstruct both the low-frequency and high-frequency information. In human pose estimation, Sun et al. [19] demonstrated the

VOLUME 8, 2020 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/
effectiveness of HRNet, which contains multi-resolution sub-networks to extract different resolution representations and achieves the state-of-the-art results. Inspired by Sun et al.’s work, we propose a novel multi-resolution network to reconstruct low-frequency and high-frequency information, separately.

For SISR, the low-frequency and high-frequency information of the super-resolution (SR) images are not only affected by the depth of the network, but also by the resolution of the features. To the best of our knowledge, deeper networks have larger receptive fields to extract global low-frequency semantic information, while shallower networks have smaller receptive fields which focus on local regions containing more high-frequency information. As far as the resolution of features is concerned, high-resolution features contain more pixels and high-frequency information than the low-resolution ones. In other words, the high-resolution features can better reflect the regions where the frequencies or signals change rapidly.

The key challenge of SISR is that low-frequency and high-frequency information can not be reconstructed well at the same time and thus makes SISR prone to failure. To tackle this problem, we propose a novel SISR method based on Multi-resolution space-Attended Residual Dense Network (MARDN) to extract different frequency representations in parallel for reconstructing the final super-resolved images. Specifically, we start from a low-resolution sub-network, gradually add low-to-high resolution sub-networks in several stages. In our model, the high-resolution sub-network with few stages is used to extract the local high-frequency representations, while the low-resolution sub-network with more stages is applied to extract the global low-frequency representations. Additionally, we propose a channel-wise sub-network attention block to adaptively fuse the features from different sub-networks. To accelerate the end-to-end training, we use a Global Skip Connection (GSC) between the shallow features and the deep features, a Local Skip Connections (LSC) in the space-Attended Residual Dense Blocks (ARDB), and a Direct Connection (DC) from the shallow features to the sub-network whose features have the same resolution with the original input image. As for our basic block, the ARDB, we apply residual connection, dense connection and an attention module to enhance feature representation ability. As shown in Fig. 1, with the help of channel attention, RCAN [8] can recover part of the lines in the circled region but still produces severely blurred texture. While our MARDN recovers all the lines rightly and is closer to the HR image, benefiting from the design of multi-resolution sub-networks.

Overall, the contributions of our method can be summarized as follows:

- We design a novel multi-resolution space-attended residual dense network for effective image super-resolution. Specifically, different sub-networks extract different frequency information by controlling their resolution and depth.

- We propose an adaptive fusion block based on channel-wise sub-network attention for better fusing the feature maps from different sub-networks.

- Extensive experiments on benchmark datasets demonstrate the effectiveness of our model and the results are superior to the state-of-the-art super-resolution methods.

The rest of the paper is organized as follows: Section II introduces the related work. Section III elaborates the structure of the proposed network. Section IV discusses the difference between our model and similar works. Section V shows our experimental results on benchmark datasets, including quantitative and visual comparisons with other methods as well as the effectiveness of each components in our network. Finally, Section VI draws the conclusions.

II. RELATED WORK

Recently, deep convolutional neural networks (CNNs) have brought remarkable improvements in computer vision owing to their powerful learning ability. This section first reviews several CNN-based super-resolution (SR) methods, and then briefly describes HRNet [19], a method using multi-resolution parallel sub-networks in human pose estimation, which inspired our method.

A. CNN-BASED IMAGE SUPER-RESOLUTION

Dong et al. [4] first proposed a CNN model (SRCNN) to learn the end-to-end mapping between interpolated low-resolution (LR) and high-resolution (HR) image patches, which outperforms the traditional SR methods in restoration quality and speed. VDSR [5] and DRCN [7] increase the network depth and introduce residual blocks to alleviate the training difficulty, demonstrating that the network depth has an important impact on the reconstruction performance. Using the interpolated LR image as the network input results in redundancy computation, particularly when the networks are very deep. Instead, some works take the original LR image as input, then extract feature maps from the LR image, and finally use different upscaling modules to rescale the feature maps to the desired size, such as deconvolutional layer [21], sub-pixel layer [22], [23], and EUSR [24]. The LapSRN method [10] progressively reconstructs the sub-band residuals of the high-resolution images at multiple pyramid levels and uses the
robust Charbonnier loss functions. The authors also adopted the multi-scale training strategy to train a single model for handling multiple upsampling scales (MSLapSRN) [25]. The EDSR method [11] removes the batch normalization layers in the ResNet [26] and constructs a very deep network, which has won the first place in the NTIRE 2017 Super-Resolution Challenge [27]. Haris et al. [28] utilized mutually-connected up- and down-sampling stages with an error feedback mechanism for reconstructing SR images. Zhang et al. [13] proposed residual dense block (RDB) as building module to set up the SR Network (RDN), which could generate comparable or even better results than that of EDSR, but reduced nearly half of parameters. Zhang et al. [8] also proposed RCAN with residual-in-residual structure and channel attention mechanism to produce appealing SR results. Furthermore, Dai et al. [29] presented a second-order attention network (SAN) with a non-locally enhanced residual group structure, which achieved the state-of-the-art performance with less parameters.

The aforementioned methods aimed to minimize the mean square error (MSE) for higher PSNR value, but sometimes produced over-smoothed edges. To produce photo-realistic SR images, SRGAN [6] first uses the generative adversarial networks (GAN) [30] for SISR. Wang et al. [15] proposed enhanced GAN-based method (ESRGAN) with Residual-in-Residual Dense Block (RRDB) and used RaGAN [31] as the discriminator to produce more photo-realistic images. In addition, ESRGAN has won the first place in the 2018 PIRM Challenge on Perceptual Super-Resolution [32]. Recently, Zhang et al. [33] proposed Super-Resolution Generative Adversarial Networks with Ranker (RankSRGAN) to optimize the generator in the direction of perceptual metrics, which achieved visually pleasing results and reached state-of-the-art performance in perceptual metrics. Zhou and Susstrunk [34] proposed KMSR to build the realistic blur kernel pool with GAN, and used it to construct real photograph paired dataset for unsupervised training. Lugmayr et al. [35] also presented an unsupervised SR model through CycleGAN, which can transfer the LR image from bicubic degradation domain to real-world degradation domain, in this way, real world LR-HR paired training and testing datasets are formed.

The above-mentioned methods focus on increasing the network depth or designing efficient network structure to extract more effective features for improving the network reconstruction performance either in accuracy or perceptual quality. However, they do not take multi-resolution features in consideration and use only a single network to extract the same resolution information. Our proposed method pays attention to different frequency representations and uses a multi-resolution model with several sub-networks to extract different frequency information in parallel.

### B. HIGH-RESOLUTION NETWORK

Most existing methods for pose estimation usually recovered high-resolution representations from low-resolution representations produced by a high-to-low resolution network. Similar to hierarchical neural networks [36], Sun et al. proposed the HRNet [19] to maintain high-resolution representations during the whole process, which achieved superior pose estimation results over two benchmark datasets: the COCO keypoint detection dataset [37] and the MPII Human Pose dataset [38].

Most networks had the high-to-low and low-to-high processes, where the former process aimed to produce low resolutions and high channel counts and the latter was designed to produce high-resolution representations and reduce the resolution of the feature maps. Compared with these networks, HRNet had two benefits: i) It formed a high-resolution sub-network as the first stage and gradually added high-to-low resolution sub-networks to form more stages. Then it connected high-to-low resolution sub-networks in parallel rather than in series as most networks did. In this way, HRNet can maintain the high-resolution through the whole process for spatially precise heatmap estimation. ii) It conducted multi-scale fusions across parallel sub-networks such that each sub-network repeatedly received the information from other sub-networks. Consequently, the high-resolution representation was boosted and the keypoint heatmap was potentially more accurate and spatially more precise, resulting in better performance compared with other pose estimation networks.

### III. PROPOSED METHOD

This section introduces our MARDN model. The overall architecture is presented firstly, followed by detailed description of the space-Attended Residual Dense Block (ARDB) and the fusion block.

#### A. ARCHITECTURE

As shown in Fig.2, MARDN reconstructs a super-resolution image \( I_{SR} \) from its low-resolution counterpart \( I_{LR} \). The MARDN consists of four modules: shallow feature extraction, deep feature extraction, upsampling, and reconstruction modules. And the pseudo-code is shown in Algorithm 1.

Firstly, the original low-resolution images are fed into the shallow feature extraction module, which transfers those images from the color space to the feature space, resulting in the shallow features \( F_{sf} \).

\[
F_{sf} = H_{SF}(I_{LR}),
\]

where \( H_{SF} \) represents the function of the shallow feature extraction module, which is constructed by a \( 3 \times 3 \) convolutional layer.

The shallow feature \( F_{sf} \) then passes through the deep feature extraction module, and produces the deep feature maps \( F_{df} \),

\[
F_{df} = H_{DF}(F_{sf}),
\]

where \( H_{DF} \) denotes our multi-resolution sub-networks, which is inspired by HRNet [19].
Algorithm 1: Multi-Resolution Space-Attended Residual Dense Network

// Training Phase

Require: Training set \([I_{LR}, I_{HR}^{r=1}]\).

Require: Learning rate \(\eta\), mini-batch size \(m\) and iteration epoch \(E\).

Ensure: Parameter \(\Theta\).
1: Initialize parameter \(\Theta\)
2: for \(e = 1, \ldots, E\) do
3: \(n \leftarrow 1\)
4: while \(n \leq N - m\) do
5: \([F_{sf}]_{n+m} \leftarrow H_{SF}(I_{LR}^{n+m})\)
6: \([F_{df}]_{n+m} \leftarrow H_{DF}(F_{sf}^{r=1})\)
7: \([F_{up}]_{n+m} \leftarrow H_{UP}(F_{df}^{r+1}) + (F_{sf}^{r=1})_{n+m}\)
8: \([I_{SR}]_{n+m} \leftarrow H_{REC}(F_{up}^{r+1})\)
9: \(L_\Theta \leftarrow \frac{1}{m} \sum_{i=n+m} \|I_{SR}^{r=1} - I_{HR}^{r=1}\|_p\)
10: Update parameter with ADAM optimizer:
\(\Theta \leftarrow \text{Optimizer}(L_\Theta, \eta)\)
11: \(n \leftarrow n + m\)
12: end while
13: end for

// Testing Phase

Require: Image \(I_{LR}\).

Require: Parameter \(\Theta\).

Ensure: Super-resolved image \(I_{SR}\).
1: Loading MARDN model with parameter \(\Theta\)
2: \(F_{sf} = H_{SF}(I_{LR})\)
3: \(F_{df} = H_{DF}(F_{sf})\)
4: \(F_{up} = H_{UP}(F_{df} + F_{sf})\)
5: \(I_{SR} = H_{REC}(F_{up})\)

Different from existing SISR methods, which use single channel networks to extract deep features and reconstruct SR images, in order to extract features of different frequencies effectively, we adopt multi-channel sub-networks to extract features. We first downsample the shallow features into low-resolution features through bi-linear interpolation. Taking the low-resolution features as input, the deep feature extraction module starts from a low-resolution sub-network as the first stage, and then adds low-to-high resolution sub-networks gradually in several stages, where each sub-network consists of several ARDBs. The features from different resolution sub-networks join together at the start of each stage via our specifically designed fusion blocks, which apply channel-wise sub-network attention blocks for adaptive fusion. Each of the low-to-high resolution representations adaptively receives information from other parallel representations over and over.

Considering the difference in complexity between the low-frequency and the high-frequency information, we extract different frequency information by controlling the depth and resolution of the sub-networks. To make the low-resolution sub-network extract the global low-frequency information with large receptive field, the data flow in this sub-network passes through more stages and more ARDBs than that in the high-resolution one. \(N_{sf}\) is the sub-network with \(r\)-resolution in the \(s\)-th stage, whose resolution is \(2^{s-1}\) times of the first sub-network. As shown in Fig.2, the deep feature extraction module \(H_{DF}\) contains 3 different resolution sub-networks in parallel, and can be represented as:

\[
F_{sf} \xrightarrow{N_{11}} [N_{21}] \xrightarrow{N_{31}} [N_{32}] \xrightarrow{N_{33}} F_{up} \uparrow \uparrow \uparrow \uparrow \uparrow (3)
\]

After obtaining deep features \(F_{df}\) from the deep feature extraction module, we use the upscaling module to produce the features \(F_{up}\) with the same size as the desired super-resolved images, expressed as:

\[
F_{up} = H_{UP}(F_{df} + F_{sf} \uparrow), (4)
\]
where \( H_{UP} \) is the function of sub-pixel convolution proposed in ESPCN [22], \( F_{sf} \uparrow \) represents bi-linear upsampling of the shallow features \( F_{sf} \), and the symbol + represents the global skip connection (GSC). Specifically, we only construct one \( \times 2 \) higher-resolution sub-network because it does not introduce much computational burden. For \( \times 2 \) output we deactivate the upsampling module, and for \( \times 4 \) and \( \times 8 \) output we activate the upsampling module.

At the end of the network architecture, we use the reconstruction module to output the desired super-resolved image \( I_{SR} \):

\[
I_{SR} = H_{REC}(F_{up}),
\]

where \( H_{REC} \) is the function of the reconstruction module consisting of a convolutional layer for mapping feature space into color space.

Our MARDN is a supervised model, and a certain loss function is benefit for training our model. Some loss functions have been used in existing SR methods, such as \( L_1 \) [10], [11], [13], \( L_2 \) [5], [21], [39], [40], perceptual and adversarial losses [6], [15], [41]. For balancing the convergence rate and performance, we follow the optimization process as Zhao et al. [42]. We first train our model with the \( L_1 \) loss function, which is shown to speed up the convergence of training compared with \( L_2 \) loss function. At the second step, we use the \( L_2 \) loss function to fine-tune the model, which can achieve better PSNR performance. Given the training set of \( N \) pairs of LR images and their HR counterparts \( \{ I_{LR}^{i}, I_{HR}^{i} \}_{i=1}^{N} \), the goal of optimizing MARDN is to minimize the objective:

\[
L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \| H_{MARDN}(I_{LR}^{i}) - I_{HR}^{i} \|_p,
\]

where \( \Theta \) represents the parameter set of the MARDN model. The loss function is \( L_1 \) when \( p \) is 1 and \( L_2 \) while \( p \) is 2. More training details can be found in subsection V-A.

**B. SPACE-ATTENDED RESIDUAL DENSE BLOCK**

We now present more details about our basic block, the space-attended residual dense block (ARDB). As shown in Fig.3, we remove the batch normalization (BN) layers in our network since Lim et al. [11] and Wang et al. [15] have proved that batch normalization not only consumes large computational resources but also brings BN artifacts in SISR.

An ARDB consists of three space-attended residual dense units (ARDU) and a local skip connection (LSC) represented by dash arrow-line.

![Space-attended Residual dense Block (ARDB)](image-url)

**Figure 3.** Space-attended Residual dense Block (ARDB). An ARDB consists of three space-attended residual dense units (ARDU) and a local skip connection (LSC) represented by dash arrow-line.
C. FUSION BLOCK

In each stage, multi-resolution feature maps are generated from different sub-networks in parallel. Instead of directly fusing these feature maps via concatenation and 1×1 convolutional compression, our proposed fusion block (FB) fuses them with attention mechanism to enhance feature aggregation. As an example, we illustrate the last fusion block in the network, which receives three different resolution feature maps as input. The structure of this fusion block is shown in Fig.4. The first step of the fusion operation is to interpolate those low resolution feature maps to produce feature maps with the same size as the high resolution ones. The interpolated feature maps are first passed to global average pooling and squeeze-and-excitation layers (the convolutional layers), respectively. Then, the obtained three vectors with size of 64×1×1 are concatenated and processed with a softmax layer at the same channel index to generate an attention matrix, whose size is 64×3×1. After that, we split the matrix into three attention vectors with the size of 64×1×1 for the three sub-networks. Finally, we multiply the interpolated feature maps and the corresponding attention vectors, and sum them up to obtain the fused output.

The general representation of the fusion blocks is

$$F_{fb} = \sum_{r=1}^{n} v_r \otimes F_r,$$

(13)

where $v_r$ and $F_r$ denote the attention vectors and the interpolated feature maps of different sub-networks, respectively, $F_{fb}$ is the fused block output, and $n$ is the number of feature maps to be fused. For the case shown in Fig.4, $n = 3$.

IV. DISCUSSION

In this section, we compare the proposed model with similar works and discuss the difference between them.

A. DIFFERENCE BETWEEN MARDN AND CNF

CNF [18] and the proposed MARDN belong to multi-subnetwork designs. There are two differences between them. First, MARDN is an end-to-end training paradigm, while CNF trains different sub-network structure separately and fine-tunes through weighted assemble. Second, MARDN has different resolution sub-networks, while CNF only reconstructs information in the same resolution space.

B. DIFFERENCE BETWEEN MARDN AND RDN

Both RDN [13] and MARDN both utilize residual and dense connection to effectively extract features and stabilize the training of deeper network. However, MARDN is different from RDN in three aspects. First, MARDN has a spatial attention block after residual dense unit to focus on most relevant features, while RDN does not. Second, MARDN has a multi-resolution sub-network architecture, while RDN only contains the same resolution data flow. Third, MARDN uses a specifically designed channel-wise sub-network attention to better fuse the feature maps from different resolution sub-networks. In fact, RDN is a special case of MARDN when the space-attended weights are set to 1 and only the original-resolution sub-network is kept.

C. DIFFERENCE BETWEEN MARDN AND RCAN

Recently, the attention mechanism is applied to super-resolution task, and RCAN [8] is an exemplary work using channel attention. There are also three differences between RCAN and MARDN. First, RCAN uses residual skip connection to stabilize the training process, while MARDN contains both residual and dense skip connection. Second, RCAN only considers channel-wise correlations and rescales features with channel attention, while MARDN applies both spatial attention in the ARDBs and channel-wise sub-network attention in the fusion blocks. Third, RCAN utilizes a single network to extract feature from the same resolution features, while MARDN uses several sub-networks to extract features of different resolutions, handling low- and high-frequency information in different sub-networks.
V. EXPERIMENTS
In this section, we first detail our experiment setting. Then we conduct ablation studies and present quantitative and visual comparisons between our method and some state-of-the-art methods in bicubic and blur-downscale degradation model. Finally, we describe the model analyses.

A. EXPERIMENTS SETTING

1) DATASETS
Timofte et al. [27] published a high-quality image dataset DIV2K for image super-resolution, which contains 800 training, 100 validation, and 100 test images. Following [6], [8], [13], we train and validate our model on DIV2K dataset and test its performance on 5 benchmark datasets including Set5 [43], Set14 [20], B100 [44], Urban100 [45] and Manga109 [46].

2) DEGRADATION MODEL AND EVALUATION METRICS
We obtain the LR images by down-sampling the HR images leveraging the MATLAB bicubic (BI) kernel function. The experiments are conducted with the scaling factors of ×2, ×4, and ×8 between the HR and the LR images. The commonly used evaluation metrics PSNR and SSIM are utilized for quantitative comparisons with other methods on the Y channel. The visual results for qualitative comparisons are also provided.

3) TRAINING SETTING
As shown in Fig. 2, the proposed network is a full convolutional network. The convolutional layer has the kernel size 3 × 3 in default. It is followed by a Leaky ReLU activation layer. The model is first trained with $L_1$ loss with learning rate initialized as $1 \times 10^{-4}$ and then decreases by half every $1 \times 10^3$ iterations of back-propagation. After the training process converges, we change the objective function to $L_2$ to fine tune the parameters. For optimization, we use Adam [47] with $\beta_1 = 0.9, \beta_2 = 0.999, \text{ and } \epsilon = 10^{-8}$. Following Lim et al. [11], we randomly crop the input LR and HR images into 48 × 48 LR patches and HR patches according to the scaling factor, and then, these patches are augmented with random horizontal flipping and 90° rotation. Finally, we train the network in RGB channels with mini-batch size 16. Our MARDN model is implemented on the PyTorch framework [48] on 2 NVIDIA GeForce RTX 2080 GPUs.

B. ABLATION INVESTIGATIONS
To verify the effectiveness of the proposed MARDN model, we conduct a series of ablation investigations, including the effects of the components of the sub-networks, the skip connections and the fusion blocks. Table 1 shows the best PSNR(dB) values of different ablation cases on Set5 in 100 epochs with ×4 upscaling factor and Fig. 5 shows the corresponding convergence curves.

Table 1. Ablation investigations of sub-networks component, with/without skip connections and fusion blocks. We report the best PSNR(dB) values on Set5(×4) in 100 epoch.

| No. | SubN1 | SubN2 | SubN3 | LSC | GSC | FB | PSNR |
|-----|-------|-------|-------|-----|-----|----|------|
| 1   | ✓     |       |       | ✓   | ✓   | ✓  | 32.06|
| 2   |       | ✓     |       | ✓   | ✓   | ✓  | 32.06|
| 3   | ✓     |       |       | ✓   | ✓   | ✓  | 32.10|
| 4   | ✓     | ✓     |       | ✓   | ✓   | ✓  | 32.19|
| 5   | ✓     | ✓     | ✓     | ✓   | ✓   | ✓  | 32.15|
| 6   | ✓     | ✓     | ✓     | ✓   | ✓   | ✓  | 32.27|

Figure 5. The training convergence curve of ablation investigations on Set5. The comparisons are conducted on Set5 with the ×4 upscaling factor over 100 epoch.

1) COMPONENTS OF SUB-NETWORKS
To demonstrate the effect of our proposed multi-resolution sub-networks structure, we conduct two ablation experiments (No.1 and No. 2 in Table 1) by removing high- and low-resolution sub-network, respectively. Although these two experiments reach the same PSNR = 32.06, the No. 1 experiment is hard to train without high-resolution sub-network. As shown in Fig.5, the curve of the No.1 experiment oscillates heavily sometimes. In comparison, our MARDN achieves the best PSNR value, which demonstrates that our multi-resolution sub-networks can extract more effective features for reconstructing the super-resolved images.

2) GSC AND LSC
To verify the effectiveness of global skip connection (GSC) and local skip connection (LSC), we remove them separately in the No.3 and the No.4 experiments. It can been seen from Table 1 and Fig. 5 that their PNSR values are lower than the complete MARDN (No.6) in both cases. Furthermore, removing the LSC results in slower convergence and lower PNSR than removing the GSC, which indicates that the GSC passing shallow feature information can improve the performance.
3) FUSION BLOCKS
In the No.5 experiment, we remove all the fusion blocks, and replace them by concatenation and $1 \times 1$ convolution. Compared to the complete MARDN, the model without adaptive channel-wise sub-networks fusion blocks reduces the PSNR by 0.12dB. It demonstrates that the fusion blocks are effective in joining different feature maps from different sub-networks.

These ablation investigations validate the effectiveness of the proposed MARDN designs, including multi-resolution sub-networks, GSC, LSC and fusion blocks. All of them can bring performance improvements.

C. RESULTS WITH BICUBIC (BI) DEGRADATION MODEL
To evaluate the powerful reconstruction ability of our MARDN, we first compare it with 12 state-of-the-art CNN-based SR methods with BI distortion, including SRCNN [21], FSRCNN [39], VDSR [5], LapSRN [10], CNF [18], MemNet [23], EDSR [11], SRMDNF [14], D-DBPN [28], RDN [13], RCAN [8], and SAN [29].

1) QUANTITATIVE RESULTS
Table 2 presents all the quantitative results on the benchmark datasets with various scaling factors. Our MARDN performs similar to RCAN and SAN and better than the other state-of-the-art CNN-based methods. Specially, the MARDN performs relatively better for scale factors $\times 4$ and $\times 8$, but a little worse for scale factor $\times 2$. We consider that the inferior performance of our MARDN for scale factor $\times 2$ may be caused by deactivating the upscaling module, which increases the burdens of the previous sub-networks. For upscaling with large-scale factors, the input image lacks effective information for image super-resolution. To exploit limited information, we use the multi-resolution sub-networks architecture to separate and reconstruct different frequency information effectively, resulting in high-quality super-resolved images.

2) QUALITATIVE COMPARISON BY VISUALIZATION
Fig. 6 shows the visual comparison of different CNN-based methods. For the images ‘Babara’ in Set14 and ‘img_092’
in Urban100, recent methods RDN, RCAN and SAN suffer from artifacts with wrong textures to some extent, while our MARDN recovers more correct textures similar to the HR images and has the highest PSNR/SSIM values. In the image ‘UchiNoNyansDiary_000’, our MARDN keeps the original texture and achieves sharper result, while the other methods generate blurred or wrong textures of the cat ear. These observations demonstrate the effectiveness and superiority of our MARDN, which not only obtains more faithful results but also brings no or less artifacts via the multi-resolution structure.

D. RESULTS WITH BLUR-DOWNSCALE (BD) DEGRADATION MODEL
To further prove the robustness of our method, we also compare our MARDN with 8 CNN-based SR methods with BD distortion, including SPMSR [50], SRCNN [21], FSRCNN [39], VDSR [5], IRCNN [51], RDN [13], RCAN [8], and SAN [29].

1) QUANTITATIVE RESULTS
As commonly used in [8], [29], we compare ×3 SR results with BD distortion. As shown in Table 3, our MARDN achieves the best performance in each dataset. Compared with the very deep networks with second order attention in SAN, MARDN reconstructions high-quality images with multi-resolution sub-network structure and indicating our method can reconstruct super-resolved images with high quality from more complex degraded images.

2) QUALITATIVE COMPARISON BY VISUALIZATION
Similar to BI degradation, we also provide visual comparisons in Fig. 7. For reconstructing the challenging texture in image “253027”, RDN and RCAN recover some texture information in some degree, but still suffer from over-smoothness. In contrast, our MARDN recovers the correct pattern of zebra and reconstructs a clear super-resolved image like the original high resolution image. It clearly demonstrates the efficiency of our multi-resolution sub-network to separate the frequency information for high quality images.

E. MODEL ANALYSES
1) MODEL PERFORMANCE ANALYSIS
To find the best performance of our MARDN model, we first investigate the correlation between the performance and
TABLE 3. Quantitative results with the blur-down degradation model. Best and second best results are highlighted and underlined.

| Method       | Scale | Set5 PSNR | Set5 SSIM | Set14 PSNR | Set14 SSIM | B100 PSNR | B100 SSIM | Urban100 PSNR | Urban100 SSIM | Manga109 PSNR | Manga109 SSIM |
|--------------|-------|-----------|-----------|------------|------------|-----------|-----------|---------------|---------------|---------------|---------------|
| Bicubic      | × 3   | 28.78     | 0.8308    | 26.38      | 0.7271     | 26.33     | 0.6918    | 23.52         | 0.6862        | 25.46         | 0.8149        |
| SPMSR [50]   | × 3   | 32.21     | 0.9001    | 28.89      | 0.8105     | 28.13     | 0.7740    | 25.84         | 0.7856        | 29.64         | 0.9003        |
| SRCNN [21]   | × 3   | 32.05     | 0.8944    | 28.80      | 0.8074     | 28.13     | 0.7736    | 25.70         | 0.7770        | 29.47         | 0.8924        |
| PSRCNN [39]  | × 3   | 26.23     | 0.8124    | 24.44      | 0.7106     | 24.86     | 0.6832    | 22.04         | 0.6745        | 23.04         | 0.7927        |
| VDSR [5]     | × 3   | 33.25     | 0.9180    | 29.46      | 0.8244     | 28.57     | 0.7893    | 26.61         | 0.8136        | 31.06         | 0.9234        |
| IRCNN [51]   | × 3   | 33.38     | 0.9182    | 29.63      | 0.8281     | 28.65     | 0.7922    | 26.77         | 0.8154        | 31.15         | 0.9245        |
| RDN [13]     | × 3   | 34.58     | 0.9280    | 30.53      | 0.8447     | 29.23     | 0.8079    | 28.46         | 0.8582        | 33.97         | 0.9465        |
| RCAN [8]     | × 3   | 34.70     | 0.9288    | 30.63      | 0.8462     | 29.32     | 0.8093    | 28.81         | 0.8645        | 34.38         | 0.9483        |
| SAN [29]     | × 3   | 34.75     | 0.9290    | 30.66      | 0.8466     | 29.33     | 0.8101    | 28.83         | 0.8646        | 34.46         | 0.9487        |
| MARDN (ours) | × 3   | 34.78     | 0.9299    | 30.77      | 0.8471     | 29.33     | 0.8106    | 28.91         | 0.8660        | 34.48         | 0.9491        |

**FIGURE 7.** Visual comparison for ×3 SR with the BD model on the BSD100 dataset. The best result is highlighted.

**TABLE 4.** PSNR & Parameters v.s. number of ARDBs of low(L) resolution, original resolution (O) and high resolution(H) sub-networks on Set5 (×4) in 100 epoch.

| Model          | L5O4H3 | L2O7H1 | L4O7H2 | L5O7H3 | L6O7H4 |
|----------------|--------|--------|--------|--------|--------|
| PSNR (dB)      | 31.92  | 32.17  | 32.25  | 32.27  | 32.23  |
| Param. (M)     | 19.1   | 15.5   | 20.6   | 23.5   | 26.3   |

the number of ARDBs. As shown in Table 4, the model L5O7H3 with 5, 7 and 3 ARDBs in the low-, original- and high-resolution sub-network, respectively, performs better than those with fewer ARDBs. However, it can also be seen that the L4O7H2 and L5O4H3 have similar number of parameters, but obviously L4O7H2, which has more ARDBs in its original-resolution sub-networks, performs better than L5O4H3. On the other hand, L6O7H4 has a deeper network, but performs worse than L5O7H3. For balancing the calculation burden and performance, we use the L5O7H3 as our final setting.

2) MODEL SIZE ANALYSIS

Fig. 8 shows the performance and model sizes of some state-of-the-art deep CNN-based SR methods. Our MARDN model has the best performance, but it has a large number of parameters. To reduce the parameters, we remove a space-attended residual dense unit from the ARDB, and name it as MARDN lite model. It can be seen that the MARDN lite model has similar number of parameters to the SAN model, but performs better than the other state-of-the-art methods.

3) TIME COMPLEXITY ANALYSIS

Considering the execution time, we use the public official implement testing code of recently CNN-based SR methods, including D-DBPN, SAN and RCAN on the machine with 4.2GHz Inter i7 CPU (32G RAM) and Nvidia RTX 2080 platform. We conduct five times inference over the 100 images of Urban100 dataset, and the mean values of per image are showed in Table 5. It is noteworthy that our MARDN is a real-time model, whose inference time is the least but has the highest PSNR.

4) FEATURE MAPS VISUALIZATION IN MULTI-RESOLUTION SUB-NETWORKS

To verify whether our MARDN recovers the features of different frequencies separately, we visualize the feature maps before the last fusion block, which are depicted in Fig. 9.
In our model, we utilize the low-resolution sub-network to process the low-frequency information, and the high-resolution one to learn the high-frequency information. It is obvious that as the resolution increases, the feature map exhibits more high-frequency texture information, which demonstrates that our multi-resolution sub-networks have powerful ability to process different frequency information, separately.

VI. CONCLUSION

This paper proposed a multi-resolution space-attended residual dense network (MARDN) for image super-resolution. Specifically, different frequency feature maps are produced through controlling the resolution and the depth of the sub-networks. To effectively fuse the feature maps from different sub-networks, a channel-wise sub-network attention fusion block was introduced to adaptively join the features. Extensive experiments on SR with bicubic (BI) and blur-downscale degradation models showed the effectiveness of our MARDN in terms of quantitative and visual results. Ablation investigations also validated the effectiveness of multi-resolution sub-networks, skip connection and fusion blocks in our MARDN. Furthermore, visualization of feature maps in multi-resolution sub-networks verified that the high-resolution sub-network with few stages can extract high-frequency information, while the low-resolution sub-network with more stages can extract low-frequency information. In our future work, we aim to design a light-weight backbone network to offset the parameter increase within multi-sub-networks.

REFERENCES

[1] J. M. Haut, R. Fernandez-Beltran, M. E. Paololetti, 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.
[2] A. S. Chaudhari, Z. Fan, K. 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.
[3] W. W. Z. Zou and P. C. Yuen, “Very low resolution face recognition problem,” IEEE Trans. Image Process., vol. 21, no. 1, pp. 327–340, Jan. 2012.
[4] 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. Springer, Sep. 2014, pp. 184–199.
[5] 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.
[6] C. Ledig, L. Theis, F. Huszar, 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.
[7] 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.
[8] 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. Euro. Conf. Comput. Vis. (ECCV) Springer, Sep. 2018, pp. 286–301.
[9] T. Tong, Q. Li, X. Liu, and Q. Gao, “Image super-resolution using dense skip connections,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4799–4807.
[10] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 624–632.
[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] 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.
[13] 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.
[14] 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.
[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV) Springer, Sep. 2018, pp. 63–79.
[16] Y. Chen, J. Wang, X. Chen, M. Zhu, K. Yang, Z. Wang, and R. Xia, “Single-image super-resolution algorithm based on structural self-similarity and deformation block features,” IEEE Access, vol. 7, pp. 58791–58801, 2019.
[17] Z. Tang, S. Li, L. Luo, M. Fu, H. Peng, and Q. Zhou, “Image super-resolution via simplified dense network with non-degenerate layers,” IEEE Access, vol. 7, pp. 24775–24787, 2019.
[18] H. Ren, M. El-Khamy, and J. Lee, “Image super-resolution based on fusing multiple convolution neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 54–61.
[19] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5693–5703.
[20] R. Zeyde, M. Elad, and M. Porta, “On single image scale-up using sparse-representations,” in Proc. Int. Conf. Curves Surf. Springer, Jun. 2010, pp. 711–730.
[21] 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.
[22] W. Shi, J. Caballero, F. Huszar, 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.

[23] 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.

[24] J.-H. Kim and J.-S. Lee, “Deep residual network with enhanced upsampling module for super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 800–808.

[25] 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.

[26] 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.

[27] 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. 114–125.

[28] M. Haris, G. Shakhnarovich, and N. Ukiti, “Deep back-projection networks for super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1664–1673.

[29] 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. Workshops (CVPRW), Jun. 2019, pp. 11–065–11–074.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[31] A. Jolicoeur-Martineau, “The relativistic discriminator: A key element missing from standard GAN,” 2018, arXiv:1807.00734. [Online]. Available: http://arxiv.org/abs/1807.00734

[32] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, “The 2018 PIRM challenge on perceptual image super-resolution,” in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), Sep. 2018.

[33] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, “Ranksrgan: Generative adversarial networks with ranker for image super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3096–3105.

[34] R. Zhou and S. Susstrunk, “Kernel modeling super-resolution on real low-resolution images,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2433–2443.

[35] A. Lugmayr, M. Danelljan, and R. Timofte, “Unsupervised learning for real-world super-resolution,” 2019, arXiv:1909.09629. [Online]. Available: http://arxiv.org/abs/1909.09629

[36] L. Zhang, Y. Zhu, and W. X. Zheng, “Synchronization and state estimation of a class of hierarchical hybrid neural networks with time-varying delays,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 2, pp. 459–470, Feb. 2016.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, Sep. 2014, pp. 740–755.

[38] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 3686–3693.

[39] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, Oct. 2016, pp. 391–407.

[40] 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.

[41] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, “EnhanceNet: Single image super-resolution through automated texture synthesis,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4491–4500.

[42] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for neural networks for image processing,” 2015, arXiv:1511.08861. [Online]. Available: http://arxiv.org/abs/1511.08861

[43] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” in Proc. Proceedings Brit. Mach. Vis. Conf. (BMVC), 2012, pp. 1–10.

[44] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV), vol. 2, Jul. 2001, pp. 416–423.

[45] 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.

[46] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, “Sketch-based manga retrieval using manga109 dataset,” Multimedia Tools Appl., vol. 76, no. 20, pp. 21811–21838, Nov. 2016.

[47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, “Pytorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8024–8035.

[49] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super-resolution with sparse prior,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 370–378.

[50] T. Peleg and M. Elad, “A statistical prediction model based on sparse representations for single image super-resolution,” IEEE Trans. Image Process., vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[51] 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.
LONGCUN JIN (Member, IEEE) is currently an Associate Professor with the School of Software Engineering, South China University of Technology, China. His research projects include the National Natural Science Foundation of China, the Doctoral Program Fund of Ministry of Education, Science and Technology Program of Guangdong Province, the Open Research Fund of State Key Laboratory, the Science and Technology Program of Guangzhou City, and the Key Projects of Basic Scientific Research Operating Expenses of Central Universities of Ministry of Education. His research interests include computer vision and artificial intelligence. He is a member of ACM, AAAI, AAAS, and SIAM. He has received the Baosteel Award, the Hong Kong Scholars Award, and the Pearl River Science and Technology Star.

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 nine million. Among them, nine projects above the provincial level have passed the acceptance inspection. His research interests include artificial intelligence and data mining.

***